Phase-I Study of the Effects of Professional Development and Long-Term Support on Program Implementation and Scaling Up: Final Report

Paul R. Brandon, Alice K. H. Taum, Carlos C. Ayala,
Donald B. Young, Mary E. Gray, Thomas W. Speitel,
Thanh Truc T. Nguyen, and Francis M. Pottenger III

Curriculum Research & Development Group
College of Education
University of Hawaiʻi at Mānoa
Honolulu, Hawaiʻi

May 2007

Phase-I Study of the Effects of Professional Development and Long-Term Support on Program Implementation and Scaling Up: Final Report

Paul R. Brandon, Alice K. H. Taum, Carlos C. Ayala,
Donald B. Young, Mary E. Gray, Thomas W. Speitel,
Thanh Truc T. Nguyen, and Francis M. Pottenger III

Curriculum Research & Development Group
College of Education
University of Hawaiʻi at Mānoa
Honolulu, Hawaiʻi

May 2007

This is the final report of a project funded by the National Science Foundation (Grant No. REC0228158), Paul R. Brandon, Principal Investigator (PI), Donald B. Young, co-PI, and Alice K. H. Taum, Project Manager. Paul Brandon oversaw the project, with Donald Young's assistance, and was in charge of developing the teacher questionnaire and the instrument for judging program quality. In addition to managing the project, Alice Taum was in charge of developing the observation coding protocol. Carlos Ayala of Sonoma State University developed the student assessment. Thomas Speitel and Mary Gray oversaw the development of the alternative version of professional development (PD) that is described in the study. Thanh Truc Nguyen contributed significantly to the development of the multimedia DVD-ROM that was developed as part of the alternative PD package. Francis M. Pottenger served as a representative, with Donald Young, of the developers of Foundational Approaches in Science Teaching program. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation. An electronic copy of the report, given in one file for the body of the report and another for the appendixes, is available at http://www.hawaii.edu/crdg/programs/pre/scup.html .

Contents

TABLES

FIGURES

ACKNOWLEDGMENTS

The authors are indebted to many contributors to this project. Richard Shavelson, Maria Ruiz-Primo, Jane Kahle, Paul Lemahieu, J. Bradley Cousins, and Thomas Guskey served as our advisory board. Byron Inouye and Paul Sosso contributed significantly to the development of the multimedia DVD-ROM. Brian Lawton and Lance Linke helped in instrument development and provided statistical support. Carolyn Glass structured and monitored data collection. Mary Lee, Rachael Jones, and Qi Zeng served ably as graduate assistants. Paula Takei and Tamara Teffeteller helped develop our observation instrument and served as our primary observation coders. Tom Scarlett participated with the curriculum developers in reviewing drafts of data collection instruments. James Bell, Susan Boynton, James Harpstrite, and Mathias Kainrath served as videographers. Christina Justiniano, Angela Leung, Cassandra Payne, and Jenny Lee served as student assistants. We are grateful to all of them for their dedicated assistance during the project.

Phase-I Study of the Effects of Professional Development and Long-Term Support on Program Implementation and Scaling Up: Final Report

EXECUTIVE SUMMARY

Paul R. Brandon, Alice K. H. Taum, Carlos C. Ayala, Donald B. Young,
Mary E. Gray, Thomas T. Speitel, Thanh Truc T. Nguyen, and Francis M. Pottenger III

Curriculum Research & Development Group, College of Education, University of Hawaiʻi at Mānoa

May 2007

This is the executive summary of the final report for a National Science Foundation (NSF) project (Grant No. REC022818). The project was conducted by researchers and curriculum developers at Curriculum Research & Development Group (CRDG), College of Education, University of Hawaiʻi at Mānoa. It was funded by the Interagency Educational Research Initiative (IERI), a collaboration among NSF, the U. S. Department of Education, and the National Institutes of Health. The project was a two-component, preparatory phase for a randomized study of the effects of variations in professional development (PD) duration and long-term training, with multimedia support, on (a) the implementation and student outcomes of middle-school inquiry-based science (here called *inquiry science*) on the wide dissemination ("scaling-up") of inquiry science. The version of inquiry science that we examined was CRDG's Foundational Approaches in Science Teaching (FAST) program. The project's two components were (a) the development of an alternative version of FAST PD for the first year of FAST to examine in the second phase of the study and (b) the development and validation of data collection instruments to use in the second phase.[1]

In this summary, we describe the development of the alternative version of FAST PD, describe the development of instruments, and summarize the findings of studies of the validity of data collected with the instruments.

Development of the Alternative Version of FAST PD

The alternative version of FAST PD that we developed is called FASTPro. It was developed by a team from CRDG's Learning Technology Section, including the section head, an instructional designer, two videographers, a professional video editor, and a graphic designer. The instructional designer and graphic designer also served as multimedia programmers. FASTPro consists of a one-week face-to-face institute (FASTStart), an electronic resource in the form of a multimedia DVD-ROM for the trained teachers (FASTeR), and an on-line course (FASTForward). FASTStart addresses the essential skills and concepts that are best taught in face-to-face PD. The knowledge and skills that are traditionally covered in the second week of FAST PD institutes are addressed in FASTeR and FASTForward. FASTeR includes video of FAST institutes, including video of the FAST trainer and of teachers in the role of students during the institutes; video showing FAST being taught in the classroom, including segments on students and segments on teachers; photographs and animation (in the form of slide shows) of setting up student investigations; and photographs of laboratory materials and equipment. FASTeR has one or more of these PD aids for 19 FAST 1 student science investigations. A survey of teachers about the extent to which they would use FASTeR showed strongly positive opinions about the DVD-ROM. The development of FASTForward was partially completed during the project and as of May 2007 was ongoing.

[1] No Phase-II grants were funded in the year that we applied for our second, follow-up grant. New funding for all IERI grants has been discontinued; the program is being phased out.

Instrument Development

The instruments that were developed for comparing FASTPro with traditional FAST PD included the Inquiry Science Observation Code Sheet (ISOCS), the Inquiry Science Teacher Questionnaire (ISTQ), the Inquiry Science Questioning Quality (ISQQ) method, and the Inquiry Science Student Assessment (ISSA). The ISOCS is a method for coding and analyzing videotaped observations of teachers in inquiry science classrooms, with a focus on the interaction between teachers and students that is initiated by teachers' questions of students. It was developed in close consultation with FAST developers in more than 40 review-and-revision cycles over a two-year period. The ISTQ is a self-report instrument for collecting data on (a) the implementation of inquiry science in the classroom and (b) the context within which teachers implement inquiry science, including teacher demographics; teacher perceptions, behaviors, attitudes, opinions, interests, and beliefs; some classroom variables; and the support the school provides teachers to implement inquiry science. Implementation was measured with the ISTQ's Inquiry Science Implementation Scale. The context scales on the instrument included the Collaboration Frequency Scale, the Collaboration Benefits Scale, the Teacher Participation in Science Activities Scale, and the School Support for Inquiry Science Scale. Coefficient alphas were all high, and a test-retest study of the Implementation Scale showed high reliability. The ISSQ uses the paired-comparison method, conducted by expert judges, for measuring the quality of program implementation. A group of five FAST trainers served as judges and tried out the method. The ISSA is our student outcome measure. It includes multiple-choice items, a performance assessment, and attitudinal items that can be examined not only as outcome measures but also as measures of the participant responsiveness aspect of program implementation. It was developed and validated by Carlos C. Ayala of Sonoma State University, who collected validity data from over 400 students in the classrooms of 10 Hawai‘i teachers .

Validity Data Collection and Analyses

After the lengthy development period, data for conducting validity studies were collected. Data for studies of the ISOCS and the ISQQ were from videotapes of 107 classroom periods of 16 FAST teachers in Hawai‘i. Data for studies of the ISTQ were from two rounds of instrument administration to two samples of FAST teachers nationwide (*N*s = 79 and 156). Validity data for examining the ISSA were collected in 10 FAST classrooms in Hawai‘i. An overview of the extent to which the findings of the data collection and analyses support conclusions favoring the validity of the data is are shown in Table ES-1.

Inquiry Science Observation Code Sheet (ISOCS) Validity

Evidence is given in the full report for the content-related validity, concurrent validity, and criterion-related validity of data collected with the ISOCS. The care with which the ISOCS development was conducted and the thoroughness of the process are evidence of content-related validity; they show that the instrument is designed to collect data that are (a) relevant to the measurement task (i.e., the data reflect what is intended to be observed) and (b) representative of the content domain. Videotapes were analyzed for nine of the 16 teachers whose classrooms had been observed. Analyses of the consistency of ratings between two coders showed a Pearson correlation of .99 between the numbers of codes assigned to the coding categories and a correlation of .53 between the total number of codes assigned to the teachers. We believe that these correlations are evidence of reliability. The consensus findings were less favorable. Agreement on the percentage with which two coders assigned identical codes in the first, independent round of coding ranged from 5% to 50%. These percentages are not high. Clearly, the reconciliation step in the coding process is essential for collecting ISOCS data. However, we believe that ISOCS inter-coder agreement results cannot be compared with the desirable or typical results for observations in which events are recorded in time periods (say, one code for every five-minute period). Our method is more stringent than this method, because our coders had to agree on precisely on codes at each observed moment of a class.

For the ISOCS concurrent validity analyses, we correlated the ISOCS results for the nine teachers with the results on the ISQQ. The Spearman's rho correlation of the ISQQ quality ranks with the percentage of codes that had been assigned in the student-comment code category = .52, and the Spearman's rho correlation of ISQQ ranks with the percentage that the teachers used follow-up statements and probing questions = .45. (The unit of analysis was the teacher; each teacher's percentage for a given code = the percentage of all the codes assigned for the teacher.) We believe that these correlations show a relationship between the two sets of results, thus supporting the validity of the ISOCS data.

For the ISOCS criterion-related validity analyses, we correlated the ISOCS results (in the form of teacher-student question-response exchanges, coded as the percentage of the total number of observed behaviors) with mean achievement test scores for six teachers. The correlation = .96—not a conclusive result, because the number of teachers analyzed is small, but nevertheless strongly suggestive of the validity of the observation data.

Inquiry Science Teacher Questionnaire (ISTQ) Validity

We reported evidence for the content-related validity, concurrent validity, and criterion-related validity of data collected with the ISTQ. Analyses were conducted of the ISTQ scales that measure inquiry science implementation and the context within which inquiry science is implemented. Data for the validity analyses were collected from a sample, closely representative of K–12 teachers nationwide, of 79 FAST teachers. Test-retest data for the Implementation Scale were collected from 156 FAST teachers nationwide.

The procedures with which the instrument was developed provide evidence of content-related validity. The results of reliability analyses, including coefficient alphas, factor analyses, and a test-retest study also provide content-related validity evidence. The alphas and the results of the factor analyses show that the implementation scale and the context scales are reliable. Because of the small number of teachers, we supplemented the factor analyses with parallel analyses; the results of all of them supported the appropriateness of our decision to conduct factor analyses. The results of the test-retest analysis of the Implementation Scale (Pearson correlation between the two administrations of the test = .76, and no variance due to occasion, as shown in a generalizability theory analysis) strongly support the reliability of data collected with the scale.

The results of concurrent validity analyses of ISTQ data show that the pattern of correlations among three context scales was as expected. The Pearson correlation of the Implementation Scale results with data collected on an implementation teacher log, which we developed and administered on multiple occasions to 66 of the teachers who completed the ISTQ = .66—substantial evidence for concurrent validity. The Pearson correlation of results on Teacher-Student Interaction factor of the ISTQ Implementation Scale with the ISOCS teacher-student interaction results =.50 and with the ISQQ ranks = .39. Both of these correlations are evidence of concurrent validity, although the strength of the evidence in both cases is tempered by outliers.

The results of the criterion-related validity analyses show a Pearson correlation of .37 between the Teacher-Student Interaction Subscale of the implementation scale with the ISSA multiple-choice posttest and a correlation of .43 with the ISSA extended-response item posttest. These correlations provide some evidence of criterion-related validity, although they were somewhat affected by an outlier on the ISSA.

Inquiry Science Questioning Quality (ISQQ) Validity

We gathered evidence of the content-related validity and concurrent validity of data collected with the ISQQ. The procedures with which the ISQQ was developed provide some evidence of content-related validity, although the strength of the evidence is qualified somewhat by the feedback of the five ISQQ judges: Some of the judges reported that it was difficult to make holistic judgments about quality, and some tended to add quality criteria of their own to those specified in the ISQQ procedures. Reliability findings, generated by five analyses, each from a different measurement tradition, also were mixed. Further analysis of the results showed that the judgments of two of the five ISQQ judges were less reliable than the judgments of three of the others, suggesting that the criteria for judging quality need revision and that the training period needs to be lengthened.

The results of concurrent validity analyses that are reported above in the section for the ISOCS provide good evidence for the validity of data collected with the ISQQ.

Inquiry Science Student Assessment (ISSA) Validity

We reported evidence for the content-related and concurrent validity of the achievement test, performance assessment, and seven attitudinal scales on the ISSA. Content-related validity evidence is found in the careful development of the instruments and in the alpha coefficients for components of the instrument. Concurrent validity evidence is found in the correlations among performance assessment components, of attitudinal scale scores among each other, and of the attitudinal scale results with the achievement test results, which all showed the expected pattern.

Table ES-1

The Extent to Which the Findings of Validity Studies Conducted During the Phase-I Study of the Effects of Professional Development and Long-Term Support on Program Implementation and Scaling Up Suggest Validity

Instrument or scale	Content-related validity evidence	Concurrent validity evidence	Criterion-related validity evidence
Inquiry Science Observation Code Sheet	*Strongly suggestive of validity:* •Careful, thorough development procedures *Moderately suggestive of validity:* •Reasonably high correlations between raters *Slightly suggestive of validity:* •Consensus results show lower percentages than desirable (although results are not comparable to observation rating studies)	*Strongly suggestive of validity:* •Correlations of .52 and .45 with quality ranks from the ISQQ	*Strongly suggestive of validity:* •Correlation of .96 with ISSA scores
Inquiry Science Teacher Questionnaire (ISTQ) Implementation Scale	*Strongly suggestive of validity:* •Careful, thorough development procedures •High test-retest correlation; no variance due to occasion (G-theory analysis) •Factor analysis and coefficient alpha results strong for Factor 1 (Teacher-Student Interaction) •Coefficient alpha results strong for Factor 2 (Connecting to the Outside World) *Moderately suggestive of validity:* •Coefficient alpha results acceptable for Factor 3 (Introducing the Investigation) *Contrary evidence:* •Small eigenvalues for Factors 2 and 3	*Strongly suggestive of validity:* •Correlation of .66 with teacher log results *Moderately suggestive of validity:* •Correlation of .50 between the Teacher-Student Interaction factor and the ISOCS teacher-student interaction results (small *N*; results perhaps affected by outlier) •Correlation of .39 with teacher quality rank (ISQQ) (small *N*; results perhaps affected by outlier)	*Moderately suggestive of validity:* •Correlation of .43 between the Teacher-Student Interaction factor and the ISSA total extended-response score and of .37 between the factor and the ISSA multiple-choice posttest
ISTQ context scales	*Strongly suggestive of validity:* •Careful, thorough development procedures	*Strongly suggestive of validity:* •The pattern of correlations about the total scale scores was as expected.	—
Inquiry Science Questioning Quality (ISQQ) method	*Strongly suggestive of validity:* •Careful, thorough development procedures *Slightly suggestive:* •Reliability results were mixed	*Strongly suggestive of validity:* •Correlations of .52 and .45 with teacher-student interaction on the ISOCS	—
Inquiry Science Student Assessment	*Strongly suggestive of validity:* •Careful, thorough development procedures •High coefficient alphas	*Strongly suggestive of validity:* Positive correlations found among attitude scales.	—

CHAPTER I
PURPOSE, RATIONALE, BACKGROUND, AND THEORETICAL MODEL

PURPOSE

This report describes Phase I of what was intended as a two-phase project addressing variation in two aspects of K–12 PD: (a) its duration and (b) long-term support, in the form of an online course supported by a multimedia DVD, provided to teachers after initial training institutes. The project was funded by the National Science Foundation Interagency Education Research Initiative (IERI) (Grant No. REC 0228158).

Most of the conclusions reported in the literature on duration and long-term computer-based support are not from randomized experiments. Our project was the preparatory phase, with two components, for a randomized study of the effects of variations in PD duration and long-term training, with multimedia support, on the implementation and student outcomes of middle-school inquiry-based science (here called *inquiry science*), as well as the long-term effects of these variations on scaling-up. The focus of the preparatory phase of the project was on inquiry science in the form of Foundational Approaches in Science Teaching (FAST), an award-winning middle school program developed and disseminated by Curriculum Research & Development Group (CRDG), University of Hawaiʻi at Mānoa (UHM). The project included two major components: (a) the development of an alternative version of FAST PD for the first year of FAST and (b) the development and validation of data collection instruments to use in the second phase of the study.

The IERI program was begun in 1999 in response to the call of the President's Committee of Advisors on Science and Technology to fund studies that "scale-up" programs that have previously shown success in the laboratory or in a few sites. As NSF said in the program solicitation, IERI was "designed to help educators integrate the insights of scientific research on educational improvement into the realities of varied educational contexts to produce sustainable improvements in learning for diverse student populations" (National Science Foundation, 2004, Introduction, ¶1). Preferably, the programs were to be interdisciplinary and were to incorporate technology for teaching or learning (Brown, McDonald, & Schneider, 2006).

Our two-year Phase I project began in March, 2003. In the second year of the project, we submitted an IERI proposal for the second phase: a five-year randomized experiment that would use the instruments developed in the first phase to compare the traditional version of FAST with the alternative version developed in Phase I. However, NSF did not fund the second phase. (Indeed, it did not fund any Phase II IERI projects that year, and it began to phase out the program thereafter.) Because we were no longer on a tight timeline to prepare for Phase II, and because we needed additional time to complete the development of an observation protocol, we requested and received two one-year, no-cost project extensions.

This is the final report for the fours years of the project. We begin in this chapter with a presentation of the rationale for the Phase II study. This is followed by a presentation of the foundation for the PD-development component of the study, including background on K–12 inquiry science and a description of FAST. We conclude this chapter with a description of the theoretical model underlying the instrument development component of the project. In Chapter II, we describe the PD-development component. In Chapters III and IV, we describe the instrument component of project, including the development (Chapter III) and validation (Chapter IV) of a classroom observation protocol, teacher questionnaire, method for studying the quality of teachers' implementation of inquiry science, and suite of student assessments. Earlier versions of much of the material given here were presented as conference papers, as noted throughout the report.

RATIONALE FOR THE PROJECT

K–12 teacher PD has been the focus of an increasing number of research studies, articles, papers, and reports. Many recommendations about developing and conducting PD are found in this literature (see Joyce & Showers, 1995 and Supovitz, 2001, among others, for summaries). For example, the literature states that effective PD (a) models inquiry teaching (Arons, 1989; Bybee, 1993; Little, 1993; McDermott, 1990); (b) is aligned with standards (Garet, Porter, Desimone, Birman, & Yoon, 2001); (c) addresses theory, demonstration, practice, feedback, and coaching (Joyce & Showers, 1987) and focuses on content (Cohen & Hill, 1998); (d) demonstrates the connection of the PD material to student performance standards (Hawley & Valli, 1999); (e) engages teachers in specific teaching tasks and gives them opportunity for extensive practice (Darling-Hammond & McLaughlin, 1995); (f) is sustained (Supovitz, 2001; Supovitz, Mayer, & Kahle, 2000) and conducted in conjunction with teachers' ongoing classroom duties (Zigarmi, Betz, & Jennings, 1977); (g) provides long-term "support coupled with pressure" (Joyce, Showers, & Rolheiser-Bennet, 1987, p.23); (h) provides teachers with regular feedback on their efforts (Guskey, 1995); and (i) considers teachers' educational context (Guskey, 1995; Little 1993; Loucks-Horsley, Hewson, Love, & Stiles, 1998; McLaughlin & Marsh, 1990). Garet et al. and Supovitz et al., among others, have reported positive relationships between the characteristics of effective PD and program implementation. Furthermore, research shows that the greater the fidelity of implementation of treatments such as PD, the better the effects (e.g., Mowbray, Holter, Teague, & Bybee, 2003).

The foundation for the rationale for our project rests on two pillars: that more needs to be known about the effects of the duration of PD and that more needs to be known about the effects of online and computer-based PD.

The Duration of PD

Many conclusions can be drawn from the PD research literature, but more needs to be known about the effects of the duration of PD. The findings on this topic are inconclusive. Some have shown a positive relationship between the duration and the effects of PD. Garet et al. (2001, pp. 921–922), for example, claimed that

> almost all of the recent literature on teacher learning and professional development calls for professional development that is sustained over time. The duration of professional development is expected to be important in two ways. First, longer activities are more likely to provide an opportunity for in-depth discussion of content, student conceptions and misconceptions, and pedagogical strategies. Second, activities that extend over time are more likely to allow teachers to try out new practices in the classroom and obtain feedback on their teaching.

In their study of a national probability sample of teachers, Garet et al. (2001, p. 933) concluded that “professional development is likely to be of higher quality if it is both sustained over time and involves a substantial number of hours.” They found that time span and contact hours had a substantial positive effect on opportunities for active learning and on the coherence of the PD and that they had a moderate positive effect on the emphasis on content knowledge. Supovitz and Turner (2000) analyzed data collected on the effects of Local Systemic Change Initiative and found that teachers with more than 80 hours PD reported using inquiry-based practice about .2 st. dev. more than the average teacher. These are slight differences, however, as is apparent in a comparison of the percentages of teachers in the two groups reporting positive changes (see Weiss, Montgomery, Ridgway, & Bond, 1998).

Other studies, however, have shown no positive relationship between duration and good effects of PD. Desimone, Porter, Garet, Yoon, and Birman’s (2002) longitudinal study of 207 teachers showed no effect of the number of PD contact hours or of the span of PD on teachers’ (a) self-reported use of technology, (b) use of higher order instructional practices, or (c) use of alternative student assessment practices. Kennedy (1998) reviewed eight mathematics and four science PD studies that examined the effects on student achievement and found that total contact time and concentrated vs. distributed contact hours did not affect achievement. The study in which teachers had a concentrated, four-week summer institute showed somewhat less of an effect on achievement than the other three studies. Our Phase II study was planned to help resolve questions about the effects of PD scheduling and duration.

Online PD

Online PD occurs using computers over a network, usually the Internet. Sometimes it occurs between people and computers, without interaction with other people, and sometimes it occurs among people using computers. Most of the research on online education has addressed post-secondary course-taking, and most of the research on how K–12 teachers learn online has addressed the formation of communities of learning and practice (e.g., Barab, MaKinster, & Scheckler, 2004; Schlager & Schank, 1997) and patterns of discourse (e.g., Anderson & Christiansen, 2004; Polin, 2000). The jury is out on the extent to which online education is effective. Jones and Paolucci (1998) concluded in an extensive literature review that research findings did not strongly support the effectiveness of technology-mediated instruction (see also Matthews, 1998; Price, 1996; van Dusen, 2000). Furthermore, some evidence, although anecdotal, suggests that teachers prefer face-to-face institutes over online PD (Brunvand, Fishman, & Marx, 2003; Honey & Moeller, 1990; Little, 1993). However, the distance-learning literature (e.g., Hannafin, Hill, Oliver, Glazer, & Sharma, 2003; Hara & Kling, 1999; Harasim, Hiltz, Teles, & Turoff, 1995) and the online PD literature (e.g., Barab, MaKinster, Moore, & Cunningham, 2001; Fishman, Marx, Best, & Tal, 2003; Kabilan, 2004; Marx, Blumenfield, Krajcik, & Soloway, 1998) suggest that online PD can successfully supplement institutes when it focuses on teacher motivation, skills, knowledge, self-directed learning, and technology skills, including competence in interactive computer methods. In particular, Web-based courses have been shown to increase teachers' self-efficacy (Huai, Braden, White, Elliot, 2003), help them learn new skills, and help them implement new pedagogical approaches (Schlager & Schank, 1997). One of the best available studies compared online and face-to-face versions of the same class (Harlen & Doubler, 2004) and showed that they had about the same effects on teacher beliefs. Some researchers (e.g., Vrasidas & Glass, 2004) have concluded that the optimal approach is to combine face-to-face institutes with online PD.

BACKGROUND

Inquiry Learning in Science Education

Although conclusions about the effects of inquiry science teaching and learning are not unanimous (e.g., Kirschner, Sweller, & Clark, 2006), many years of research has demonstrated their positive effects. The Inquiry Synthesis Project, a large-scale review of the research on the topic, is expected to publish its findings soon.[1] Until then, we are relying on summaries of the research by Tamir (1983) and the National Research Council (1996), who reported that the research "suffers from the lack of a shared, precise definition of inquiry" but that reviewing the research yields "patterns that show up across studies" (National Research Council, p. 124). The National Research Council and Tamir showed that inquiry science has positively affected student achievement and attitudes; process skills; problem solving and creativity; vocabulary knowledge; conceptual understanding and critical thinking; inquiry abilities; and "scientific ways of thinking, talking, and writing" (National Research Council, p. 125). Studies conducted since the

[1] Preliminary, informal reports of the results of the project have shown that inquiry science has had positive effects.

compilations by Tamir and the National Research Council (e.g., Lee, Hart, Cuevas, & Enders, 2004; Wu & Hseih, 2006) have continued to show positive effects. Furthermore, *Education Week* reported a survey of 1,000 college of education deans and elementary school teachers: "Most of the deans and teachers, 95 percent and 93 percent, respectively, reported that inquiry-based science lessons, which include hands-on activities, are the most effective way of teaching the subject because they engage students in the lessons" (Galley, 2004, p. 12).

There are a multitude of descriptions of inquiry science education (Inquiry Synthesis Project, 2004). Traditionally, these descriptions focus on the steps of a simplified process of scientific inquiry, as given in the National Science Education Standards (National Research Council, 1996); these steps include developing questions, developing a plan to collect evidence addressing the questions, collecting the evidence, explaining the evidence, connecting the explanations to existing scientific knowledge, and communicating and justifying the explanations. Variations in these steps reflect the degree of student independence in conducting inquiry. Teachers who lead students through the process typically provide them with materials and instruments, use various questioning strategies to elicit students' understanding, develop opportunities for students to learn in mini-scientific communities, and so forth (Harlen, 2004). The teaching approach is founded on the constructivist theory that all learners incrementally develop knowledge and understanding from their experiences and that shared knowledge is developed and clarified through interactions with others.

Traditional descriptions of the steps of inquiry-learning methods, however, do not adequately depict the full range of inquiry in scientists' practice; nor do they fully reflect inquiry as it is taught in FAST PD. According to FAST's conceptualization (Pottenger, 2005), inquiry takes the form of any of several modes or combinations thereof. Some of these modes are (a) simple curiosity, (b) replication inquiry (copying natural phenomena), (c) technological inquiry (inventing and engineering), (d) authoritative inquiry (drawing on existing knowledge to answer questions), (e) descriptive inquiry (in which domains are described and agreed upon), (f) explanatory deductive inquiry, (g) explanatory inductive inquiry, (h) explanatory experimental inquiry (the primary focus of the traditional approach to inquiry science, as described above), and (i) Socratic inquiry (in which teachers use questioning to "open vistas on explored ideas, build ideas to an intended end point, jog memory, and explore the students' knowledge status" [Pottenger, p. 19]). Students' simple curiosity is tickled across the breadth of science investigations. Replication inquiry occurs when a small group of students adopts another group's investigative methods. Technological inquiry occurs when students invent devices and develop solutions to carry out investigations. Authoritative inquiry happens when students inquire of authoritative people or documents. Descriptive inquiry happens when students describe the results of their data collection. Deductive inquiry occurs when students check their ideas against data or identify hypotheses for further inquiry. Inductive inquiry happens when students synthesize and generalize their results. Socratic inquiry occurs when teachers guide students through the other forms of inquiry.

Experienced inquiry science teachers using the FAST model employ variations of Socratic questioning strategies while they make use of the various forms of inquiry. Students in the midst of an experimental inquiry are led through the other forms of inquiry, as appropriate. As they help students develop hypotheses, design experiments, describe data, develop conclusions, and generate new hypotheses, teachers loop from one inquiry mode to another. Viewing and using inquiry in this manner captures the richness of practices that scientists use when addressing societal needs and problems with scientific methods.

Findings about the effects of teachers' use of questions vary among studies, but research in general has shown that teachers' proficient use of the appropriate questioning strategies improves student learning (e.g., Gall, 1970; Gall, 1984; Redfield & Rousseau, 1981; Samson, Sirykowski, Weinstein, & Walberg, 2001). Hamilton and Brady (1991, p. 253) stated,

> Research indicates that frequency of teacher questioning is a reliable, if not precise, predictor of student achievement (Brophy & Evertson, 1976; Coker, Lorentz, & Coker, 1980; Soar, 1973; Stallings & Kaskowitz, 1974; Weil & Murphy, 1982). That is, higher frequencies of teacher questions have been found to be related to higher levels of student achievement (as measured by either standardized achievement tests or course mastery tests). The link between frequent questions and increased levels of achievement can be explained by the high levels of student involvement (engagement) which occur in response to directed teacher actions; higher levels of student engagement have been linked repeatedly to higher levels of achievement (Morine-Dershimer, 1985; Pratton & Hales, 1986).

Overview of FAST

Description of the Program

FAST is an interdisciplinary middle-school science program consisting of three inquiry science courses entitled "The Local Environment," "Matter and Energy in the Biosphere," and "Change Over Time." The program emphasizes the foundational concepts and methods of the physical, biological, and earth sciences. The program is aligned with the National Science Education Standards (CRDG, 1996; Rogg & Kahle, 1997). It has been disseminated in 36 states and 10 countries, translated into Japanese, Russian, Slovak, and Hawaiian, and produced in Braille. Through 2001, over 7,000 teachers were trained in FAST. CRDG estimates that over 3 million students have taken one or more years of FAST.

FAST reflects several assumptions:

1) A science program reflecting the workings of the scientific disciplines gives students an authentic view of science and has a high probability of success.
2) Passively received knowledge transmitted by the teacher is an ineffective approach to learning. Learners incrementally develop knowledge and understanding from their experiences. Scientific findings are first learned in laboratory and later confirmed outside the classroom.
3) The program models scientific disciplines and the inquiry that happens within them. Shared knowledge is developed and clarified through interactions with others. Students learn that science proceeds through a process of constant reconstruction of explanation in the light of new findings.
4) Student learning should repeat the historical sequence of scientific discoveries and should draw upon the disciplines' way of developing knowledge.
5) Students should learn how to use scientific tools as they are needed in investigations.
6) The teacher should guide instruction through Socratic questioning methods.

FAST models the experience of practicing scientists, with students working in research teams to generate theoretical content of the program. Students are researchers who create hypotheses, do physical experiments, organize and analyze data, and develop a team consensus about conclusions. Students often work in small collaborative groups sharing data, ideas, and experiences; planning and executing experiments; and summarizing and drawing conclusions. The class identifies and clarifies generalizations following each investigation.

FAST teachers are "research directors," stimulating, facilitating, and probing students. The FAST research-team approach tolerates temporary student misconceptions because the contexts of investigation are carefully sequenced so that hypotheses and conclusions are constantly retested. Concepts are presented to the students in a carefully planned programmatic sequence of tasks and contexts of inquiry.

To achieve successful group generation of an acceptable conceptual structure, students spend 70-80% of their time in laboratory or field studies. These planned encounters allow time to define categories of events, generate hypotheses, test hypotheses, correct misconceptions, and ultimately, come to a consensus on the adequacy of explanations. The remainder of students' time is devoted to data analysis, small group or class discussions, literature research, and report writing.

FAST program developers knew that "the success of any new program rests heavily on the degree of understanding that the teacher has of its philosophy, objectives, and subject matter" (Hawai'i Science Curriculum Council [HSCC], 1967, p. 11). They conceived of FAST as a “completely articulated system reliant on the teacher training package. . .[that] will not be distributed piecemeal” (p. 26). Ultimately the FAST project team settled on a unique “participate in order to purchase” policy (Yamamoto, 1996) that required teachers to enroll in teacher institutes if they wished to use the program. This approach is supported by Berman and McLaughlin’s (1978) findings that financial resources were not a condition for a project’s growth and survival. Instead, professional development was key to putting an innovation into practice.

Initially, FAST developers were unable to find a publisher willing to support the required professional development component. Therefore, the CRDG developers decided to self-publish and distribute their own materials. This decision, combined with the need to develop a “format for effective supervision during the initial year of use” (Hawai‘i Science Curriculum Council, 1967, p. 65), led the project developers to create new structures for dissemination, professional development, and long term follow-up support as a comprehensive package.

The dissemination model that CRDG developed is made up of a number of components that work together to support successful, long-term change, including two-week pre-implementation institutes for teachers, regular structured follow-up during the first year of implementation, and varied forms of follow-up thereafter. The schedule for the FAST 1 two-week institute is shown in Table I-1. The FAST teacher institutes are designed to prepare participants to successfully teach the program by developing participants’ (a) knowledge of the program’s philosophy and objectives, (b) abilities to use a variety of instructional strategies, (c) understanding the physical, biological, and earth sciences content that is necessary to teach the course, and (d) excitement and enthusiasm for teaching science. Approximately half of the institute focuses on concepts and skills of science developed in FAST. The remaining time is devoted to teaching and developing effective instructional strategies. The institutes, which were recognized by the National Staff Development Council (1999) as an effective teacher learning program for improving student learning, focus on laboratory and field exercises that (a) students carry out during the year, (b) inquiry teaching methods and strategies, (c) classroom organization and management, and (d) methods for dealing with mathematics, reading deficiencies, and evaluation. FAST follow-up support services over the years have included administrative follow-up, responses to hot-line inquiries, telephone and teleconference consultations, a newsletter, written responses by U. S. mail or e-mail, locally certified trainers, a one-to-three year implementation support program, site visits, and renewal workshops.

In the FAST two-week face-to-face institute, participants conduct all FAST investigations for that one-year course of study. The institutes immerse participants in inquiry investigations that model the variety of teaching behaviors inherent in FAST and provide opportunities for

Table I-1
Schedule of Traditional Two-Week FAST Professional Development Institute

Day 1	Day 2	Day 3	Day 4	Day 5
		Morning		
•Registration •Introduction to FAST: What is inquiry •PS 1–4 lab safety	•FAST overview •Science standards student books •Ecology 4 & 5 •Oral/written reports	•PS 8 & 9 •Discussion on grouping •PS 10–12	•Ecology 29 & 30 •Field mapping •Ecology 26 animal care	•Discuss questioning strategies in FAST •PS 15–17
		Afternoon		
•Ecology 1 & 2 •Flow diagrams	•PS 5, 6, & 7 •CGs •Assign reading on grouping	•Teacher's guide, format & content •Evaluation; Evaluation Guide; PS evaluation 1 •Assign plant propagation	•PS 13–14 •Balloons in water submarine •Assign reading on questioning strategies	•Ecology 6, 7, & 8 soils •Ecology 17 weather station

Day 6	Day 7	Day 8	Day 9	Day 10
		Morning		
•The FAST instructional system•Ecology 31 & 32	•PS 23–28	•Ecology 18–25	•PS 34–40	•Relational study air pollution or water resource management
		Afternoon		
•PS 20–22	•Ecology 9–16	•PS 29–33 •Classroom organization & management	•Collect and analyze seed scarification data •Plant propagation reports	•Summary of FAST 1 •Planning for the academic year •Evaluation of institute

reflective discussions of the learning, teaching, and assessing experiences. Trained, certified instructors model successful teaching strategies while participants conduct the investigations that their students will undertake. Attention focuses on concept development and on the feelings of frustration and elation that accompany inquiry learning. After each of several series of investigations, the instructors and participants discuss instructional strategies (grouping, cooperative learning, inquiry questioning techniques, listening, paraphrasing, consensus building, ongoing assessment, and so forth). The instructors present techniques for handling problems with safety, unexpected data, alternative procedures, lack of science equipment, and students who have difficulties or special needs. The institute participants are recruited by district notices, CRDG mailings, the CRDG web site, and word of mouth.

Previous Research on FAST

FAST has been evaluated in several settings using a variety of designs and outcome measures. CRDG (2000) described the full body of these studies; the nine most methodologically sound studies are summarized here and are shown in greater detail in Appendix A. The studies are a sound, multimethod body of evidence of FAST's effectiveness.

Study 1: Tamir and Yamamoto (1977)

Using a comparison-group posttest-only design, Tamir and Yamamoto (1977) studied a sample of 614 high school biology students (spread across the four high-school grades), 36% of which had studied FAST for one year and 31% of which had studied it more than one year. Achievement was measured using final Grade-9 science grades and expected biology grades; interest was measured using questionnaire items on hobbies and intended college majors; and cognitive preference was measured using a 40-item questionnaire on preferences among four modes of attending to information in biology. Chi-square and analysis-of-variance findings on interest in science hobbies and class grades favored FAST students at statistically significant levels. Significant differences on intended college majors were not found. On the cognitive-preference instrument, FAST students showed significantly less preference for recall than non-FAST students.

Study 2: Young (1993)

Young (1993) reported a randomized posttest control-group study, with the pretest serving as the covariate. Of the 7th-grade students in a private school, 130 were randomly assigned to FAST classes and 123 were assigned to the control group. The pretest was the Comprehensive Test of Basic Skills (CTBS) science subtest; posttests included the pretest measure plus the verbal and figural batteries of the Torrance Tests of Creative Thinking and the Stanford Achievement Test science subtest. After adjusting the scores with the CTBS covariate, a statistically significant multivariate *F* value favoring FAST was found. Univariate analyses showed significant differences between FAST and non-FAST students on the CTBS and on the Torrance Verbal Originality and Figural Elaboration subtests. These results were interpreted to show that statistically significant differences were not found on the Stanford scores because the instrument measures recall, which FAST does not emphasize, whereas the CTBS scores showed significance because the instrument measures some higher-level cognitive processes, which FAST does emphasize. Results on the Torrance subtests were speculatively interpreted to show the effects of FAST on observing phenomena in detail and actively processing information.

Study 3: Young (1982)

Young (1982) reported a posttest-only comparison-group study. Random samples of 25 FAST classes and 25 non-FAST classes in Hawai‘i public schools, stratified by student ability level and school socioeconomic level, were selected. Compared with the sampled FAST classes, students in about half the non-FAST classes showed superior reading and mathematics achievement. The study used the Laboratory Skills Test (LST), a measure of student performance on laboratory tasks (six items), science-process skills (e.g., observing, predicting, and providing evidence) (four items), and knowledge and understanding of science (four items), with a total-instrument reliability (KR_{20}) of .93 and content validity established by expert review (Young, 1982); and the Comprehensive Test of Basic Skills science subtest. Analyses (*t*-tests) showing statistically significant results favoring FAST were found for both instruments.

Study 4: CRDG (2000)

CRDG (2000) reported a 1988 posttest-only comparison-group study with a covariate. The sample included intact groups of 6th-graders (FAST *N* = 38 and non-FAST *N* = 47) and 7th-graders (FAST *N* = 58 and non-FAST *N* = 83). Three instruments were used, including (a) the LST, with results examined on the three subscales, (b) the Performance of Process Skills (POPS) Test, a 21-item instrument measuring science process skills; and (c) the Fukuoka, Ishikawa, and Nakayama (FIN) test, also a measure of science process skills. Results on the California Achievement Test (CAT) total battery were collected for use as the covariate in the statistical analyses. A multivariate analysis of variance (MANOVA) was conducted to examine the significance of the tests overall, followed by univariate analyses of covariance (ANCOVA). The MANOVA showed a statistically significant overall *F* value in favor of FAST. In the subsequent ANCOVAs, favorable statistically significant differences were found on the three LST subtests for Grade 6; for Grade 7, favorable statistically significant differences were found on the LST laboratory skills and process skills subtests but not on the third LST subtest. No significant differences were found on the POPS or FIN tests.

Study 5: CRDG (2000)

CRDG (2000) reported two annual posttest-only studies in which a small group of FAST schools and a small group of non-FAST schools in a California district were compared with state standards. The FAST sample included 472 students in six schools in the first year and 254 in three schools in the second; the non-FAST sample included 360 students in four schools in the first year and 366 in three schools in the second. Data were collected with a statewide student science assessment that was briefly used in California in the late 1980s. In *t*-tests, the means of FAST schools in a district were compared with “expected high scaled scores” for the schools; the means for the non-FAST schools in the district were compared likewise. (The scores that the state expected for the schools varied among the schools because of school demographics.) In the first of the two years, the results for the FAST schools were greater than the expected school results at a statistically-significant level, but no differences were found for the non-FAST schools. In the second of the two years, FAST schools' results were not significantly different from the expected scores, but the non-FAST schools' results were significantly *below* the expected scores ($p < .01$).

Study 6: Pauls, Young, and Lapitkova (1999)

Pauls, Young, and Lapitkova (1999) reported a posttest-only study with a "comparison-group" consisting of all students, including FAST students. The sample included 333 students (ages 13–14) in FAST schools in Slovakia, with a "comparison group" of 7,524 students in 145 Slovakian schools. The results on a total of 14 items from the 1995 Third International Mathematics and Science Study, each addressing one of four science topics, were examined. Comparisons of percentages correct were made between FAST students and Slovakian students. FAST students outscored the group of all Slovakian students on all four topics. Only item percentages were reported, with no aggregation. For this report, means of item percentages correct for each of the four areas were calculated, and these means in turn were averaged. The results showed overall means (rounded to the whole number) of 48% for all students and 77% for FAST students (with standard errors of percentages, similarly averaged, of .56 and 2.28, respectively).

Study 7: Dekkers (1978)

Dekkers (1978) reported a posttest-only comparison of 101 FAST high-school students with 88 students instructed in the Australia Science Education Project (ASEP). Instruments included the Science Cognitive Preference Inventory developed by Tamir (see the description of Study 1 above) and the 30-item Activity Preference in Science, which judges preference for reading activities, designing experiments, or discussion activities (test-retest correlation = .62). Analysis of variance results on cognitive preferences showed FAST students with significantly higher scores on preferences for recall, with the highest scores for both programs on questioning. The results on science activities showed that FAST students' scores were significantly higher than ASEP students in laboratory work, discussion, and field work but not in reading or project work.

Study 8: CRDG (2000)

CRDG (2000) reported a post-test only study for each of three years (1995, 1996, and 1998). Results on a criterion-referenced, standards-based annual statewide test were reported on all 10th-graders (*N* unreported) in a Connecticut district in which students had been instructed with FAST during their middle school years. Mean scores were compared with the "state goal" for each of three years. Scores from the FAST district were greater than the state goal in each of the three years. For two of the three years, the percentages of the district's students that were above the state goal were 52 and 61, respectively; the percentages for all students statewide during these two years were 32 and 34. (The percentages for the third year were unavailable.) In one of the three years, scores in science (compared with the results for tests given in four other subjects) were the only ones that improved.

Study 9: CRDG (2000)

CRDG (2000) reported a pre/posttest study of 45 6th-graders and 45 7th-graders randomly selected from students classified as average or above-average on the Comprehensive Test of Basic Skills science subtest and by teacher judgment. Analyses (*t*-tests) were conducted between FAST student posttest means and the expected means, as specified in norms tables. Significant differences favoring FAST were found for both grades.

External Reviews of FAST

Other evidence of FAST's success is the body of reviews by various organizations. To the best of our knowledge, FAST has received more national recognitions than any other middle school science program. After extensive review of the program's materials, program evaluation design and methodology, and program effects over multiple years in multiple sites, the U. S. Department of Education's Expert Panel on Mathematics and Science Education (2001) designated FAST as one of only two exemplary science programs; furthermore, it is the only one so widely implemented. After extensive review of FAST's findings on the effects of its professional development component, the National Staff Development Council (1999) called it an effective teacher professional development program for improving student learning. The Building Science and Engineering Talent (2004) project rated FAST as "notable." *Best Practices From America's Middle Schools* (Watson, 1999) described FAST as a successful instructional program. In the *Catalog of School Reform Models* (Northwest Regional Educational Laboratory, 1998), it was described as one of three research-based effective science reform models. OERI (1994) designated it as a science program that works, and the Program Effectiveness Panel of the U.S. Department of Education's National Diffusion Network (OERI, 1990) designated it as an exemplary program. Pauls, Young and Lapitkova (1999) stated that FAST was the only program known to them that showed promise of significant impact in data collected in the Third International Mathematics and Science Study.

THE THEORETICAL MODEL UNDERLYING THE INSTRUMENT DEVELOPMENT COMPONENT OF THE PROJECT

The Phase-II study for which we prepared in this project is based on a theoretical model of the relationship among teachers, program implementation, student achievement, and the school and community context within which the program is presented. Successful implementation is affected by the characteristics, participation, and support of teachers and administrators and by the organizational and socio-political context within which the school exists.

The model, which guided our instrument development and validation in this project, is pictured in Figure I-1. It shows between-school and within-school effects. Within schools, it

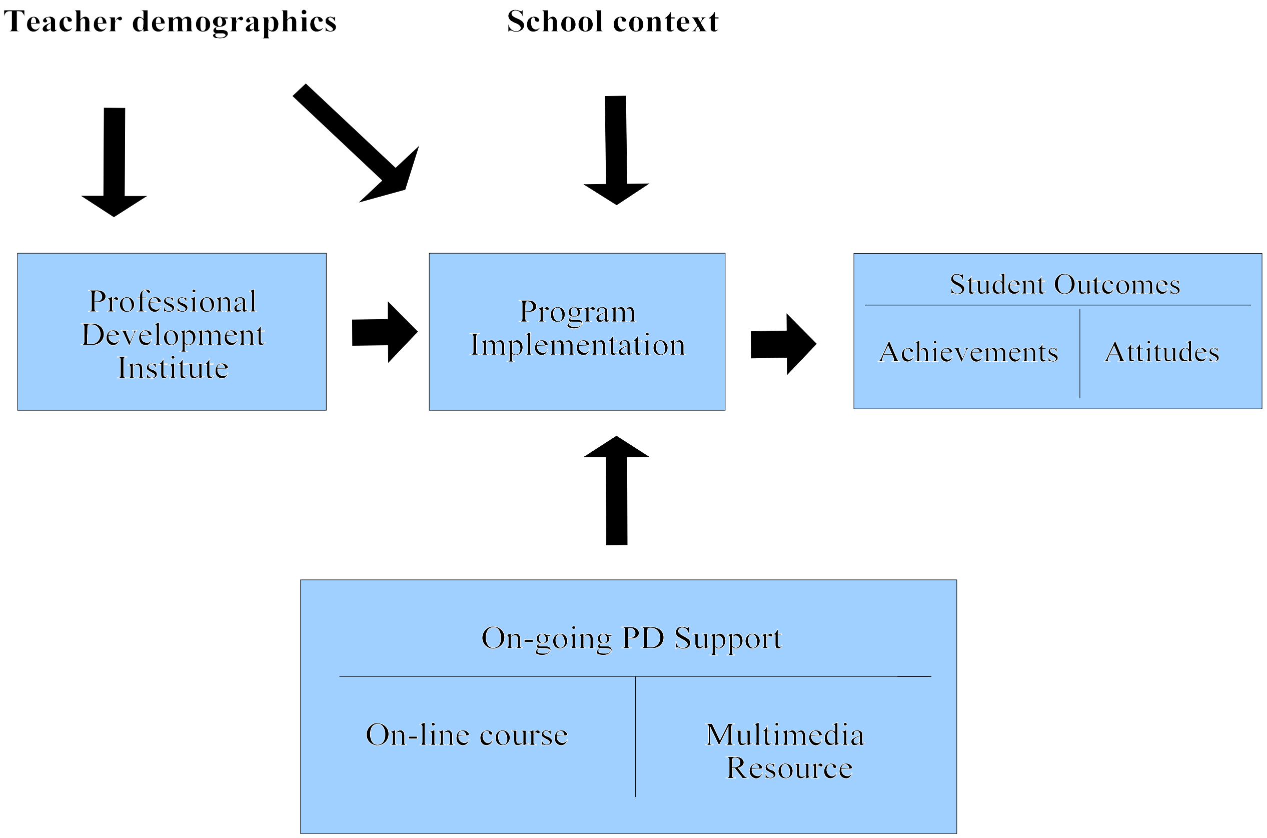


Figure I-1. Theoretical model of the effects of professional development on program implementation and outcomes.

suggests the direct effects of teachers on program implementation level. Between schools, it suggests these direct effects and also suggests the indirect effects of the context (i.e., school, district, and community) on implementation level. At the school level, teachers' participation in planning, decision making, professional development activities, and so forth affect implementation, while at the individual level (i.e., within their classes), teaching practices and fidelity to program content and methods affect implementation. In turn, implementation levels affect student learning levels.

This simple model cannot do justice to the complexity of program implementation, of course. It does not accurately reflect the myriad interactions occurring among people within and across organizations (Datnow & Stringfield, 2000). Nor does it reflect the strong links between professional learning communities, teacher learning, and student performance (Newmann & Wehlage, 1995). For example, our analyses might show that some characteristics of individuals are best analyzed as organizational characteristics (Brandon & Heck, 1998; Ogawa & Bossert, 1995).

With the exception of student achievement, we describe the classes of variables addressed in the model in this section.

Program Implementation

In the United States, evaluators began attending to program implementation a few years after the federal government funded large-scale, post-Sputnik curriculum reforms in the 1950s and Great Society social and educational programs in the 1960s. The earliest studies, which used "black-box designs" that assumed that programs were uniformly and fully implemented, had not focused on implementation, but this changed when many of these studies failed to show effects. Researchers and evaluators began to assess implementation because the lack of effects shown in previous studies might have been due to poor implementation. By the mid-1970s, theorists, researchers, and evaluators began to publish meta-studies of implementation: For example, Hall and Loucks (1978) studied the implementation of educational innovations, Fullan and Pomfret (1977) reviewed evaluations examining implementation, and Patton (1978), among others, described how to evaluate implementation. Many of the implementation studies in these early days were about educational programs, as Scheirer and Rezmovic (1983) showed in their literature review. In the late 1980s and the 1990s, the focus of reviews of studies of program implementation published in refereed journals shifted to social programs, such as public or mental health programs (e.g., Bond, Evans, Salyers, Williams, & Kim, 2000). Educational researchers and evaluators began to contribute more on the topic again after 2000, including some researchers and evaluators funded by the NSF Interagency Educational Research Initiative and other NSF grants (Lynch, 2007; Lynch & O'Donnell, 2005; O'Donnell, 2007; Ruiz-Primo, 2005).

Our description of the theory of program implementation draws primarily from the publications of the last 20 years, as well as from other pertinent program evaluation literature that usually has been ignored in the implementation literature. The description addresses two questions: What are the purposes of studying program implementation? What aspects of implementation should be examined? These are answered in the following two subsections.

What Are the Purposes of Studying Program Implementation?

Several purposes of studying program implementation—or *fidelity of implementation*, as it commonly is called—are commonly described in the implementation literature. The first of these purposes has to do with good research design: to collect data about a key variable—the extent to which a program is implemented—in a causal chain ending in program effects. This is necessary to inform evaluation conclusions, particularly when studies show programs to have none of the intended effects. The term *fidelity* implies that the purpose of measuring implementation is to examine how close program implementation is to the ideal. However, implementation need not be measured relative to an absolute standard. Findings on measures of implementation can be useful in causal studies irrespective of whether they are compared to the ideal. Patterns of implementation can be tied to varying levels of outcomes (Ruiz-Primo, 2005). Second, the findings of studies can inform education and learning theory. Theorists can learn about the effectiveness of their conceptions about education, the degree to which programs are implemented as intended, whether adapted programs are more successful than faithfully implemented programs, and so forth. Third, implementation findings can help program developers revise their materials, procedures, staffing, and so forth, and they can inform program administrators about the extent to which programs are being delivered as intended. This is a formative purpose of studying implementation. The fourth purpose is summative in nature: to inform program funding organizations about how well their money is being spent. A fifth purpose—not often found in most of the literature on program implementation—is to conduct evaluability assessments (e.g., Wholey, 1994). The study of the level of implementation of a program can inform evaluators whether the program is ready for a summative study. The sixth purpose, related to the fifth, is to examine the feasibility of interventions (Dusenbury, Brannigan, Falco, & Hansen, 2003).

What Aspects of Implementation Should Be Examined?

An issue often discussed in the literature on studying program implementation is the choice of aspects, features, characteristics, or dimensions of programs that should be examined. Several authors have stated that implementation studies should examine program *components*. Ruiz-Primo (2005, p. 6) cited Hord, Rutherford, Huling-Austin, and Hall (1987), who defined components as the "major operational features or parts of the program." Blakely et al. (1987, p. 260) defined a component as "an observable activity, material, or facility" and "logically discrete from other components." Gresham (1989) called for defining them in behavioral terms. Components can be of "global, intermediate, or molecular" specificity (Gresham, 1989, p. 40).

The approach we have taken in our study has been to collect data on a variety of program aspects, mostly at a fine-grained degree of specificity. As we describe later in some detail, we collected data on aspects across the *breadth* of the program, some in greater *depth* than others. Consistent with the program evaluation literature (e.g., Bickman, 1985; Brandon, 1993; Scriven, 1991), which defines program components as broad "spatio-temporal separate regions" of programs (Scriven 1991, p. 43) or sets of "related activities directed toward reaching some common objective" (Bickman, 1985, p. 192), the aspects that we examine are not *components*. A curriculum might include, say, a professional development component and a teaching component. We examine activities, strategies, materials, quantity, and duration (Moncher & Prinz, 1991), and so forth within components such as these. However, summarizing the findings about these features of programs at the component level is not useful for our theoretical model of implementation. Because we do not examine implementation at a global level, we therefore do not refer elsewhere in this paper to components in the description of our study of program implementation.

It has been recommended in the implementation literature that studies of implementation are incomplete without considering program context (e.g., Ruiz-Primo, 2005). Considering covariates, mediating variables, and moderator variables is a necessary design feature of many studies, of course. We consider context variables to be moderators or predictors of implementation but not, strictly speaking, aspects of implementation. Therefore, in our model, contextual variables are outside the boundaries of program implementation.

To our knowledge, two schema for characterizing the study of program implementation levels have been discussed in the recent educational and social science research and evaluation literature. The first includes five aspects of implementation: *adherence*, *exposure* (sometimes called *dose*), *quality of delivery*, *participant responsiveness*, and *program differentiation* (Dane & Schneider, 1998; Dusenbury et al., 2003). Adherence is the extent to which program implementation follows the prescribed sequence, procedures, lessons, steps, and so forth; exposure is the number of procedures, lessons, or steps that are implemented and their duration; quality is the implementation skill and knowledge shown by the service deliverer; participant responsiveness is “a measure of participant response to program sessions, which may include indicators such as levels of participation and enthusiasm” (Dane & Schneider, 1998, p. 45); and program differentiation is the extent to which the program is delivered in a manner that differentiates it from other inventions, particularly those to which it is compared, and the extent to which it avoids treatment drift. Implementation researchers developed this taxonomy of aspects by reviewing the literature on the implementation of prevention research programs. The choice of terminology, particularly exposure (dose) and program differentiation, shows the influence of these programs and the designs (experimental or quasi-experimental) often used in studies of the programs. Of the five aspects, we have not addressed program differentiation.

The second schema for characterizing the study of program implementation levels includes aspects of implementation in the general categories of *structure* (i.e., the program framework) and *process* (Mowbray, Holter, Teague, & Bybee, 2003). The aspects include length, intensity, and duration of service delivery; content, procedures, and activities; staff roles, qualifications, and activities; and “inclusion/exclusion characteristics for the target service population” (Mowbray et al., 2003, p. 315). The overlap with the first schema is apparent. We consider staff qualifications to be contextual, and we do not address the final criterion because it does not address studies of intact classrooms.

Teacher Variables

Teacher characteristics and the manner in which teachers participate in program implementation are central to successful program implementation (Smylie, 1992). Teacher characteristics affecting this success have been shown (among many others) to include age (Huberman, 1989), gender (Datnow, 1998), the number of years the teachers have taught at their current schools and the total number of years they have taught (Heck, Brandon, & Wang, 2001), the stage of the teachers' careers (Kirby, Berends, & Naftel, 2001), previous teaching experiences (Tyack & Cuban, 1995), subject-matter expertise (Brandon & Heck, 1998), familiarity with and attitudes toward the program (Evans, 1986; Roberts-Gray, 1985), and qualifications and motivation for implementing the program (Evans, 1986; Pinto & Prescott, 1990; Roberts-Gray, 1985). Teachers who have previously participated in several unsuccessful implementations are unlikely to support implementing new programs. Forms of formal teacher participation (Firestone & Corbett, 1988; Heck, 1993; Pinto & Prescott, 1990) that affect the success of interventions include collaboration among faculty (Heck & Brandon, 1995, Little, 1981; Pinto & Prescott, 1990) and participation in decision making about implementation (Heck, Brandon, & Wang, 2001). As we describe in Chapter III, we identified additional teacher variables when developing our instruments.

School and Community Variables

Considerable research has shown that school and community characteristics affect the likelihood of program success. Some of these school characteristics include school size (Lee & Smith, 1997), the number of interventions implemented ("innovation overload") (Kirby et al., 2001), and the number of years the intervention has been implemented (Berends, Kirby, Naftel, & McKelvy, 2001). As described in Chapter III, we identified additional school characteristics when reviewing the literature while developing instrumentation in this project. Community characteristics affecting program implementation include parent support (Fullan, 2001) and student prior achievement (Coleman et al., 1966). Community socio-economic status can affect students' aspirations and the availability of support for teachers seeking to engage in collaborative relationships (Berends et al., 2001). Community support can enhance the degree to which a novel program, such as an inquiry-based program that replaces a traditional content-based program, will thrive; active community opposition can mean the death of an intervention.

CHAPTER II
DEVELOPMENT AND TRIAL OF FASTPRO, THE ALTERNATIVE VERSION OF FAST PROFESSIONAL DEVELOPMENT[2]

The purpose of the SCUP project was to prepare for a randomized experiment comparing the effects of two versions of K–12 inquiry science PD. One of the versions is the traditional FAST model, consisting, as described in Chapter I, of a two-week in-person institute for each of the three years of the program, with follow-up support in the form of e-mail, telephone, and newsletters. The alternative version, called FASTPro, was developed in this project. It consists of a one-week face-to-face institute (FASTStart), an electronic resource in the form of a multimedia DVD for the trained teachers (FASTeR), and an on-line course (FASTForward). In this chapter, we describe FASTPro and its development, beginning with a discussion of how it differs from traditional FAST PD in the extent to which it addresses the literature and the NSES PD standards for teaching inquiry science.

COMPARISON OF TRADITIONAL FAST PD AND FASTPRO

As described in Chapter I, the traditional FAST PD model reflects many of the characteristics of effective PD that have been identified in research. Traditional FAST PD is inquiry-based training that provides demonstration, feedback, science content, and the opportunity for extensive practice. It provides some long-term post-institute support, but the support is less sustained and conducted less in conjunction with teachers' ongoing classroom duties than is suggested in the literature (Supovitz, 2001; Supovitz, Mayer, & Kahle, 2000; Zigarmi, Betz, & Jennings, 1977). Furthermore, over the long term, the traditional version of FAST PD does not regularly provide teachers with the suggested degree of feedback about how to implement the program at the various teachers' sites (Guskey, 1995; Little 1993; Loucks-Horsley, Hewson, Love, & Stiles, 1998; McLaughlin & Marsh, 1990).

In Table II-1, we highlight how the alternative version addresses some of these deficiencies. The table shows the extent to which the traditional and the alternative versions of FAST PD address several of the features of effective PD that are discussed in the literature. Traditional FAST PD addresses 8 of the 10 features at least to some extent, and, as a group, the three components of FASTPro address all the features. FASTPro particularly increases the capacity of FAST PD to (a) allow for participation in follow-up during the school year, (b) consider the teachers' educational context when providing guidance to them, (c) coach the teachers, and (d) give them regular feedback and long-term support, coupled with pressure to implement inquiry science fully and well.

In Table II-2, we show how both the traditional version of FAST PD and the alternative

[2] This chapter draws heavily from Gray, Nguyen and Speitel (2005) and Nguyen, Speitel, and Gray (2007).

Table II-1
Professional Development Features That Are Identified in the Literature and How They Are Addressed in Traditional FAST Professional Development and in FASTPro

Feature of FAST professional development	Traditional FAST	FASTPro		
		FASTStart	FASTeR	FAST Forward
1. Provides discussion and reflection of the *FAST* content, philosophy, and instructional and assessment strategies.	Yes	Yes	Yes	Yes
2. Involves participants in the activities that students will experience and has them work through the sequence of program activities in the role of the students that they will teach.	Yes	Some	Some	No
3. Provides background to, and application of, the concepts and skills of integrated science.	Yes	Some	Yes	Yes
4. Includes discussion after each series of activities of the instructional strategies used (e.g., grouping techniques, collaborative and cooperative learning, questioning strategies, curriculum integration, and evaluation) and addresses classroom issues such as reading, writing. and mathematical difficulties; unexpected outcomes; safety; and lack of equipment.	Yes	Some	Yes	Yes
5. Models inquiry teaching.	Yes	Yes	Yes	No
6. Engages teachers in specific teaching tasks and gives them opportunity for practice.	Yes	Some	Yes	Yes
7. Allows for participation in follow-up during the school year to reflect and share class experiences and review activities, philosophy, and instructional strategies.	No	No	Yes	Yes
8. Considers the teachers' educational context.	Some	No	No	Yes
9. Provides coaching.	Some	No	No	Yes
10. Provides teachers with regular feedback and long-term support, coupled with pressure.	No	No	No	Yes

version address the NSES inquiry science PD standards. The primary differences between the two versions reflect the comparison between traditional FAST and the characteristics of PD that are suggested in the PD literature. The differences are that

1) traditional FAST PD teaches all the student investigations in a two-week institute, whereas FASTPro teaches about half the investigations in a one-week institute.
2) the FASTForward component of the alternative version allows learning to be applied to teachers' contexts while they are still being guided by PD trainers.
3) FASTForward provides the teachers with guidance while they lead their students through the

Table II-2

How Traditional FAST Professional Development (PD) and FastPro Address the National Science Education Standards (NSES) for Professional Development (PD) in Science Inquiry, Grades 5–8

NSES standards for PD in inquiry	Traditional FAST PD	FASTPro
A1. Involve teachers in actively investigating phenomena that can be studied scientifically, interpreting results, and making sense of findings consistent with currently accepted scientific understanding.	Teachers experience investigations as students, developing research questions and collecting and analyzing data. A scientific community is established in which ideas, data and findings are reported, discussed and analyzed.	Teachers experience half of the investigations as students, developing research questions and collecting data and analyzing data. A community is established in FASTStart and extended and in FASTForward.
A2. Address issues, events, problems, or topics significant in science and of interest to participants.	Teachers investigate relevant problems, issues and questions and apply content to ecological issues and to science technology and societal issues. Connections between science and other disciplines are discovered and studied.	Teachers become engaged and investigate relevant problems, issues and questions, make connections with the disciplines through the FASTForwar's assignments and threaded discussions. They apply science to their local environments through sharing descriptions, photos, and videos.
A3. Introduce teachers to scientific literature, media, and technological resources that expand their science knowledge and their ability to access further knowledge.	Teachers choose an investigation; design, plan, and conduct it, and share the results during training. Teachers use the library, Web, and other resources to conduct their research.	Participants select a research problem, design an investigation on paper, discuss it with their institute, examine other research in the literature during FASTForward, and conduct the investigation.
A4. Build on teacher's current science understanding, ability and attitudes.	The instructor models scientific teaching which teachers later practice. The instructor gets to know the teachers and employs various teaching techniques as appropriate to engage a variety of learners.	Teachers experience inquiry, then practice it in the classroom and reflect with FASTeR and FASTForward. The determines teachers' technological abilities and address individual needs during FASTeR and FASTForward.
A5. Incorporate ongoing reflection on the process and outcomes of understanding science through inquiry.	Working at a fast pace, teachers keep a journal and discuss the investigations. They have additional modeling, coaching and practice during the second week.	Teachers use FASTeR at own pace, and participate in FASTForward for further reflection. Teachers keep electronic journals and share them in threaded discussions.
A6. Encourage and support teachers in efforts to collaborate.	Teachers work together, learning, investigating, questioning, planning, with no formal mechanism for continued collaboration.	Collaboration that began in FASTStart continues in FASTForward.

FAST program’s student investigations.

4) FASTeR provides the teachers in the alternative version with a rich electronic resource of text, video, photographs, animation, and audio to consult while they are teaching students.
5) the community that is built during the institute is continued during FAST Forward.

DEVELOPMENT OF FASTPRO

During the project, a schedule for one-week FASTStart institutes was developed, many of the tasks in developing FASTeR were conducted, and a preliminary version of FASTForward was prepared. FASTeR was the most resource-intensive component of the development of FASTPro, and it is the primary focus of this chapter. FASTStart and FASTForward were piloted in a project that was not part of the proposed and planned NSF IERI grant. The development and implementation of FASTForward and the results of the pilot test of FASTStart and FASTForward will be reported in a forthcoming dissertation.

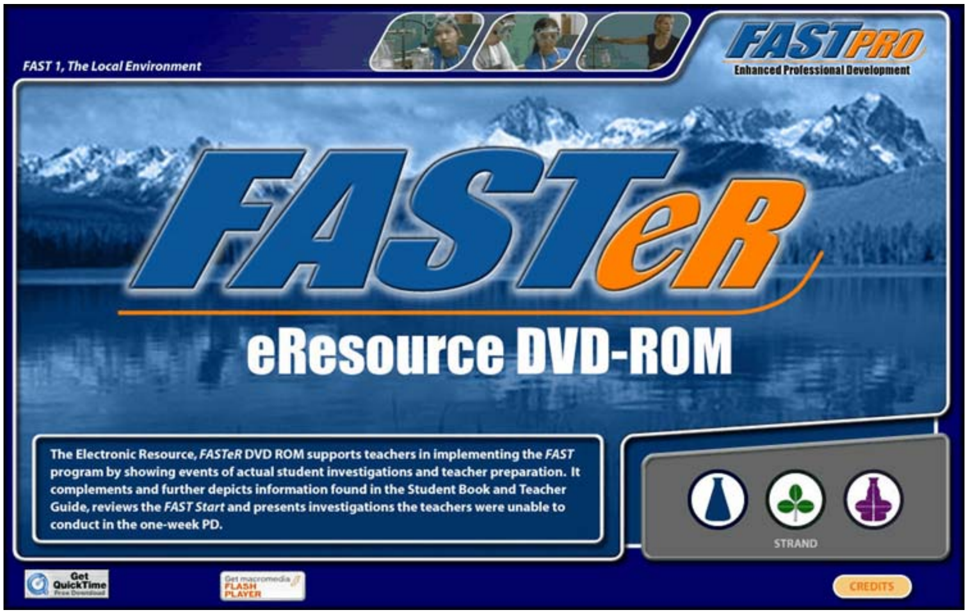
Development of FASTStart

The first step in developing FASTPro was to prepare a schedule for the one-week institute to replace the two-week institute that is shown in Table I-1. The revised schedule had to attend to the skills and concepts that were essential for teachers to learn; the PD could not simply be cut in half. The FAST program developers and an expert FAST trainer/teacher prepared and iteratively reviewed and revised the schedule. It is shown in Table II-3.

Table II-3

Schedule for FastStart, the One-Week FAST Professional Development Institute

Day 1	Day 2	Day 3	Day 4	Day 5
		Morning		
•Introduction to FAST: What is inquiry? •PS 1–4	•FAST Overview •Science standards student books •Ecology 27–29 •Ecology 3–5	•PS 8 & 9 •Brief discussion grouping •PS 10– 12	•Ecology 30 •Ecology 26 (talk) •Ecology 6–8	•Discuss questioning strategies •PS 15–17
		Afternoon		
•Ecology 1 & 2 •Flow diagrams •Homework: Instructional Guide 1–35	•PS 5–7 •Enhancement web/media training •Homework: Grouping; Inst. Guide 36–56	•Teacher's Guide •Evaluation in FAST	•PS 13–14 •Homework: Questioning strategies; Inst. Guide 56–59	•Relational Study •Analyze Ecology 2 •Submarines •Plans for follow-up


FAST 1, The Local Environment
FASTPRO
Enhanced Professional Development
FASTeR
eResource DVD-ROM
STRAND
Get QuickTime
Get macromedia FLASH PLAYER
CREDITS

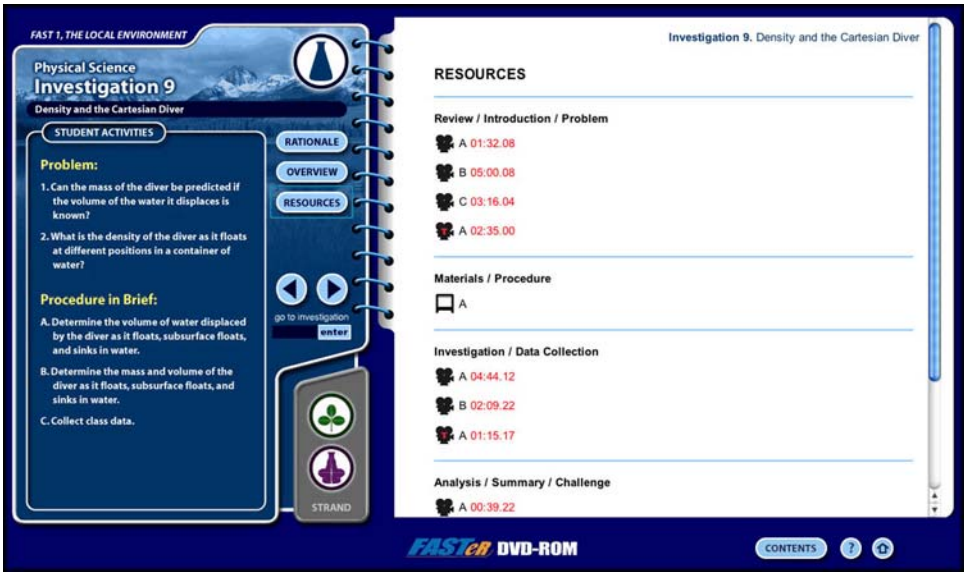
The Electronic Resource, *FASTeR* DVD ROM supports teachers in implementing the *FAST* program by showing events of actual student investigations and teacher preparation. It complements and further depicts information found in the Student Book and Teacher Guide, reviews the *FAST Start* and presents investigations the teachers were unable to conduct in the one-week PD.

Figure II-1. The opening screen of FASTeR.

Development of FASTeR

Overview of FASTeR

FASTeR is a multimedia resource on a digital video disc (DVD-ROM) for use on a computer. The opening page of FASTeR is shown as Figure II-1, a sample page is shown as Figure II-2, and the beginning pages of a slide show and two video examples are shown in Figure II-3. The DVD-ROM includes video from FAST institutes, video showing FAST being taught in the



Problem:

1. Can the mass of the diver be predicted if the volume of the water it displaces is known?
2. What is the density of the diver as it floats at different positions in a container of water?

Procedure in Brief:

A. Determine the volume of water displaced by the diver as it floats, subsurface floats, and sinks in water.

B. Determine the mass and volume of the diver as it floats, subsurface floats, and sinks in water.

C. Collect class data.

Investigation 9. Density and the Cartesian Diver

RESOURCES

Review / Introduction / Problem

A 01:32.08

B 05:00.08

C 03:16.04

A 02:35.00

Materials / Procedure

□ A

Investigation / Data Collection

A 04:44.12

B 02:09.22

A 01:15.17

Analysis / Summary / Challenge

A 00:39.22

Figure II-2. A page of FASTeR.

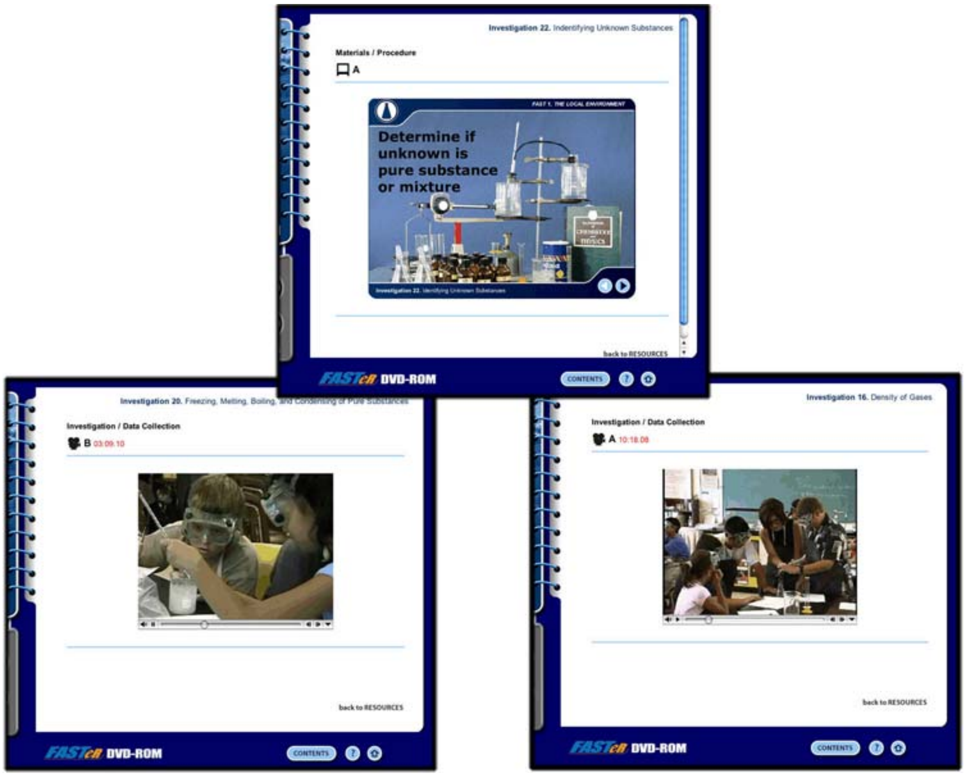

Investigation 22. Indentifying Unknown Substances
Materials / Procedure
A
FAST 1. THE LOCAL ENVIRONMENT
Determine if unknown is pure substance or mixture
Investigation 22. Identifying Unknown Substances
back to RESOURCES
FASTeR DVD-ROM
CONTENTS
Investigation 20. Freezing, Melting, Boiling, and Condensing of Pure Substances
Investigation / Data Collection
B 03:09:10
back to RESOURCES
FASTeR DVD-ROM
CONTENTS
Investigation 16. Density of Gases
Investigation / Data Collection
A 10:18:08
back to RESOURCES
FASTeR DVD-ROM
CONTENTS

Figure II-3. The beginning page for a FASTeR slide show and the beginning pages for two video examples.

classroom, video of teachers in the role of students at the FAST institutes, photographs and animation (in the form of slide shows) of setting up student investigations, and photographs of laboratory materials and equipment. (See Figures II-2 and II-3.) Each type of PD aid (video segment, animation, and so forth) is categorized according to the phases of FAST I investigations that are shown in the student book, including the introduction to the investigation, the procedures followed in the investigation, the data collection and analysis phase, and the summary phase. The investigations included are shown in Table II-4.

Developers

Intensive development of FASTeR occurred over about a two-year period. Development was overseen by the head of CRDG's Learning Technology Section, who supervised an instructional designer, two videographers, a professional video editor, and a graphic designer. The instructional designer and graphic designer also served as multimedia programmers. Student assistants also helped with some of the videotaping. (Further development and refinement is continuing, despite the end of the grant, as resources permit.) A FAST trainer/teacher served as the content expert, with the assistance, when necessary, of the two primary FAST program developers. A substantial portion of the development resources were donated by CRDG, because FASTeR eventually will have value to the organization beyond its use in studies of FAST PD.

Development Goals

The development team established several goals for FASTeR. First, it was to be made available on a DVD-ROM and not the Internet because of uneven availability of high-speed broadband Internet connections across the country. The DVD was to include no more than 20 hours of compressed video. The quality of the multimedia was chosen because it allowed for small windows on the computer screen and thereby conserved space on the disk. The second goal was to provide several types of media. Slide shows were deemed appropriate for step-by-step instructions. Movies in particular were deemed to be helpful, as shown by the results of previous projects (e.g., Barab, MaKinster, Moore, Cunningham, 2001; Callahan & Switzer, 1999; Speitel & Nguyen, 2001; WGBH Educational Foundation, 2005). The movies were sufficiently long to present the desired procedure, pedagogical method, content, and so forth; each movie was to correspond to a FAST-investigation phase. The third goal was for FASTeR to be designed as a World Wide Web interface and navigation scheme and catalogued by investigation name and number. Fourth, FASTeR was to provide the rationale for each FAST student investigation that was included on the DVD-ROM, as well as an overview of the investigation and the problem to be addressed in it. This was to allow the teacher to quickly review the essential features of the units without referring to the lengthy printed teachers' guide. Fifth, it was to show teachers using the program in the classroom. Teachers trained in traditional FAST had no opportunity to see it in action in the classroom during training. FASTeR was designed to provide numerous such examples. Finally, it was to cover the investigations that are addressed in the traditional two-week FAST institute but not the one-week institute. This goal has yet to be achieved.

Technical Characteristics

The DVD-ROM was prepared using Hypertext Markup Language, a programming language of the World Wide Web. The videos segments, ranging in duration from 1¼ to 14 minutes, are presented with Quicktime software. As seen in Table II-4, of the 22 student investigations addressed on FASTeR, 18 have videos from teacher institutes and 9 have videos of students in the classroom. (The number of investigations showing classroom footage is to be increased in future development.) They can be stopped, started, and rewound at will. Animations, and still photographs (*N* investigations = 12) are shown with FLASH software. Teachers advance them manually.

For more information about the technical characteristics and development of FASTeR, see Nguyen, Speitel, and Gray (2007).

Table II-4

Numbers of Videos and Slide Shows Developed for FASTeR Student Investigations, by Investigation Phase (Introduction, Procedures, Data Collection and Analysis, and Summary)

Investigation	Teacher institute video				Classroom video				Slide show
	Intro	Proc	Data	Sum.	Intro	Proc	Data	Sum.	
Physical Science:									
1. Liquids and Vials	1		1	1					2
2. Sinking a Straw	1	1	1	1					
3. Graphing the Sinking Straw Data	1	1		1					
4. Mass and the Sinking Straw	1	1	1	1					1
5. Sinking Cartons	1		1	1					
6. Volume and the Sinking Cartons	1	1	1	1					
7. Floating and Sinking Objects			1		1	1		3	1
8. Introduction to the Cartesian Diver	1		1		1		1		
9. Density and the Cartesian Diver	1		1	1	3		2	2	1
10. Density of Objects									
11. Density of Liquids									
12. Buoyancy of Liquids						1	2	1	
13. Balloons in Water									1
14. Submarine Project				1					
15. Bubbles in Gas					2		2	2	1
16. Density of Gases						2	2		1
17. Weather Balloon Project									
18. Boiling Water									1
19. Heating Ice in a Balloon		1	1	1	1	1	1	4	1
20. Freezing, Melting, Boiling, and Condensing of Pure Substances					2	1	5	3	1
21. Freezing, Melting, Boiling, and Condensing of Mixtures	1	1	2	1					1
22. Identifying Unknown Substances					1		1	1	1
Ecology:									
1. Seeds with Hard Coats	1		1	1					
2. Scarifying Seeds	1	2	1	2					
3. Propagating Plants	1	1		2					
4. Oral Scientific Reports	1								
5. Written Scientific Reports	1								

FASTeR Evaluation Results

In an effort to collect data on the usefulness of FASTeR, two of the FASTeR developers conducted an e-mail questionnaire survey. Of the questionnaire items, four were Likert-scale questions about the usefulness of FASTeR. The instrument was distributed to 214 FAST teachers (55 by e-mail and 159 by U.S. mail) who were identified during the instrument development that is described in Chapter III. Each teacher who completed a questionnaire was mailed a $10 bookstore gift certificate. Of the 214 teachers, 25 (11.7%) responded after as many as four reminders by e-mail and one by postal mail. This response rate, comparable to the rate we experienced with our teacher questionnaire, as reported in Chapter III, is typical for questionnaire distribution of this nature.

The teachers' responses are shown in Table II-5. As seen in the table, the responses were uniformly favorable about the usefulness of FASTeR. The teachers' average responses indicated that if they had had FASTeR during their FAST PD, they would have found it useful for setting up student investigations, understanding classroom interaction, and presenting concepts to the students. They also agreed that they would have used it during their first years of teaching.

Table II-5
*Teachers' (*N *= 25) Responses to Five Likert-scale Items about the Usefulness of FASTeR*

Item	Mean	St. dev.	S.e.$_M$
Reflecting back to my FAST professional development institute, I would have found the video examples to be useful for setting up investigations.	4.28	0.74	0.15
Reflecting back to my FAST professional development institute, I would have found the video examples to be useful for considering classroom interactions.	4.20	0.76	0.15
Reflecting back to my FAST professional development institute, I would have found the video examples to be useful for presenting concepts to students.	4.32	0.69	0.14
Reflecting back to my FAST professional development institute, I would have used the FAST electronic resources during my first years of teaching FAST.	4.28	0.94	0.19

Note: (*1* = strongly disagree, *2* = disagree, *3* = unsure, *4* = agree, and *5* = strongly agree)

CHAPTER III
DEVELOPMENT OF THE INQUIRY SCIENCE IMPLEMENTATION AND OUTCOME MEASURES

In Chapter II, we described the development of significant portions of FASTPro, the alternative version of FAST PD that we plan to study in future experiments and other studies. Developing FASTPro was one of two major components of our study; the other was the development of the instrumentation for conducting these studies and the validation of data collected with the instruments. The instruments that we developed addressed the implementation and intended outcomes of inquiry science and the context within which it is implemented.

In this chapter, we describe the development of our measures, including the Inquiry Science Observation Code Sheet (ISCOS), the Inquiry Science Teacher Questionnaire (ISTQ), the Inquiry Science Questioning Quality (ISQQ) method, and the Inquiry Science Student Assessment (ISSA). The ISOCS is a method for coding and analyzing videotaped observations of teachers in inquiry science classrooms, with a focus on the interaction between teachers and students that is initiated by teachers' questions of students. It addresses the adherence aspect of implementation. The ISTQ is a self-report instrument for collecting data on (a) the implementation (exposure and adherence) of inquiry science in the classroom and (b) the context within which teachers implement inquiry science, including teacher demographics; teacher perceptions, behaviors, attitudes, opinions, interests, and beliefs; some classroom variables; and the support the school provides teachers to implement inquiry science. The ISSQ uses the paired-comparison method, conducted by expert judges, for measuring the quality aspect of program implementation. The ISSA is our outcome measure. It includes multiple-choice items, extended-response (i.e., written-response) items, and a performance assessment. It also includes attitudinal items that can be examined not only as outcome measures but also as measures of the participant responsiveness aspect of program implementation. Our focus in this chapter is on instrument development; in Chapter IV, we conclude with our description of studies of the validity of the data collected with the instruments.

DEVELOPMENT OF THE INQUIRY SCIENCE OBSERVATION CODE SHEET[3]

We developed and tried out the Inquiry Science Observation Code Sheet (ISOCS), one of our measures of the adherence aspect of implementation. It is used primarily to count the frequency of instances of teacher behaviors as they interact with students. The ISOCS is included in the Inquiry Science Observation Guide, an in-depth, comprehensive, stand-alone manual that includes information about observation data collection ranging from videotaping to coding. A complete copy of the Observation Guide is provided as Appendix B to this report.

[3] Some of the material in this section was presented by Taum and Brandon (2005a, 2005b, 2006).

Much of what is described in this section is also reported in the Guide.

When preparing the instrument, we examined two observation instruments: the FAST Classroom Observation Instrument, a FAST research instrument (based on the Instrument for the Observation of Teaching Activities [National IOTA Program, 1970]) that has been used to collect observation data in previous FAST studies, and an observation protocol and teacher log developed at the Stanford Educational Assessment Laboratory, with whom CRDG collaborated on another NSF project (Grant No. ESI 0095520) about FAST. Our reviews of these two instruments proved helpful as we thought through our conceptualization of the ISOCS, but we did not base it on them.

Rationale and Focus

The ISOCS is a protocol for identifying low-inference behaviors or events—that is, specific, unambiguously observable behaviors or events that are simple to identify, in contrast to high-inference behaviors events "where characteristics being observed are more global or nebulous in nature" (Evertson & Green, 1986, p. 10). The data collected with the instrument are analyzed as frequencies and percentages.

Program implementation observation instruments typically are used for recording whether behaviors occur or for rating behaviors on some criterion (e.g., quality of delivery). Behaviors on most of these instruments are recorded as a single datum across the breadth of the observation period, whether that period be an entire class or one of a sequence of brief time increments (say, two minutes or five minutes). In contrast, using the ISOCS, observers record each behavior in the stream of classroom events as they occur across the breadth of the observation period (one class period). Observed instances are recorded as occurring but are not rated on any criteria. We chose to obtain precise counts of classroom behaviors on the assumption that data on exact frequencies are more precise and therefore more valid than observers' recording of frequencies or ratings across entire blocks of elapsed time. As we report in this section, however, we found that it is not a simple endeavor to obtain reliable records of the stream of behaviors and events as it unfolds.

The ISOCS focuses on teachers' use of questions and the interaction that follows their questions. Questioning is the preferred method of interaction in inquiry science classes, because teachers' role in these classes is not to instruct students directly but to guide them as they develop, implement, and interpret small scientific investigations. Inquiry science teachers, of course, often interact with students without questioning them, but the primary means of helping students learn in a constructivist, hands-on fashion is by asking questions. Findings about the effects of teachers' use of questions vary among studies, but research in general has shown that teachers' proficient use of the appropriate questioning strategies improves student learning (e.g., see Gall, 1970; Gall, 1984; Redfield & Rousseau, 1981; Samson, Sirykowski, Weinstein, &

Walberg, 2001).

Collecting Observation Data on Videotape

To provide data for developing and validating the ISOCS, as well as for the ISQQ, we videotaped FAST 1 physical science (PS) in the classrooms of a sample of 16 public- and private-school teachers on the four major Hawaiian islands during School Year 2004–05. The videotaping targeted five FAST 1 PS student investigations (PS4, PS7, PS10, PS12, and PS13), which occurred at key junctures in the sequence of 14 investigations that students conducted on buoyancy and density. The key junctures had been identified in a previous study of the FAST program (Stanford Education Assessment Laboratory & Curriculum Research & Development Group, 2005). We hired part-time employees on each island, trained them in how to videotape lessons, and provided them with video cameras and other equipment, including boom and lavaliere microphones, digital cassette tapes, tripods, watch and camera battery replacements, and battery chargers. Each person conducting the taping used checklists to prepare for taping and followed taping guidelines that were designed to ensure that the data were collected uniformly and that details were not overlooked. The guidelines, checklists, and logs are described in greater detail in the ISOG (Appendix B). The videotaping personnel recorded comments about events and activities at the school that affected the class during the taping sessions and noted the sessions in logs. We asked the teachers to keep the videotaping personnel apprised of their progress through the early FAST 1 investigations and to inform them when they anticipated teaching the next targeted lesson.

We did not tape all the targeted FAST investigations in all the teachers' classrooms because of unanticipated issues such as scheduling conflicts, communication problems, and faulty equipment and because for some of the investigations, the teachers integrated FAST with other programs. By the end of the year, we had videotaped a total of 135 FAST class periods—up to five full FAST investigations (PS 4, 7, 10, 12, or 13) per teacher. We transferred the videos to DVDs (one DVD per class period). We then viewed samples of every five-minute increment of every DVD and classified the quality of audio and video of the teacher on the DVDs. These quality checks showed that we had 91 DVDs classified as 100% acceptable and 16 classified as 75% acceptable, for a total of 107 class periods to use for instrument development and validation. The audio or video quality of the other 28 were deemed to be inadequate for coding the classes.

ISOCS Development

The ISOCS was developed over a two-year period during which a team of eight CRDG FAST program developers, FAST trainers, FAST teachers, and researcher/evaluators, as well as several coders and our advisory board members, collaborated in an iterative process of development, review, and revision. The instrument was revised in varying degrees ranging from minor to major over about 40 iterations. In this section, we describe the evolution of the instrument, in part to show the complexity of the development process (which was not always linear) and in part to provide the foundation for making claims about the content and construct validity of data collected with the instrument.

The initial trials of the instrument were conducted by the lead ISOCS developer and two graduate student coders. Later, the coding was primarily conducted by the lead developer and two other coders. They continued to refine the instrument, resulting in the version shown in the ISOG. The development of the instrument is described in detail by Taum and Brandon (2005a, 2005b, 2006).

The ISOCS did not initially focus sharply on teachers' questioning behaviors. Instead, we began the development by seeking to cover the breadth of FAST variables. We reviewed FAST program materials, including the FAST Instructional Guide (Pottenger & Young, 1992a), FAST student book (Pottenger & Young 1992b), and FAST Teacher's Guide (Pottenger & Young, 1992c). We prepared an outline of each of the 88 investigations in FAST 1, highlighting both teacher and student activities. Ultimately, however, we decided that the outlines were much more detailed than necessary for the proposed project and that it would be more feasible to identify those variables that were manifested in all FAST 1 PS investigations. From our review of these documents, we developed a list of 21 observable inquiry science classroom activities and behaviors, which formed our first list of items on the instrument.

In the next step of the development, we examined the extent to which the 21 activities and behaviors reflected good teaching practice as defined in the Five Standards for Effective Pedagogy model (Tharp, Estrada, Dalton, & Yamauchi, 2000) and reflected in the Standards Performance Continuum (SPC) (Hilberg, Doherty, Epaloose, & Tharp, 2004). Underlying both FAST and the Five Standards model is the Vygotskian theory that learning takes place through collaboration, whether through informal social interaction or a more formal scientific community of classroom learners (Taum, 2004). Therefore, the SPC was a logical starting point to begin to refine our inquiry science observation prototype.

The five standards include *Joint Productive Activity* (defined as teachers and students working together on classroom activities), *Language and Literacy Development* (developing the language of instruction), *Contextualization* (connecting school to students' prior knowledge), *Challenging Activities* (challenging students with cognitively complex activities), and *Instructional Conversations* (engaging students in classroom dialogue). Inquiry science students arranged in small groups produce Joint Products in the form of laboratory investigations. Each FAST lesson begins with a whole-class discussion in which students are expected to produce definitions for new vocabulary words, illustrating the Language and Literacy Development standard. The Contextualization standard is manifested in the sequence of FAST investigations, which ensures that students build on their prior knowledge. Challenging Activities occur throughout the steps of FAST investigations, in which students develop hypotheses, design experiments, describe data, develop conclusions, and generate new hypotheses.

After many rounds of observation coding, discussions among the FAST program developers and the ISOCS developers, and protocol revisions, it became increasingly clear that the standard that best fit our focus on teaching science through inquiry was Instructional Conversations. We had intended when we began the development process to address as many observable FAST activities as possible, but in the interest of validity, as well as reliability and cost-effectiveness, we came to focus on teacher-student interaction that is guided by teachers' questions. The theory underlying FAST holds that the aspect of teacher behavior that is most likely to enhance student learning is teachers' use of questioning strategies to guide classroom discussions. FAST teachers ask questions that lead students through the steps of the investigation while requiring them to address the proper choice of method, reason through the steps of the analysis, and interpret the findings of the investigations. Because teacher questioning of students is a central part of the Instructional Conversations standard, we limited ourselves to this standard.

The other major revision that we made was to restructure the instrument so that it provided for coding in a manner that mirrored the chronological occurrence of activities and behaviors that occur in FAST student investigations. The SPC rubric did not allow us to record sufficient detail about teachers' behavior in inquiry science classes, because SPC data are recorded with a 0–4 score for each standard over an entire observation period. We revised the instrument to record data about teachers' use of questioning in "strings" that showed how teachers began classroom discussions with questions, the topic of the discussions, how students responded to teachers, and how teachers in turn responded to students. The revision required that we add items to the original list of 21.

The revised ISOCS was divided in six major types of activities that were found in the FAST investigations (three of which had to do with types of teacher-initiated questions), with follow-on activities for each. The new structure allowed us to code the parts of investigations during which the teacher initiated discussions with questions. Coders began with one major activity and then looped among activities.

Our coders reported that the revised structure helped them code more reliably, but more improvements were needed. Showing specific activities for each of the six types required that we have duplicate codes across sections, because observable activities or behaviors could follow more than one major type of activity. The duplication resulted in coding disagreements among coders. The duplicates were eventually eliminated. We also revised the labels for the six major types of activities because of ambiguities among them.

Throughout the development period, the researchers monitored the length of time required to code and the length of time for the coders to come to agreement. In an effort to make the instrument as resource-efficient as possible, we constantly strove to reduce its complexity. By the end of the development process, the total number of items on the instrument = 31. The instrument, as shown in the ISOG in Appendix B, has five sections (columns). Coding begins when the observer views a *clarifying*, *lifting*, or *summarizing* question (Young & Pottenger, 1983) and records the code using the list in Column A. In Column B, the science investigation activities in which the students are engaged—allowing for multiple activities to occur simultaneously—are shown. Column C shows the types of student responses, ranging from no response to a comment or question. Column D lists possible teacher responses to the students, including (a) no response, (b) non-verbal acknowledgment, (c) verbal acknowledgment, (d) repeating, (e) rephrasing, (f) using a follow-up statement, (g) goal-oriented directing, and (h) probing further. Finally, Column E shows codes for the teacher actively moving throughout the classroom, making contact with individual groups, or addressing the class as a whole. (Eventually, Column E was dropped from the ISOCS.) The variables that the instrument addresses are shown in Appendix C.

When coders view a teacher asking one of the three types of questions shown in Column A, they note the minute and second on the DVD player elapsed-time clock and then begin a chain of codes that continues until the interaction ends or the teacher asks another clarifying, lifting, or summarizing question. For each of the three types of initiating questions, the coders continue the chain in several steps. First, they select the activity in which the students are engaged (Column B), choosing one or more from a list of 14 different activities, with allowance for multiple activities occurring together. Second, they code the type of student response to the teacher's question (Column C). Third, they select a code for the teacher's ensuing response to the student (Column D). Each chain of activities that begins with the initial teacher question can include multiple loops of student-teacher interactions.

DEVELOPMENT OF THE INQUIRY SCIENCE TEACHER QUESTIONNAIRE[4]

The purpose of the Inquiry Science Teacher Questionnaire (ISTQ) is to collect information on the exposure and adherence aspects of teachers' implementation of inquiry science and on the context within which inquiry science is taught. In this section, we describe the development of the instrument and analyses of the validity of data collected with it.

Preparing a Draft List of the Variables That Address Teaching Science with Inquiry Methods

To identify variables about the implementation of inquiry science teaching, we reviewed

[4] Some of the material in this section was originally presented by Brandon and Taum (2005a, 2005b).

documents and conducted an iterative review and revision of the list of variables with FAST developers and teachers. We began by reviewing pertinent FAST inquiry science program documents, including the instructional guide (Pottenger & Young, 1992a), the student book (Pottenger & Young 1992b), and the teacher's guide (Pottenger & Young, 1992c). We then reviewed a monograph describing inquiry and the manifestation of inquiry within FAST that had been prepared for the project by the FAST senior developer (Pottenger, 2005). From these documents, we identified and prepared a list of the variables that all the FAST student investigations had in common. We subjected the list of variables to multiple iterations of review, discussion, and revision. A research team member and a FAST trainer reviewed the variables that had to do with FAST student investigations and classified them by investigation phase (introduction, conducting the investigation, and interpretation). A FAST teacher with several years experience in the program reviewed the list of variables and identified instances in the FAST student book and teacher guide in which the variables were manifested. She prepared descriptions of these instances, which served to flesh out the meaning of the variables. A research team member and a FAST trainer reviewed the descriptions and fleshed them out further. Finally, the resulting list of variables was reviewed several times by FAST developers, resulting on each occasion in revisions, deletions, or additions. Variables were (a) revised because of vagueness or inaccuracy, (b) deleted because they were insufficiently central to FAST or to ensure that the data collection would be feasible, or (c) added to ensure that data would be collected on the aspects of student inquiry that are essential for student learning. The list of variables is shown as Category A in Appendix C.

Identifying the Variables That Address the Context Within Which Inquiry Science Is Taught

In addition to addressing the implementation of inquiry science, the ISTQ addresses the context within which inquiry science is taught. By context, we mean the teacher, classroom, and school variables that previous research has shown to affect program implementation and outcomes. We identified a few community variables as well.

These aspects of context cover a large swath of the educational research literature, making identifying them a substantial task. Therefore, we turned to summaries of the literature on these variables, particularly the curriculum-indicator literature and the school effectiveness literature. We identified, reviewed, and in many cases, outlined about 55 pertinent books, monographs, and articles that widely reviewed these bodies of literature (e.g., Blank, Porter, & Smithson, 2001; Blank, 1993; Bosker & Scheerens, 1994; Carey & Shavelson, 1989; Carey, 1989; Creemers, 1993; Creemers & Reezigt, 1996; Creemers, Reynolds, & Swint, 1996; Creemers & Scheerens, 1994; Darling-Hammond & Hudson, 1989; Fullan & Stiegelbauer, 1991; Heck, Larsen, & Marcoulides, 1990; Heck & Mayor, 1993; Klein et al., 2000; Mayer, Mullens, Moore, & Ralph, 2000; Mortimore, Sammons, Stoll, & Lewis, 1989; Murname, 1981; Muthen et al., 1995; Oakes, 1989a, 1989b; Oakes & Carey, 1989; Porter, Kirst, Osthoff, Smithson, & Schneider, 1993; Porter, 1993; Reezigt, Guldemon, & Creemers, 1999; Scheerens & Creemers, 1989, 1993; Scheerens, Vermeulen, & Pelgrum, 1989; Shavelson, McDonnell, & Oakes, 1989; Slater & Teddlie 1992; Teddlie & Reynolds, 2001; Wahlberg & Shanahan, 1983; Wang, 1998; Willms & Kerckhoff, 1995). This review resulted in a 24-page table showing a total of about 325 entries about students, parents, teachers, classrooms, principals, schools, and communities that have been found to affect program implementation and student achievement. The table included the topics, the sources in the literature, and the conclusions about the topics that were drawn from the sources. To help organize the table, we classified the entries according to Creemer's (1993) model of educational effectiveness.

The next step was to apply our criteria for selecting variables identified in the literature. table. Our first selection criterion was that the variables had to be supported by multiple studies, and our second criterion was that it had to address our theoretical model for a group-randomized study of the effects of variations of PD on middle school inquiry science classroom implementation and student outcomes.

The first criterion was satisfied because we had identified most of the variables from literature reviews. The second criterion required that we delete some of the topics that we had identified in our reviews of the summaries of the literature, for one of three reasons. First, we deleted some variables because they were beyond the scope of the study. To be sure, all the variables on our initial list had been shown to affect program implementation and student achievement. However, some were not essential for examining our theory of the effects of variations of PD in middle school science classrooms because they were not important enough to the theory. Others were not essential because their effects could not reasonably be expected to affect achievement in the group-randomized experiment that we planned for Phase II of our study. For example, school variables such as the financial resources available to the school or district resources such as the support given to the program by the district have been shown to affect school programs, but they are not essential to the success of inquiry science and their effects are unlikely to be sufficiently strong to be shown to affect inquiry science implementation or outcomes. Our Project Advisory Board (J. Bradley Cousins, University of Ottawa; Thomas Guskey, University of Kentucky; Jane Kahle, Miami University of Ohio; Paul LeMahieu, University of California at Berkeley; Maria Ruiz-Primo, Stanford University; and Richard Shavelson, Stanford University) concurred that we need not delve deeply into contextual variables beyond the classroom, with a few exceptions such as socio-economic status, attendance, school size, and ethnic distribution. Therefore, many of the school-level and district and community variables were eliminated from consideration.

The second reason that we deleted some variables was because the limitations of inferential statistical techniques require that we address a minimum number of variables in analyses of the results of our future group-randomized experiments. We endeavored to select as small a set of variables as possible so as to avoid having too many covariates in our planned future group-randomized experiment.

The third reason for eliminating some variables was that we were aware of the financial limitations for conducting a future group randomized study. The available resources will require us to limit the overall amount of data that we will collect to those that are essential to addressing the theory underlying the study. Collecting data that will be analyzed only descriptively will be helpful, of course, because it will help define our sample, but we expect that in future studies most of our financial resources will be devoted to collecting the primary data that we need for studying the effects of variations in PD.[5]

After preparing the list of 325 topics, we combined similar topics into variables and assigned them names. The final list of variables is shown in Appendix C. They are presented in five categories, including (a) inquiry science classroom implementation level; (b) inquiry science teacher characteristics; (c) inquiry science student characteristics; (d) school resources for, and constraints on, inquiry science implementation; and (e) district, community, and state resources for and constraints on inquiry science implementation. The variables that are addressed in the ISTQ (as shown in Appendix C) and the implementation aspects that the variables address, are shown in Table III-1.[6] A few of the variables shown in the list of 325 topics were added when we noted them after preparing the list.

Item Selection, Item Development, and Cognitive Interviews

The next step was to identify existing items addressing the ISTQ variables and write items addressing other variables. To identify potentially useful items, we reviewed instruments from several national studies and ongoing federal data-collection efforts such as Reform Up Close (Porter, 1993), the Survey of Enacted Curriculum (Council of Chief State School Officers, 2000), the Trends in International Mathematics and Science Study, the National Longitudinal Educational Survey:1988, the Study of Instructional Improvement (2001), the Schools and

[5] Furthermore, because of financial limitations, we did not develop some of the instruments that we had considered when the project began. For example, it would have been appropriate to ask all the teachers in middle schools some of the questions that we ask only of the science teachers in the ISTQ, such as the questions about school leadership or the collaboration among teachers. Data collected from all teachers in a school would be more reliable than data collected only from the school's inquiry science teachers. However, in the current study we chose to devote our resources to developing a good inquiry science teacher questionnaire.

[6] Appendix C includes other variables from the original list of 325 topics. Some are addressed on our observation instrument or our student instruments, and some will be addressed in our future studies by obtaining data from Internet sites such as www.schoolmatters.com, For some others, instruments must be identified or developed.

Staffing Survey (National Center for Education Statistics, n.d.), the Longitudinal Evaluation of School Change and Performance in Title I Schools (LeBlanc & Turnbull, 2001), and the USDOE's Fast Survey Response System (National Center for Education Statistics, 2007) for potential items. Items that addressed the variables were selected from these instruments and edited to fit the purposes of our questionnaire; other items addressing key features of inquiry science were written by project staff. Multiple items were prepared for most variables that could

Table III-1
Variables Addressed on the Inquiry Science Teacher Questionnaire, the Type of Variable (Adherence, Exposure, or Context) Addressed by the Variables, and Corresponding Item Numbers in Appendix C

Variables	Implementation aspect & Appendix C variable no.
1. The FAST investigations the teacher has taught or plans to teach during the year.	*Exposure:* A13
2. Teachers' implementation of 26 key features of inquiry science. The items (18a–z) addressing these variables comprise the *Inquiry Science Implementation Scale*.	*Adherence:* A1–A12
3. The extent to which the teacher (a) plans for and customizes FAST, (b) provides students with extra support, (c) integrates inquiry science with teaching other subjects, and (d) assigns homework.	*Adherence:* B17, A20, B20, B21, B22
4. Teacher demographics, including age, gender, the number of years they have taught in K–12 schools, the number of years the teachers have taught K–12 science, highest degree obtained, the number of undergraduate or graduate science courses taken, salary, certification, and training in FAST 2 or 3.	*Context:* B1–B10
5. Teacher attitudes toward science.	*Context:* B13
6. The extent to which the teacher shows interest in science by participating in science activities outside of the classroom including taking science PD courses (*Teacher Participation in Science Activities Scale*).	*Context:* B15
7. Classroom and school climate variables such the extent to which the teacher (a) has high expectations of students, (b) participates in school decision making, and (c) has opportunities to interact with colleagues at the school (the *Collaboration Frequency Scale* and the *Collaboration Benefits Scale*).	*Context:* B14, B19, B23
8. School support for teaching as manifested by the availability and adequacy of science the equipment, textbook resources, and other materials labs); school leadership support for inquiry science and support for teacher PD (the *School Support For Inquiry Science Scale)*; and the number of inquiry science teachers in the school.	*Context:* D1–D5
9. The teacher's perception of some student characteristics such as behavior and perseverance.	*Context:* C2
10. Miscellaneous descriptive teacher, classroom, and school characteristics, including the grades in which the inquiry science teachers are teaching inquiry science, the extent to which the inquiry science teachers are proficient with the Internet and computers, the number of inquiry science classes that the teachers teach, the total length of the inquiry science classes; whether the school groups students by ability level, the number of students in inquiry science classes, the grade levels served by the school, and whether the school is public or private .	*Context:* B11, B16, B18, D6, D7, E6, E7

be measured with scales, with the caveat that we endeavored to develop scales only for the context variables that we deemed to be most essential to effective schooling.

Drawing on the cognitive-interview procedures of Desimone and others at American Institutes for Research (Desimone & Le Floch, 2004), we developed cognitive interview procedures for our project.

> Messick (1989) maintains that traditional approaches to test development—those that are limited to examining patterns of relationships among item scores or between test scores and external measures—offer the weakest form of construct validation. Messick (1989) argues that a stronger form of construct validation, and perhaps the 'most illuminating,' of the approaches involves probing and modeling the cognitive processes underlying test responses. Based on work in the field of survey research, the cognitive laboratory method and think aloud procedure are new tools currently being explored for informing test development. The cognitive laboratory method utilizes procedures intended to assist in understanding respondents' thought processes as they respond to questions. . . . Interviewers ask respondents to think aloud as they respond to survey or test items. Interviewers also use probes to understand the cognitive processes respondents use in responding to questions (Desimone & Le Floch, 2004, p. 3)

The procedures were pilot-tested with two project staff members and revised as appropriate. Cognitive interviews then were conducted with six FAST teachers. Contrary to procedures recommended by AIR and others, the interviews were not taped, but extensive notes were taken. The results were reviewed immediately after each interview. By the end of the cognitive interviews, a total of 83 items were revised, 4 were added, and 9 were deleted.

As seen in Table III-1, the questionnaire included items that formed a total of five scales—the Inquiry Science Implementation Scale, the Collaboration Frequency Scale, the Collaboration Benefits Scale, the Teacher Participation in Science Activities Scale, and the School Support for Inquiry Science Scale. The first of these addresses the adherence aspect of implementation and the other four address program context. Other items that might be considered appropriate for developing scales (e.g., teacher attitudes toward science) were not formed into scales because we included too few items on the instrument. The instrument was prepared as an online questionnaire using Remark software. It is shown in Appendix D. It has three sections (A–C), with the option to pause data entry at the end of each section. Section A includes items about the use of FAST in the classroom, Section B asks about science activities at the teachers' school, and Section C asks about the teachers' background and experience. Teachers were prompted at the end of each section to answer any items that they missed. The instrument includes a total of 122 items about context, 28 items about the adherence aspect of implementation (Section A, Nos. 18–20), and, for the exposure aspect, a checklist of 43 FAST investigations that were taught or that the teacher planned to teach during the current year (Section A, No. 16). The checklist is our sole measure of exposure as we define it.

DEVELOPMENT OF THE INQUIRY SCIENCE QUESTIONING QUALITY METHOD[7]

To address the quality aspect of program implementation, we developed the ISQQ. The quality aspect of implementation is defined as the skill and knowledge shown by the service deliverer. In contrast to the evaluation of the adherence and exposure aspects of implementation, in which evaluators address the question, "How fully was the program implemented?", when evaluating quality, evaluators answer the question, "How well was the program implemented?"

The ISQQ is a paired comparison method (David, 1963; Torgerson, 1958). In paired comparisons, each object in a set of objects is paired with each other object, and trained judges select the member of each pair that addresses a specified criterion the most. This is a *preference vote*. For example, Heath and Brandon (1982) used the paired comparison method to compare a group of schools (the *objects*) on each of several criteria that defined effective special education; analyses of the results showed that the paired comparisons were conducted by two observers reliably. The method yields scaled objects, with unequal distances between objects. With the ISQQ, expert judges evaluate the quality of the implementation of questioning strategies by a sample of FAST teachers. The judges record a preference vote for each pair of teachers. They make holistic judgments, which are more feasible for addressing many characteristics than an analytic method such as an observation checklist. Judges of quality using the ISQQ keep in mind not only all the characteristics of good questioning but also the context within which each pair of teachers ask questions.

The development of the ISQQ included development of the description of the criteria on which teachers were to be compared and development of the procedures for conducting the paired-comparisons.

Development of the Description of the Criteria

The first step in developing the ISQQ was to describe the criteria that the judges were to address when comparing teachers. The goal of this step was to prepare a statement about one page in length that described the characteristics of high-quality questioning in sufficient depth to ensure that judges can accurately and reliably compare a sample of videotapes of FAST teachers. The senior FAST developer prepared a draft description, which was reviewed and revised in several iterations by the other members of the research team. The description drew in part on a monograph by the senior developer (Pottenger, 2005) and in part on a list of 26 key features of implementing FAST that we had previously identified, iteratively discussed, and revised when developing the ISTQ.

After reviewing the senior FAST developer's draft description of the criteria on several

[7] An earlier version of the material in this section was presented by Brandon, Taum, Young, Pottenger, Speitel, and Gray (2007).

occasions, the team's final version of the criteria stated, among other things, that "questioning is the heart of inquiry-based science teacher instructional activities," "student-teacher interaction revolves primarily around questioning that supports student engagement and learning without excessive praise or criticism of student responses," and, after asking questions, "the teacher listens to the students carefully, accepts what is heard, and ties students' responses to the teacher's initiating question." Good questioning strategies include "asking clear, unambiguous questions," "using Socratic question-answer chains," and asking questions such as "What do you think?," "What might happen if did you X?," "How might that be found?," "How do these results compare with our previous results?," "How are these results different?, " and "What is the evidence for that, and what is the quality of the evidence?" The full statement of criteria is shown in Figure III-1.

Preparation of the Facilities, Equipment, Materials, and Procedures

The second step in the development phase of the project was to prepare the facilities, equipment, materials, and procedures for trying out the ISQQ and collecting validation data. Facilities and equipment were reserved for the three-day ISQQ. An outline of the procedures was prepared and reviewed by the project team. A preliminary timeline was prepared and reviewed. A participant folder, including a welcome letter briefly describing the ISQQ purpose; the agenda; a list of planned daily activities; the list of quality questioning criteria; a checklist for viewing the videotape segments; and a note-taking sheet were prepared. Judge-training and ISQQ administration guidelines, with a description of the purpose of the study; a list of the necessary facilities, equipment, and materials; an agenda and chronological description of the procedures, including a suggested script for the trainers; and copies of the judge handouts were developed and described in a manual for the researchers.

Teacher Quality Criteria

The exchange of teacher questions and student responses is the sign of good inquiry-based middle school science programs such as Foundational Approaches in Science Teaching (FAST), and teacher quality in inquiry-based science classrooms is shown by full and appropriate use of questioning strategies. Questioning is the heart of inquiry-based science teacher instructional activities; the more that teachers ask the appropriate questions at the appropriate levels at the appropriate times, the better the inquiry.

Quality teacher questioning behavior is marked by more than the use of questioning strategies, of course, but without the full and proper use of these strategies, inquiry-based science will not succeed. In FAST, the focus of the questions that teachers ask varies among the three primary phases of lessons (Introduction, Investigation, and Summary), but the questioning approach remains constant and is manifested by these primary characteristics:

1. The teacher listens to the students carefully, accepts what is heard, and ties students' responses to the teacher's initiating question.
2. Student-teacher interaction revolves primarily around questioning that supports student engagement and learning without excessive praise or criticism of student responses. Questioning strategies include:
   a. asking clear, unambiguous questions at the appropriate opportunities for the purposes of initiating discussions and encouraging student curiosity.
   b. using Socratic question-answer chains.
   c. asking the children to reflect on possible answers to their own questions; for example, "What do you think?"
   d. asking questions that gain insight into students' behavior; for example, "What might happen if did you X?"
   e. asking questions about how investigations might be conducted; for example, "How might that be found?"
   f. asking questions asking for comparisons or contrasts; for example, "How do these results compare with our previous results?" and " How are they different?"
   g. asking questions about the sufficiency of evidence; for example, "What is the evidence for that, and what is the quality of the evidence?"
   h. asking questions about connecting the findings to everyday life.

Figure III-1. Teacher questioning quality criteria.

DEVELOPMENT OF THE INQUIRY SCIENCE STUDENT ASSESSMENT[8]

The student outcomes that we address in our theoretical model include student achievementand attitudes. Student attitudes address the *participant responsiveness* aspect of fidelity of implementation, as well. We addressed the measurement of achievement and attitudes in a suite of measures entitled the Inquiry Science Student Assessment (ISSA). The ISSA was designed to examine student learning in inquiry science classrooms. It is intended to be sensitive

[8] The ISSA was developed and validated by our subcontractor and collaborator, Carlos C. Ayala of Sonoma State University (Ayala, 2005a, 2005b).

to differences in FAST program implementation that in turn may be linked to variations in FAST professional development.

Briefly, the primary targets of ISSA are the science content knowledge and science inquiry covered in FAST 1 (matter, buoyancy, states of matter and energy) linked through *relational studies* (e.g., studies of the water cycle and air pollution) to concepts of ecology (e.g., plant and animal relationships to the environment). We hypothesized that increases in the fidelity and quality of instruction would result in greater student science content learning and increased science inquiry performance. The secondary assessment targets of the ISSA are attitudinal targets (i.e., student self-efficacy towards science investigations, motivation towards science, and student views of the nature of science). We hypothesized that students' relationship to science would change as they progressed through the FAST program. If students are asked to be scientists and actually carry out their own investigations, then how they view science might change.

Content Knowledge-Type and Science Inquiry Frameworks

Knowledge-Type Framework

The content and inquiry assessment blueprints for the ISSA were based on a content knowledge-type framework (de Jong & Ferguson-Hesser, 1996; Li & Shavelson, 2001), scientific inquiry (Duschl, 2003) and the FAST curriculum (Pottenger & Young, 1992a). The knowledge-type framework provides for a broad definition of science achievement based on the types of knowledge students are expected to learn in science. The content knowledge-type framework includes declarative (knowing that; *facts and concepts*), procedural (knowing how to; *measuring and experimenting*), schematic (knowing why; *explaining models*) and strategic knowledge (knowing when and how knowledge applies; *applying a procedure from one domain to another*). Shavelson and the Stanford Education Assessment Laboratory (SEAL) faculty and staff posit effective and efficient assessment methods that correspond to the different knowledge types. For example, while the extent of declarative knowledge can be easily assessed by multiple-choice items, the structure of declarative knowledge can be assessed via concept maps. Procedural knowledge is best assessed by performance assessments or laboratory "practicals." Schematic knowledge can be assessed with multiple-choice items and extended-response items (e.g., "Why do things sink and float?"). Strategic knowledge may best be assessed via performance tasks. Once the specific content to be covered in a curriculum has been identified and the content classified into the different knowledge types, assessment methods can be identified for the different content pieces.

Science Inquiry

Duschl's (2003) conceptualization of the assessment of scientific inquiry points towards the attainment and evaluation of data and evidence and how it is used to create models and explanations in three integrated domains: (a) conceptual (scientific knowledge and reasoning), (b) epistemic frameworks used when developing and evaluating scientific knowledge, and the (c) social processes that shape how knowledge is communicated. He argues that, "The assessment of inquiry is best thought of as a set of elements that place emphasis on examining the *processes* of engaging in scientific knowing and learning as opposed to the *products* or outcomes of scientific knowing and learning" (2003, p. 44). Assessing inquiry requires designing tasks to promote inquiry activities and to capture students' reporting and sharing information and ideas.

Duschl argues that the core of the inquiry process is about collecting data and transforming those data into evidence, transforming the evidence into models, and finally transforming the models into explanations that are used to develop new questions. Assessment of student inquiry should occur at three transformation points along the Evidence-Explanation (E-E) continuum. (See Figure III-2.) For our assessment of the materials covered in FAST curriculum, we extended the continuum beyond the three transformations and included reformulation—that is, as Duschl suggests, deciding on what data is needed and what questions to ask (an *E-E-R continuum*). The first transformation is selecting and evaluating data to become evidence, the second is analyzing evidence to create models and finding patterns, and the third is determining scientific explanations that account for the models and patterns. Reformulation occurs when students suggest new questions and decide on the data and collection methods they will need. Students share their thinking at each of the transformations by engaging in "argument, representation and communication, and modeling and theorizing" (Duschl, 2003, p. 45); student conceptualizations at this stage provide us with the opportunity to evaluate students' inquiry processes. Capturing the reasoning found in student judgments and explanations is critical to the success of assessing student knowledge along the EER continuum. Therefore, it is critical to ask students to present their supporting evidence for their explanations (e.g., "What evidence do you have to support your conclusion that objects sink based on their density?"). If students are working in groups, as they do in the FAST 1 curriculum, students should work in groups in the assessment. Finally, students report their findings and conclusions. We see these transformations and reformulations as cyclical and not linear; thus, there is no starting point, as seen in Figure III-2. Students may be asked to respond and provide supporting evidence for different locations on the EER continuum at different times in an assessment task. These transformations generally represent procedural knowledge; however, we believe that schematic knowledge of the whole inquiry method may be assessable.

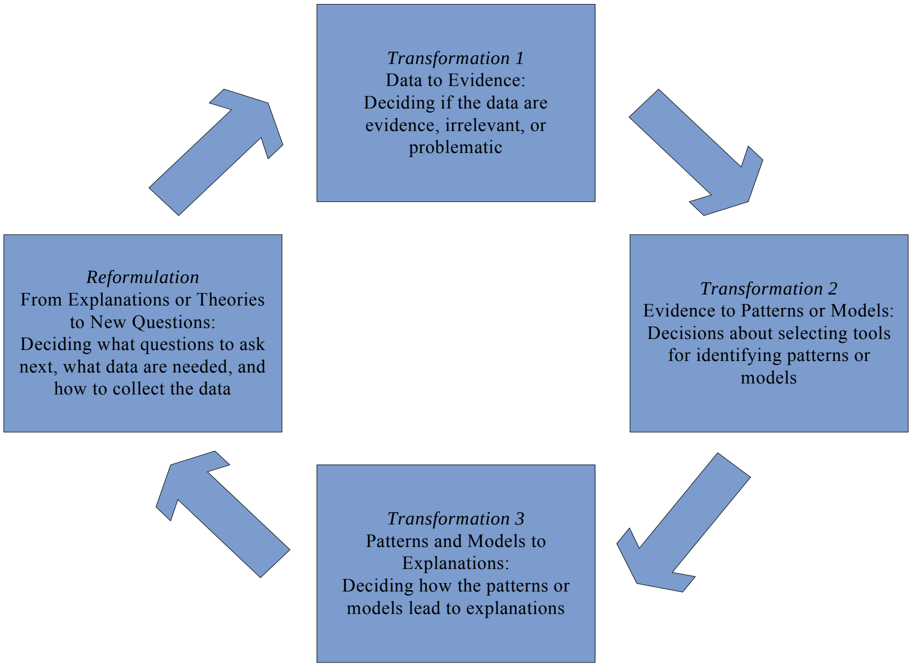

Transformation 1
Data to Evidence:
Deciding if the data are evidence, irrelevant, or problematic
Transformation 2
Evidence to Patterns or Models:
Decisions about selecting tools for identifying patterns or models
Transformation 3
Patterns and Models to Explanations:
Deciding how the patterns or models lead to explanations
Reformulation
From Explanations or Theories to New Questions:
Deciding what questions to ask next, what data are needed, and how to collect the data

Figure III-2. Three transformations and reformulation of inquiry science targets

In this study, we were interested in an individual summative assessment of student inquiry; therefore, in our performance assessment, we engaged students in the social aspects of the inquiry processes by starting off the tasks with the students in groups but then moved them to individual work. The group setting reflects FAST, in which students do group work (a validity match to the curriculum). When the students were engaged in individual work, we assessed students' ability to decide about data, create explanations from evidence with justifications, and measure their science communication. Students were asked to evaluate data and were asked questions at different times in the processes.

Shavelson's knowledge framework and Duschl's domains and transformations are not independent. Shavelson suggests that procedural knowledge and reasoning may represent these transformations.

Content Knowledge and Science Inquiry Outcome Measures Development

Content Knowledge

Following the successful assessment development process utilized by SEAL and CRDG during its previous collaboration in a NSF grant (SEAL & CRDG, 2005), the Sonoma State University team (SSU) went through the FAST student materials, instructor's guide, evaluation guide, and training guide and began the initial identification of the assessment targets for the ISSA using the knowledge framework. The key concepts, procedures and schema/models covered in FAST 1 were identified and classified into the three knowledge types. (See Table III-

2.) Ultimately, we identified 75 major potential science knowledge assessment targets. The topics that were identified were the primary foci of the lessons and materials reviewed.

To reduce the list to a smaller number of valid targets, the initial lists were presented to the CRDG team members (curriculum developers, FAST trainers, and researchers). In small groups, the CRDG team members reviewed the initial lists and began a process of selecting the most important targets covered in the program. In the previous NSF study (SEAL & CRDG, 2005), we had found this task to be quite difficult, because the curriculum developers found that all the targets were important. Therefore, in the current project, CRDG attended to the bigger ideas and content that bridge larger groups of lessons, as opposed to each individual lesson. Each small group team then wrote shorter lists of targets that reflected the most important targets. These reduced lists served as the main assessment targets.

The CRDG team then described the different elements of inquiry that were important in the FAST curriculum. The main targets identified by Duschl fit the FAST model; those were ultimately chosen. Additionally, SSU and CRDG decided that the pre-test ISSA should take no longer than two instruction periods and that the post-test ISSA should take no longer than three instructional periods.

Once the assessment targets were chosen and classified into the knowledge-type framework, specific assessment methods were identified for each target. Next, the SSU team began collecting known multiple-choice and extended-response items from national and state level tests. While we were interested in developing the measures that were closely linked to the FAST curriculum, we also believed that high fidelity implementation would lead to students overall general science achievement. In order to make this link, of the assessment items used in the ISSA, eight are from the Trends in International Mathematics and Science Study, the National Assessment of Education Progress, or state achievement tests. To achieve a balance of items, we selected those measures that were for the most part close to FAST 1 content and a few that were distal to the content (related but not explicitly part of the FAST 1 content).

For the multiple-choice and extended-response test, three different test versions were created and administered to 200 middle school students whose curriculum closely matched the FAST physical science curriculum. To validate the link between what each item was intended to target and how the students interpreted each question, a test administrator asked students about each question on the different versions and reported problematic and discrepant items. These items were dropped or revised. We then carried out analysis of the items against total score and identified additional problematic items, and revised or dropped them. Following this process, one long version of the test was created. This version was then administered to CRDG’s University

Table III-2

Partial List of Assessment Targets Identified During First Review of FAST 1 Materials by Potential Assessment Method

Targets	Knowledge Type	MC[a]	CR	PA	POE	Concept map
Absorption (Hydrology)	declarative	✓				✓
Accumulation (Hydrology)	declarative	✓				✓
Acid (Ph)	declarative	✓				
Atmosphere	declarative	✓				
Buoyancy	schematic		✓		✓	
Calorie	declarative	✓				
Climate	declarative	✓				
Collecting/Organizing	procedural			✓		
Communicating	procedural		✓			
Condensation (Hydrology)	declarative	✓				✓
Constructing Data Table	procedural		✓			
Density	declarative	✓				
Density Of Gases	declarative	✓				
Density Of Liquids	declarative	✓				
Designing Experiments	procedural		✓	✓		
Displacement	declarative	✓				
Evaporation (Hydrology)	declarative	✓				✓
Extrapolating	procedural	✓				
Gases	declarative	✓				
Graphing	procedural	✓	✓			
Ground Water	declarative	✓				✓
Heat Exchange	declarative	✓				
Mass	declarative	✓				
Matter	schematic		✓			
Mixture (Solutions)	declarative	✓				
Movement Of Gases	declarative	✓				
Percolation (Hydrology)	declarative	✓				✓

[a]MC = multiple choice, CR = constructed response, PA = performance assessment, and POE = predict-observe-explain.

Laboratory School FAST 1 students (α for multiple-choice items = .86). The program developers and trainers reviewed and validated (with respect to content) the final version of the ISSA. (See Table III-3.) The final version of the instrument is shown in Appendix E.

Science Inquiry

Once we had identified the science inquiry targets and found them to address procedural knowledge and reasoning, we selected a performance assessment format as the appropriate assessment type. A science performance assessment is a lab practical in which students carry out an investigation to solve some problem (e.g., finding the density of this block using a balance and a beaker). These assessments are valued for their congruence with what happens in the science classroom as well as for what happens in the science lab. "A science performance assessment comes as close as possible to putting a student in a laboratory, posing a problem and watching as the student devises procedures for carrying out an investigation, analyzing data,

Table III-3

Student Assessment Content Knowledge Test

Item	Type[a]	Source	Knowledge type	Description
S1	Mc	Original	Procedural	What happens to two glasses (one hot) overnight?
S2	Mc	TIMSS	Declarative	Air is made of gases?
S3	Mc	Multiple	Declarative	Primary source earth water cycle energy?
S4	Mc	Multiple	Declarative	Snowball internal temperature?
S5	Mc	TIMSS	Schematic	Coastal and inland climate reasons
S6	Mc	Original	Procedural	Variables in how much water to lettuce need study?
S7	Mc	Original	Declarative	Specific heat of object and heat transfer.
S8	Mc	Romance	Schematic	Block floats w/o hole, what happens with a hole?
S9	Mc	Romance	Declarative	What happens to floating object in larger container?
S10	Mc	Original	Schematic	Which graph represents temp of heating water to boiling?
S11	Mc	Romance	Declarative	Ball of equal mass and volume, one hallow, do they both sink?
S12	Mc	Original	Procedural	Which question is the question she wanted to answer?
S13	Mc	MOD	Procedural	Which is did not contribute to different weather readings?
S14	Mc	TIMSS	Schematic	During the day, organisms use up or give off?
S15	Cr	TIMSS	Procedural	Machine X and Y, which is more efficient?
S16	Mc	Romance	Schematic	What happens to density when block of wood is cut?
S17	Cr	Romance	Schematic	Which object can be used to determine density of second liquid?
S18	Mc	NAEP	Declarative	Temperature of freezing of different amount of water.
S19	Mc	Multiple	Declarative	What happens to salt when water evaporates?
S20	Mc	Romance	Schematic	What factor has the greatest effect on sinking or floating?
S21	Mc	Original	Procedural	How much energy to heat water?
S22	Mc	TIMSS	Declarative	What happens to atoms after animal dies?
S23	Mc	Original	Declarative	What happens to water vapor as air temp increases?
S24	Mc	Mod	Declarative	Prediction of mass of melted ice in can?
S25	Mc	Original	Declarative	What is the boiling point of a mixture?
S26	Mc	Original	Procedural	Why do scientist measure something several times?
S27	Mc	Multi	Declarative	What is the best reason why hot air balloons rise?
S28	Mc	Romance	Schematic	Estimate the density of plastic block in two liquids?
S29	Cr	Multi	Schematic	Describe the water cycle
S30	Cr	Mod	Procedural	State the relationship between Insect A and Insect B

[a] Mc = Multiple-choice item, Cr = Constructed-response

drawing inferences from the data and his prior knowledge" (Shavelson, 1995, p. 59).

Science performance assessment can be characterized by three components called "the triple:" the *task* (a hands on activity or problem that students are asked to solve), the *response format* (the nature of the response the student is expected to provide—student notebook), and the *scoring system* (the method used to evaluate student performance) (Ruiz-Primo & Shavelson, 1996; Shavelson, Solano-Flores, & Ruiz-Primo, 1998). The task invites the student to solve a problem. It requires the use of concrete materials that react and provide feedback to the student, and it addresses the covered curriculum. The response format provides a means for students to record their findings, allows students to decide how to present their findings, and requires that students justify their answers. The prompt nudges students towards the procedures but does not spell it out for them. The scoring system reflects both the goals of the task and assessment targets (i.e., science inquiry), captures the scientifically justifiable procedures, and allows for insight into students' problem solving abilities.

To develop a performance assessment that captured FAST 1 content and science inquiry that captured the EER continuum, we chose a relational study task. A FAST 1 relational study is an investigation in which students explore an ecological situation using physical science principles. Since students in FAST 1 carry out investigations about pollution in the environment, we decided to emulate this with a performance assessment. We chose the idea of factories polluting a river in which to embed our assessment items for FAST science inquiry and the EER continuum. (See Appendix E.) In this task, students sample water from different locations on Rocky River and test the samples for high levels of pollution. Students must set up comparisons using controls, pollution indicators and limited testing sites. Students record their findings on the response format (the *notebook*). In the notebook, students are prompted to record their findings and explain why they reached their conclusions. Students are given the items about FAST science inquiry and EER continuum in the notebook. For example, students are asked to decide which factory is polluting the stream (Patterns to Explanations), provide evidence for their conclusions (Patterns to Explanations), and decide if the data that they are using make sense or not (Data to Evidence). (See Appendix E.)

The scoring system (a *rubric*) links the student responses to the assessment targets and assigns values to student responses. To capture the transformations and reformulations, students are asked to carry out the investigation first in a group and then asked to repeat the investigation with new sites by themselves. There are multiple opportunities to capture student thinking in the EER's conceptual, epistemic and social domains. To assess students in the conceptual domain, we focused on the soundness of their responses. To get at the epistemic domain, we teased out whether students understand how science knowledge is developed and evaluated. Finally, to get at the social domain, we attended to the extent of the students' science communication in each transformation and reformulation. We crossed the transformation and reformulation across the three domains and embedded these crosses in the scoring rubric. (See Table III-4.)

Possible Validity Threats

The major validity threats of the claim that the performance assessment measures an individual's science inquiry knowledge are the possibility of (a) an interaction effect between group partners and individuals' performance during group work and (b) the performance assessment not measuring the intended knowledge types. We conducted think-aloud procedures on four versions of the performance assessment and found that in most cases students' performance on the group tasks was superior to the individual work. That is, working in the group, as FAST students should, students perform better in their work, and working individually, students revert to earlier procedural knowledge and reasoning to solve the problems. The think-aloud

Table III-4
The Evidence Explanation Continuum Transformations and Reformulation and the Conceptual, Epistemic and Social Process Domains

Continuum transformation	Conceptual	Epistemic	Social
Data to Evidence	Are the student's data correct?	Do students understand the purpose of standards?	To what extent are student's science communication clear, focused with minor technical errors.
Evidence to Patterns	Does the student use the appropriate evidence to describe patterns?	Do students know how to present data in an organized way to make sense to others?	To what extent are student's is the presentation of evidence relevant, clear, focused with minor technical errors. Are data tables clearly labeled?
Patterns and models to explanations	Does the student select both factories (one more than the other) as killing the fish?	Are student explanations supported by evidence?	To what extent are student's explanations clear, focused with minor technical errors.
Deciding what new questions are needed	Does the student select new sites that would be meaningful?	Do student rationales for new sites express the reason for why the new information might be valuable?	To what extent are student's rationales for choosing new sites clear, focused with minor technical errors.

proceduress were also used to review how students conduct the investigations and if the wording of the performance assessment made sense to the students. Finally, we had experts, science graduate students, and science teachers (trained biologists) conduct the performance assessment to set the standards for the response for the rubric. A total of 272 students completed the Rocky River Performance Assessment.

Student Attitudinal Measures Development

Not all student outcomes are content related. Attitudinal measures are important when considering the effectiveness of a program, especially one that is intended to engage students in science, because students' perceptions and attitudes towards science may influence their learning. Motivation (Pintrich, 1999, 1993; Haydel & Roeser, 2002), Self-Efficacy (Bandura, 1986; Pajares, 1995,1996) and, in the case of science education, views of the Nature of Science (NOS) have been found to be related to student achievement. Furthermore, based on Bandura's social cognition theory (1986, 1977), we incorporated additional motivations constructs including science anxiety (Britner & Pajares, 2001; Pajares and Urdan, 1996) and science value (Britner & Pajares, 2001; Meece. Wigfield, & Eccles, 1990).

We developed an 81-item survey addressing motivation, self-efficacy, and NOS, adapting it from the multiple sources described below.

Self-Efficacy and Student Lab Performance

Bandura (1986) argued that self-efficacy is the most influential factor in human functioning. He defined self-efficacy as "people's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances" (1986, p. 391). Self-efficacy mediates the effects of prior achievement, knowledge, and skills on subsequent achievement. Thus, it is often a better predictor of success than actual abilities. This may help explain why people with similar abilities may have different levels of achievement. Self-efficacy affects behavior by influencing people's behavioral choices, the amount of effort they expend, and the persistence they exhibit in the face of failure.

Most research on science self-efficacy has focused on science teaching self-efficacy and science self-efficacy as a predictor of career. There are few investigations of confidence in science as a predictor of subsequent science achievement, and fewer investigations focusing specifically on laboratory skills or learning through science investigations and studies focusing on the effects of a particular science curriculum. Britner (2002) investigated middle school science students' self-efficacy with respect to science and science lab grades. She found that student science self-efficacy was positively associated with grades. Furthermore, girls' grades were also associated positively with science self-concept and negatively with value of science. We hypothesized that (a) because FAST 1 students are expected to carry out and learn from their science investigations, their judgments about their capabilities to carry out science investigations and learn from these science investigations should increase after participating in the program, and (b) in pedagogically strong classrooms, FAST 1 students' judgments about their capabilities would change more as they learned from their investigations throughout the year.

Instrument Development

A student survey was created to elicit student self-efficacy based on Britner (2000). The Science Investigation Self-Efficacy Scale was assessed with the Lab Skills Self-Efficacy Scale (Britner, 2002). This scale consists of 12 items asking students how sure they are that they can perform specific science process skills commonly used in laboratory activities (National Research Council, 1996). Britner's items were adapted to match the FAST 1 language. Students estimated their confidence that they could perform each skill on a scale from 0 (no chance) to 100 (completely certain). (See Table III-5.) This was administered to the students in a pretest suite and the posttest suite.

Table III-5
Motivation Constructs and Sample Items Addressing Them

Motivation construct	Sample item
Epistemic belief	How well I do in science depends on how smart I was when I was born. You are born smart in science. I have to be really smart to do well in science.
Ego avoidance goal	It is very important to me that I do not look stupid in my science class. One of my main goals in science class is to avoid looking like I can't do my work
Ego mastery goal	I like the work in my science class best when it really makes me think. An important reason I do my science work is to master challenging concepts.
Perceived ability goal	Our teacher points out those students who get good grades as an example to all of us. Our teacher lets us know which students get the highest scores on tests.
Perceived task goal	Our teacher wants us to really understand the concepts, not just to memorize facts. Our teacher gives us time to really explore and understand new ideas.

Science Anxiety and Value of Science Measures

In addition to making comparisons between science investigation self-efficacy and science performance assessments, it is important to look at the relationship between self-efficacy and science anxiety (sample item: Just thinking about science makes me nervous) and science value (sample item: I like doing science investigations). Following the lead of Britner & Pajares (2001), we explored the relationship between science value, science anxiety and science self-efficacy.

Motivation and Science Education Measures

Snow (1994) hypothesized that individual differences in achievement can be seen as a "moment to moment" transaction between characteristics of the person and the situation itself. Snow believed that individuals bring to a task certain cognitive and motivational aptitudes that shape their performance. In order to look at the relationship between students' achievement and FAST instruction, we decided to explore the relationship between motivation and achievement in FAST. We draw on Dweck and her colleagues' (Dweck, Davidson, Nelson & Enna, 1978) theory of the organization of achievement-related goals and competence-related beliefs that are linked to academic performance. Through empirical work and logical analysis of why some students engage with and perform better on particular tasks, she found three sets of motivational processes that predict differences in achievement outcomes. The sets include a student's beliefs about the malleability of their intelligence, the intellectual confidence and their achievement goal. They proposed three motivational types: (a) mastery-oriented students, (b) ego-oriented students and (c) helpless-orientation students.

Mastery-oriented students are students who believe that intelligence is malleable and can grow over time. They pursue goals to develop their intelligence. Ego-oriented students are defined as students who believe that intelligence is fixed and adopt goals to prove their fixed ability or to hide their inability. Students with confidence in their abilities view tasks as opportunities to reinforce their sense of superior ability. Helpless-orientation students, like the ego-oriented students, believe that intelligence is fixed; however, they have low confidence in their abilities and are thought to be preoccupied with the goal of hiding their inability from others. There is some evidence to suggest that students who are members of groups that traditionally are considered inferior intellectually (e.g., females in science) may be more likely to adopt this helpless orientation (Dweck et al., 1978).

We hypothesized that in FAST classrooms with high-quality implementation, students' motivational group patterns will be different than in incomplete or low-quality FAST implementation classrooms. That is, as the students are (a) exposed to investigations where they discover their own knowledge, (b) exposed to Socratic inquiry, and (c) asked to think for themselves and come to their own conclusions, their beliefs about what they can learn will be different than in classes where they are not asked to think for themselves or to come up with their own conclusions. Furthermore, motivational types can be used to further explain differences in achievement.

In order to identify students in the different motivation pattern types, measures of students' epistemic beliefs, self-confidence in science ability, and goal orientation were developed. The scales and the number of items in each were: (a) Self Confidence in Science Ability (6 items); (b) Epistemic Beliefs (4 items); (c) Inquiry Epistemic Beliefs (4 items); (d) Peer Epistemic Beliefs (4 items); (e) Ego Avoidance Goal (5 items); (f) Ego Mastery Goal (5 items); (g) Ego Performance Goal (5 items); (h) Perceived Ability Goal (5 items); and (i) Perceived Task Goal (7 items). Following Haydel and Roeser's (2002) method (adapted from Dweck and Henderson, 1989), the results on the measures can be used to classify the students into one of the motivational types. Comparisons between pre- and post- measure proportions as well as measured differences with respect to fidelity of instruction will show differences among groups.

Student Nature of Science Measures

The final attitudinal measures that we developed address student views of the nature of science. For many years, many scientists and science educators have agreed that an objective of science education is for students to have an informed conception of the nature of science (Abd-El-Khalick, Bell, & Lederman, 1998, Duschl, 1990; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). NOS refers to the epistemology and sociology of science—science as a way of knowing, as well as the values and beliefs inherent to scientific knowledge and its development. While some aspects of the NOS are controversial (e.g., the issue of an objective reality), some are more accessible to K-12 students. Lederman et al. (2002) argue that the views of NOS that are relevant to the daily lives of students are that "scientific knowledge is tentative, empirical, theory-laden, a product of human inference, imagination and creativity, and socially and culturally embedded" (p. 499).

In the FAST program, students are considered scientists and are asked to explore the world and come up with their own knowledge. We hypothesized that students in classrooms that have a higher degree FAST implementation fidelity will have a more realistic view of the NOS than students in the lower degree of implementation fidelity classrooms. Furthermore, in a pre-post test study, we would expect students to change their opinions about the NOS from more idealistic to more realistic and that classrooms with higher degrees of FAST implementation fidelity would show the greatest gains.

We used a Likert-scale questionnaire method to carry out this investigation. Since the purpose of our instrument is to examine the extent to which students' NOS views change, we believe the Likert-scale system works because we are not necessarily interested in the absolute

NOS view value, rather the differential view. Furthermore, in order to assure validity, we used items from several sources to develop our measures, and we verified the items with FAST curriculum developers, scientists and FAST trainers. This was to assure ourselves of a broad view of the NOS. We developed 10 Likert-scale items that address the NOS issues related to validity to everyday life. (See Table III-6.) The final version of the 81-item instrument is shown in Appendix E.

Table III-6
Nature of Science Domains and Sample Items

Domain	Sample item
Scientific theories and laws as absolute	Scientists are always right.
Science as socially embedded	All people who study hard and are smart can learn to be a good scientist.
Science as amoral	Science knowledge is not good or bad.
Science does not involve creativity	Scientists always get the same results.
Science is tentative and developmental	Scientific knowledge can change over time.
Science is useful	Scientific knowledge can be useful away from school.

CHAPTER IV
VALIDITY ANALYSES OF DATA COLLECTED WITH THE IMPLEMENTATION AND OUTCOME MEASURES

In this chapter, we report the results of our validity analyses of data collected with the ISOCS, the ISTQ, the ISQQ, and the ISSA.

INQUIRY SCIENCE OBSERVATION CODE SHEET VALIDITY STUDY

Data Collection

Selecting DVDs to Code

For our studies of the validity (including the reliability) of ISOCS data, we selected a sample of the 107 classroom observations of 16 FAST teachers that had been deemed 100% usable. (See Chapter III.) Our criteria for choosing the sample was that the selected investigations were taught by a variety of teachers (only one teacher was represented twice in the set) and represented all five of the targeted PS investigations (two instances of the teaching of each investigation were chosen). We eliminated one lesson for a teacher who was taped twice, resulting in nine teachers who were included in the study. The population of videotaped investigations and teachers and the selected sample are shown in Table IV-1.

Coder Training

After we developed and refined the ISOCS with two initial coders, we developed the training procedures that are described in the ISOG (Appendix B) and hired and trained eight individuals to use the instrument. The coders ranged from a veteran science teacher to others with little or no experience working in education. We hired some coders without education experience because we needed to hire as many as possible to code the videotaped FAST class periods and because we wanted to see if a multifaceted team of observers could reliably code the videotapes.

By the third month of coding, participation by those individuals from non-education backgrounds slowly began to fade until two stalwarts remained. The two remaining coders both had formal teacher training and classroom teaching experience, with one a former teacher with experience teaching science. This suggests that the non-educators are not well-suited for coding classroom observations.

The Coding Process

Over an eight-month period, the two remaining part-time coders observed, coded, and reconciled DVDs. Each coder independently viewed and coded each DVD; they then met to identify the codes that matched, discuss their differences, and reach consensus.

On average, coders took a total of three to six hours to code a class period, beginning with an initial viewing without coding, followed first by independently coding observed behaviors and then by reconciling codes with the other coder. The purpose of the initial viewing was to

Table IV-1

Population of 100%-Usable Videotaped FAST Physical Science Investigations and the Sample (Shaded) Examined in Validity Studies [a]

Teacher	Number of class periods, by investigation					Total
	PS 4	PS 7	PS 10	PS 12	PS 13	
01	0	0	1	2	0	3
02	2	3	1	0	0	6
03	3	3	0	4	2	12
05	3	3	2	3	0	11
06	1	1	0	0	0	2
07	4	4	2	2	0	12
08	0	0	0	0	0	0
09	0	0	0	1	0	1
11	1	0	0	0	0	1
12	0	0	0	0	0	0
13	0	0	0	2	3	5
15	0	0	4	0	1	5
16	1	2	2	2	3	10
18	2	0	0	0	0	2
20	3	1	7	0	0	11
21	2	3	0	2	3	10
Total	22	20	19	18	12	91

[a] Only those investigations for which there were three taped class periods were considered for the validity studies. Some teacher numbers are missing because some teachers dropped out of the project.

allow the coders to learn about (a) the teacher's intended activities during the lesson and whether he or she conducted the activities as intended, (b) any unusual situations that may have occurred and interfered with the taping (e.g., a fire drill, student emergency, and so forth), and (c) whether the DVD was audible and the teacher was on camera (observable) for the majority of the taped lesson.

During the second, independent viewing, observed teacher behaviors were coded. Coders recorded the minute and second at which teachers asked any of the types of questions shown in Column A of the ISOCS and then recorded the activities in Columns B–D that ended with the next teacher-initiated question. Initially, this stage of the coding was time consuming, because it required that the DVD be paused while matching observed behaviors with ISOCS codes. Over time, the codes became familiar to the coders, which sped up the coding process.

The final stage of coding involved the two coders comparing their individual codes and reconciling differences between them. They first identified similar recorded start times (e.g., 12:44 for one coder and 12:42 for the second coder), and they then compared the "strings" of codes (e.g., A3, B8, C1, and D2 for one coder and A2, B3, B8, C1, and D2 for the second coder). The coders also recorded relevant notes that proved helpful in expediting the reconciliation process. This process resulted in one set of codes for each teacher. Coder 1 assigned a total of 1,653 codes and Coder 2 assigned a total of 1,466. The number assigned to each code by each teacher, as well as the reconciled code, is shown in Table IV-2.

Validity Analyses

Content Validity

Relevance and Representativeness

The content aspect of validity addresses "content relevance, representativeness, and technical quality" (Messick, 1995, p. 745). Content validity evidence for an observation instrument such as the ISOCS is found in the extent to which it can be shown that the data collected with the instrument are (a) relevant to the measurement task (i.e., they capture what is intended to be observed), (b) representative of the content domain that is to be measured, and (c) of sufficient technical quality (e.g., they are reliable).

Evidence about the extent to which the data collected with the instrument are relevant and representative is found in our description of the instrument development. This evidence has to do with the process by which the instrument was developed and the focus of the instrument on teachers' questioning behaviors and student-teacher interaction. The care with which the process was conducted and the thoroughness of the process are the strongest part of the evidence. As described previously in this report and elsewhere (Taum & Brandon, 2005a, 2005b), the instrument initially was developed in an extended iterative process that balanced the goal to collect data on the breadth of topics that conceivably could be observed in inquiry science classes against (a) the goal to take a deep look at teacher questioning and (b) the need to ensure that the observations were feasible—that is, resource-efficient and cost-effective. The iterative process began with a review of FAST documents, in which we identified topics that the instrument could address. The topics included teacher behaviors, student behaviors, classroom

Table IV-2

Number of Reconciled Codes for Each Code and Percentage on Which the Two Coders Agreed on Their First Coding

Code	Total *N*	Percentage	Code	Total *N*	Percentage
A1	46	17	B12	19	5
A2	112	28	B13	6	50
A3	41	37	B14	6	0
B1	22	23	C1	22	9
B2	11	9	C3	657	42
B3	26	37	C4	48	21
B4	4	25	D1	16	13
B5	50	20	D2	6	0
B6	21	14	D3	51	10
B7	49	6	D4	88	14
B8	54	7	D5	32	12
B9	47	13	D6	293	26
B10	27	4	D7	52	66
B11	85	27	D8	312	50
—	—	—	Total	2,203	—

events, classroom characteristics, and so forth. This was the stage at which we focused on breadth. The number of topics grew as the process continued, but after encountering difficulties in obtaining interrater agreement during the early stages of development, the number of topics shrank as we focused more on depth. At this point we began to narrow the scope of the instrument to teacher behaviors in their interaction with the students investigations, the types of tasks in which the behaviors occurred, and students' responses to teachers. Content relevance was ensured throughout the process by the FAST program developers and some of the advisory board members. These team members kept the researchers focused on the features of inquiry science classes that are most likely to affect student learning.

In this manner, the instrument increasingly came to address the student-teacher interaction in inquiry science classes that begins when a teacher questions the students. This interaction primarily is manifested in Code C3 (the code, as shown in Appendix B, for students' responses to the teacher) and Codes D6 and D8 (the codes for the teachers' use of follow-up statements and of probing questions). These aspects of student-teacher interaction are the heart of the interchang e that is initiated by teacher questions; they reflect the primary content that the instrument is intended to address. They are relevant to the purpose of the instrument, and they are representative of the intended content—indeed, they comprise that content. Data on the student science investigation tasks that occur when the teachers ask questions (e.g., making connection with previous investigations, discussing tools, dealing with data issues, and so forth) also are collected to provide information about the context within which the teacher-student interaction occurs. (See Column B of the instrument.)

The results for each code are analyzed as the percentage of all the codes recorded for a teacher for a given unit. A unit might be a class period, all the periods that it takes to complete an investigation, or all the observations coded for a teacher. Thus, if a total of 100 codes are recorded for a teacher during a given unit (say, a class period), and if 20 of the codes are C3s, the results for the teacher for C3 is 20 percent. For each teacher, the denominator for calculating the percentages for the validity analyses discussed in this report is the total of the codes across all the teacher's videotaped class periods.

Technical Quality

Evidence found in the description of procedures. Evidence for technical quality is found in our description of the procedures for training coders and for administering the instrument, as presented earlier in the body of this report, in the ISOG in Appendix B, and in previous papers (Taum & Brandon, 2005a, 2005b). The researchers conducting the training used a thorough training guide. The coders were trained over several days. Coders who were unwilling to participate or who showed diminished interest once they experienced the intensiveness and extensiveness of the coding dropped out, resulting in a team of two coders who had background in education. These two coders continued to refine the instrument over a period of several months, resulting in the final 31-item version of the instrument. Each of these characteristics of the development added to its technical quality.

Evidence from the results of reliability analyses. In addition to the procedures of the observation process, evidence about technical quality is also found in the results of reliability analyses. Reliability analyses of observation instruments should address both consistency and consensus. Consistency is the degree to which judges are consistent in their assignment of codes, and consensus is the degree of agreement among coders.

Consistency analyses are typically correlational. We conducted two correlational analyses of the two coders' assignment of codes. In the first analysis, we correlated the assignment of codes for all coded teachers and lessons combined. That is, we calculated the total number of choices of each code by each coder and calculated the correlation between totals for the two coders. The numbers of assigned codes are shown in Table IV-2. The Pearson correlation between the two sets of codes was .99, suggesting an extraordinarily strong relationship. In the second analysis, we calculated the correlation between the total number of codes assigned to each teacher (not shown here). The correlation between these two sets of codes was .53, suggesting a moderate relationship. Because the second correlation is a finer-grained analysis than the first, we believe that its results are more meaningful. However, the first correlation does suggest that across the group of teachers, the ISOCS coders were consistent.

Consensus analyses are typically shown as the percent to which judges agree with each other. We calculated the percentages that the two coders agreed on the initial coding on each code. For each code, the percentage of the total of the codes assigned by a coder is the appropriate statistic for comparing the results between the two coders. The denominator for the percentages = *N* agreed codes + the total *N* other codes, across coders. The results are shown in Table IV-2. The percentages ranged from 5% agreement for one of the codes for a science investigation task (Code B12) to 50% for teacher probing (Code D8). The percentage for student comments in response to the teacher's initial question (code C3) was 42 and the percentage for teacher follow-up was 26 (Code D6). In all but eight of the cases that required reconciliation, one of the two coders did not assign a code—that is, one of the two coders did not observe the event or behavior that the other had coded. Thus, reconciliation was rarely necessary because the coders observed different events; it was necessary because of errors of omission, not because errors of commission.

None of these percentages are high. The coders might have been somewhat lax in their initial coding because they knew that ultimately they were going to reconcile the differences between their two sets of codes. Furthermore, some of the percentages for Column A of the ISOCS might be low because, as we came to conclude over time, the distinction between clarifying, summarizing, and lifting questions is ambiguous. However, we believe that intercoder agreement results for the ISOCS cannot be compared with the desirable or typical results for observations in which events are recorded in time periods. Our method is more stringent than either of these two methods, because our coders had to agree precisely on codes at each point in a string. Recording discrete, brief, low-inference behaviors and events might initially be more fraught with the possibility of error than counts in time blocks because only one instance per block has to be seen, whereas the ISOCS requires that all instances be exactly recorded. Clearly, the reconciliation between the coders' results is essential for collecting ISOCS data.

Concurrent Validity

We correlated mean ISQQ teacher quality ranks (see p. 75) with the results for ISOCS Codes C3 and C6 for the nine teachers who videotapes were analyzed for both methods. The Spearman's rho correlation of the quality ranks with the ISOCS percentage that student comments constituted of all teacher codes = .52, and the Spearman's rho correlation of ISQQ ranks with the percentage that the teachers used follow-up statements and probing questions = .45. Confidence intervals for these correlations are substantial, of course, because of the low *N* of teachers whose results were correlated; nevertheless, we believe that these correlations show a relationship between the two sets of results, thus supporting the validity of the ISOCS data.

Criterion-Related Validity

We conducted a criterion-related validity analysis using student achievement results. Of the teachers for whom observation data were collected and coded, six were administered the ISSA. The correlation between teacher-student question-response exchanges (coded as the percentage of the total number of observed behaviors) and mean multiple-choice total posttest score was .96. The scatterplot of these values is shown in Figure IV-1. The correlation is only suggestive

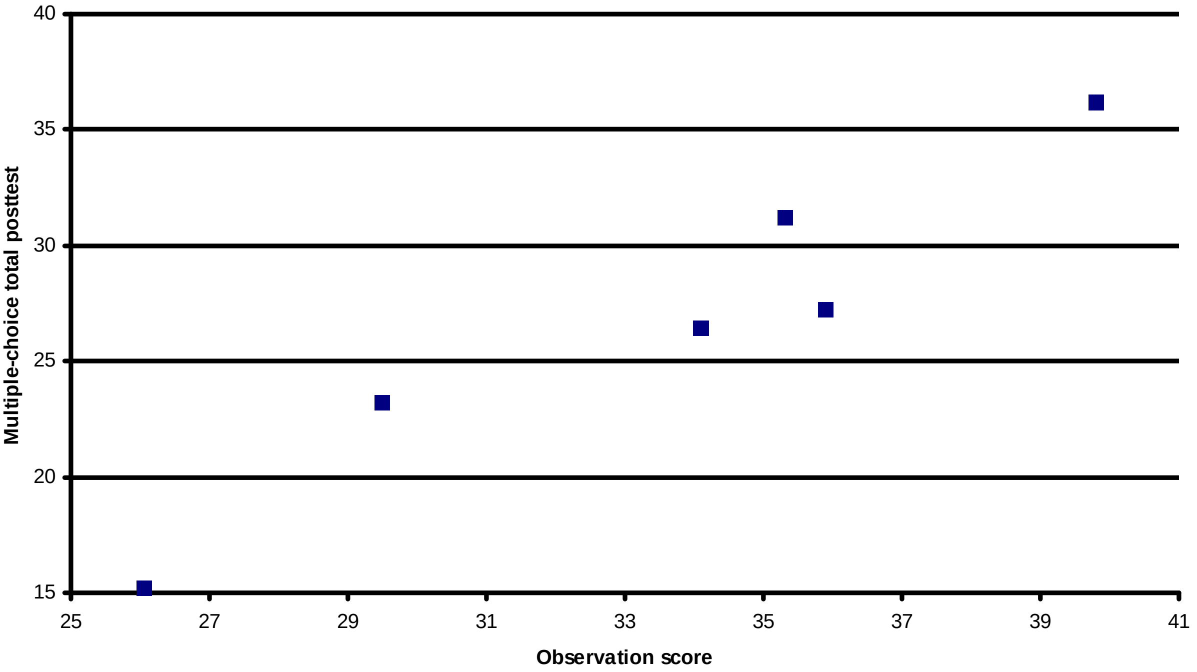


Figure IV-1. Scatterplot of ISOCS observation scores by mean ISSA multiple-choice posttest score (*r* = .96) for the six teachers for whom both types of scores were available.

because of the small number of teachers for which we have both ISOCS and student achievement results; nevertheless, its extraordinary magnitude and the clear relationship of the variables that is seen in Figure IV-1 provide evidence of the validity of the observation data.

INQUIRY SCIENCE TEACHER QUESTIONNAIRE VALIDITY STUDY

Validity has to do with the adequacy of inferences made from data (Kane, 2006; Messick, 1989). The validity aspects that we examine apply most directly to the items comprising scales.

For the purpose of having questionnaire results for conducting validity analyses, we collected original data in several ways described in this section. They were (a) administering the entire ISTQ to a sample of current teachers of the FAST program, (b) administering a reduced number of items in the Inquiry Science Implementation Scale twice to a larger sample of current or recent FAST teachers, and (c) developing and administering a teacher log that included some of the ISTQ items to a subset of the teachers who had completed the ISTQ. From these data, we conducted reliability analyses, including internal consistency analyses, test-retest analyses, and generalizability theory analyses; factor analyses of several scales; and concurrent validity and criterion-related validity (i.e., correlational) analyses.

Data Collection

The data that we collected for conducting validity studies for the ISTQ included data on the full instrument; data on a subset of the items of the instrument, administered in a test-retest study; and data on a teacher log, administered for collecting data to conduct concurrent validity analyses.

ISTQ Data Collection

For some of our validity analyses, we distributed the entire ISTQ to current FAST teachers in Hawai‘i and across the mainland in the second half of the 2004–05 school year. We first prepared a list of 948 teachers in the U. S. who had been trained in FAST 1 since 1997. We contacted by telephone, or attempted to contact, these teachers and asked if they would be willing to complete our log each time they finished a FAST investigation and the ISTQ one time only. We promised them a $30 bookstore gift card if they completed the questionnaire and a $5 bookstore gift card for each log that they completed. Of the 948 teachers, 183 agreed to participate and stated that they 380 would not participate. The remaining 475 included 14 teachers who were not teaching FAST, 257 who did not respond after a minimum of five telephone calls and a follow-up postcard, and 104 who did not respond after fewer telephone calls but were not recruited as intensively because they were trained at least seven years previously and were deemed too difficult to reach.

Internet data collection occurred over several months in the winter and spring of 2005. We began by distributing an e-mail message to each of the 183 teachers who had agreed to participate, providing them with the World Wide Web address for the ISTQ and instructions about completing the questionnaire.[9] We promised each teacher a $30 bookstore gift certificate for completing the instrument and mailed it to them within days after receiving completed questionnaires. We distributed up to three follow-up e-mail messages to teachers who did not complete the instrument after our first request. After three reminders with no response, the project Principal Investigator sent a message to the non-responding teachers; after still not receiving responses, we called teachers by telephone as many as three times each. Of the 183 teachers who agreed to participate, 79 eventually completed questionnaires. They comprised 14% of the 563 teachers who we contacted. This was clearly a smaller percentage than desirable, perhaps because it was offered online only, but it is comparable with market-research surveys that similarly rely on volunteer respondents.

We analyzed the demographic characteristics of the 81 teachers and compared them with the characteristics of the population of K–12 teachers nationwide that was described by the National Center for Education Statistics (2005). As seen in Table IV-3, the sample is fairly representative but includes more male teachers and more private school teachers than teachers nationwide. (FAST has traditionally attracted teachers from private institutions, and middle-school science teachers are likely to consist of a higher percentage of men than the entire K–12 pool.) The class size of the sample is also somewhat smaller. The difference between the sample

Table IV-3
Comparison of ISTQ Sample with K–12 Teachers Nationwide

Descriptor	Sample	K–12 teachers nationwide
Percent female	62	79
Percent that taught in public schools	70	89
Mean age	42	46
Percent that had Master’s degree	58	56
Median *N* years teaching in K–12 schools	12	14
Mean salary	$40,000–50,000	$43,262
Mean *N* students per class	28	22

[9] Having the instrument online allowed us to ensure that the teachers responded to every item before submitting the completed instruments, resulting in a dataset with no missing data—an outcome that would not have been possible had a paper version of the instrument been distributed and collected.

and population characteristics might reflect the type of schools in which inquiry science has been implemented.

For archival purposes, descriptive statistics, including (a) *N*s, means, standard deviations, and standard errors of the mean, or (b) frequency and percentage distributions, as appropriate, for all the ISTQ items are shown in Appendix F. In addition to being useful for archival purposes, many of the items are useful for examining FAST implementation and context for a small sample of FAST teachers (Brandon et al., 2006a, 2006b). The results for many of the items are not useful for validity analyses, however, except to note that the items were able to capture the data appropriately.

Test-Retest Data Collection

To collect data for a test-retest study, we prepared an instrument and distributed it by U.S. mail twice in the spring of 2006 to the group of 183 teachers who had initially agreed to complete the ISTQ. The test-retest instrument included 22 of the set of 26 items on the ISTQ's Inquiry Science Implementation scale; four of the items on the original ISTQ Implementation Scale were not included on the test-rest instrument because they were deemed to be particular to FAST. We provided incentives in the form of bookstore certificates for completing the instrument. A total of 156 FAST teachers responded to the first of the two administrations of the test-retest questionnaire; of these, 111 completed the instrument a second time. This group comprised the test-retest study sample. The mean number of days between the first and second occasions of completing the instrument was 28 (minimum = 6, maximum = 98, st. dev. = 18.5).

Log Data Collection

To collect data for conducting a concurrent validity study of the ISTQ, we developed a teacher log.

When beginning to develop the log, we reviewed the literature (e.g., Ball, Camburn, Correnti, Phelps, & Wallace, 1999; Camburn & Barnes, 2004; Mullens & Kasprzyk, 1996; Rowan, Harrison, & Hayes, 2003). Ball et al., Camburn and Barnes, and Rowan et al. tested logs that teachers completed about the instruction delivered to specific individual students. All used extensive logs requiring considerable reporting time, and all compared in-class observer results on the log with teacher results. Ball et al. collected 29 logs from seven teachers in their pilot study of a Web-based system; the generalizability of the results of their study is more limited than the generalizability of the results of the other two studies. They compensated the teachers $100 each, and the teachers were highly-motivated volunteers. They concluded that teachers need to be given strong incentives to complete logs. Teachers had some problems understanding the wording of the log. There were many "special situations" that made it difficult to record the activities of individual students. The agreement rate between teachers and observers was about 75%.

Camburn and Barnes conducted a log validation study of 31 teachers. Eight researchers observed the teachers in the classroom; both the researchers and the teachers completed their logs at the end of the day. Teachers and observers gave identical answers about half the time. Teachers tended to apply common-sense definitions to terms instead of attending to the definitions provided in the glossaries. The broader the activity, the higher the inter-rater agreement. The more frequent the instructional activities, the higher the agreement. Mullens and Kasprzyk compared seven teachers' log reports on nine items about broad instructional activities with their questionnaire responses and found that their agreement on the two instruments (with agreement defined as within one scale point) was 100% on two items, 86% on four, 71% on one, and 57% on two. Rowan et al. analyzed data from 19,999 logs completed by 509 teachers and reported acceptable levels of teacher accuracy (observer-teacher agreement was above 80% on about half the items, between 70% and 80% for two-tenths of the items, and below 70% on three-tenths of the items) after teacher training and with a hotline available for questions.

Together, the results of these studies suggest that (a) log items can be used to validate questionnaire items (Mullens & Kasprzyk, 1996), (b) teacher logs about individual student activities can present a host of difficulties, but validity can be at acceptable levels, (c) validity is greatest for instructional activities occurring frequently in the classroom, and (d) wording on the logs must be simple and unambiguous. We addressed these issues when developing our log.

Our primary purpose in developing the log was to have an instrument for correlating implementation items with some of the ISTQ Implementation Scale items. We decided to ask about both teacher activities and a few student activities in the log. We selected five items from the ISTQ Implementation Scale that addressed student-teacher interaction during student science investigations. The items asked about the extent to which (a) the teacher's students ask questions about the scientific phenomena that were addressed in the investigation, (b) the teacher uses questioning strategies to respond to students' questions about the investigations, (c) the students engage in discussions among themselves about the investigation, (d) the teacher circulates and interacts with students during the lab portion of the investigation, and (e) the teacher discusses variations in the data with the students in the summary phase of the investigation.[10] We pilot-

[10] Other topics addressed in the log and, parenthetically, the aspects of implementation or context that they address included (a) disruptions by activities inside or outside the classroom (context), (b) the number of class periods it takes to complete the investigation (exposure), (c) the customization of the investigation by using supplemental materials or any other method (adherence), (d) the adequacy of materials and equipment (context), (e) students' questioning behaviors (participant responsiveness), (f) the teacher's use of questioning strategies (adherence), (g) the teacher's circulation about the classroom (adherence), and (h) the teacher's discussions about variations in the data

tested the log using a quasi-cognitive interview method with a sample of four existing FAST teachers, one on two occasions. Based on the cognitive interview results, some revisions were made.

Teachers were asked to complete several logs each, one for each FAST 1 investigation that they conducted in the second half of School Year 2004–05. A total of 74 teachers completed logs at least once; the mean number completed was 3.8 (st. dev. = 3.40; range = 16). Of the 74 teachers, 66 also completed the ISTQ.

Validity Analyses

We conducted four kinds of validity analyses of the ISTQ data, which we introduce here and explain in detail in the remainder of the ISTQ section of the report:

1) We conducted content validity analyses that include (a) a review of the procedures for representativeness, relevance, and technical quality (Messick, 1995) and (b) two kinds of reliability analyses. For analyzing reliability, we first conducted factor analyses of items that we defined as four context scales that previous research has shown are aspects of school capacity for learning (the Collaboration Frequency Scale, the Collaboration Benefits Scale, the Teacher Participation in Science Activities Scale, and the School Support for Inquiry Science Scale) and of the 26-item Inquiry Science Implementation Scale. The purpose of the factor analyses was to examine whether we could analyze the items as scales.[11] Second, we calculated internal-consistency reliability statistics for these scales. Third, we conducted a test-retest analysis of the Inquiry Science Implementation Scale total scale score for 22 of the original 26 items. The factor analysis results provided construct validity evidence, as well.
2) We conducted two kinds of concurrent validity analyses. First, we conducted concurrent validity analyses for the four context scales. Second, we conducted concurrent validity analyses of the five items that the log and the ISTQ Inquiry Science Implementation Scale had in common. These are analyses of items addressing the adherence aspect. Third, we conducted concurrent validity analyses of subscales of the Inquiry Science Implementation Scale (identified in a factor analysis) by comparing the results for the subscales with results from the ISOCS and with the results from the ISQQ method. These are analyses of the relationship between two scales addressing the adherence aspect of implementation (from

(adherence).

[11] Item 17 in Section A of the ISTQ asked about the availability and adequacy of equipment and materials in the classroom, but the 13 questions about each topic were deemed too many for conducting factor analyses. This conclusion was confirmed in factor analyses and parallel analyses (O'Connor, 2000)..

the ISOCS and the ISTQ) and one set of items addressing the quality aspect of implementation (from the ISQQ).

3) We conducted criterion-related validity analyses of the relationship between total scores on subscales of the Inquiry Science Implementation Scale and scores from the ISSA.

Content and Construct Validity Analyses

Our description in this chapter of how we identified the variables that the ISTQ addresses and how we prepared the ISTQ items provides evidence of content validity. Content validity addresses "content relevance, representativeness, and technical quality" (Messick, 1995, p. 745). Supportive evidence for content validity is found in our description of the development of the questionnaire, including the description of (a) the extensive literature reviews and consultation with FAST experts that formed the basis for selecting item content, (b) how we helped ensure item quality by borrowing from national surveys when appropriate,(c) how we revised our items several times and had our FAST experts and advisory board review them, and (d) how we conducted think-aloud protocols, in which we asked pilot-test respondents to describe their mental processes while they answered the items.

The results of internal consistency reliability, test-retest reliability, and factor analyses provide further evidence of content validity. The factor analyses also provide evidence of construct validity.

Factor Analyses and Internal Consistency Analyses

The first step in examining our five subscales (the Collaboration Frequency Scale, the Collaboration Benefits Scale, the Teacher Participation in Science Activities Scale, the School Support for Inquiry Science Scale, and the Inquiry Science Implementation Scale) was to conduct exploratory factor analyses. For the context scales, we used the responses of the 79 teachers to the ISTQ. For the Implementation Scale, we used data from the 156 respondents to the test-retest version of the questionnaire, as well as the data from nine who responded to the first version of the instrument but not the test-retest version. For the context scales, which comprised one factor each, we conducted principal component analyses; for the Implementation Scale, we conducted a common factor analysis. (Common factor analyses, which produce latent variable scores, are possible only if there are two or more factors.) For each scale, we also examined internal consistency with coefficient alpha.

Seventy-nine respondents is not a large number for conducting factor analyses. However, recent research suggests that small sample sizes might be acceptable if communalities (i.e., the percent of variance in the item accounted for by the component or factor) are above .70 and if there is a sufficient number of variables (i.e., items) per factor (MacCallum, Widaman, Zhang, & Hong, 1999). Furthermore, parallel analyses can provide evidence about the appropriateness of conducting a factor analysis for a given set of data. In parallel analyses, the mean of eigenvalues is calculated for a large number of randomly generated data sets with the same number of respondents and variables as the actual data set; if the eigenvalues from exploratory factor analyses of the actual data exceed the mean eigenvalues generated in the parallel analysis, the appropriateness of conducting the factor analyses data is confirmed. Accordingly, we examined the communalities, the number of variables, and the results of the parallel analysis for each of the scales that we developed using factor analysis.

The number of items comprising the context scales ranged from three to five. We theorized that the items in each scale formed single factors. We accepted factors if they were interpretable and if the items loaded high on the factor (i.e., a simple structure was found). We gave the most weight to the first of these two criteria. We did not use the eigenvalue > 1.0 criterion, which is common in much research, because recent literature (e.g., Velicer, Eaton, & Fava, 2000) argues against it (although all the eigenvalues in our analyses were > 1.0). We show the number of items, factor loadings, communalities, the results of the parallel analyses, and the alpha coefficients in Tables IV-4 through IV-7.

Context scales. As seen in the four tables, the number of items for each of the context scales was small, which tends to limit reliability and common variance. However, the parallel analysis results for all four factors indicate that it the factor analysis results were appropriate.

Table IV-4
Principal Component Analysis Results for the Collaboration Frequency Scale[a]

Item	Loading
Section B, No. 5a. Frequency with which science teachers at your school meet to discuss classroom management or disciplinary issues.	85
Section B, No. 5b. Frequency with which science teachers at your school meet to discuss inquiry science teaching methods.	83
Section B, No. 5c. Frequency with which science teachers at your school meet to discuss science content issues.	86
Section B, No. 5d. Frequency with which science teachers at your school meet to discuss administrative issues.	84
Section B, No. 5e. Frequency with which science teachers at your school meet to discuss staff development issues.	90

[a]Communalities = .68, .80, .80, .66, and .78. Eigenvalue = 3.81, exceeding the parallel analysis eigenvalue. Coefficient α = .94.

Table IV-5
Principal Component Analysis Results for the Collaboration Benefits Scale[a]

Item	Loading
Section A, No. 9a. Frequency of collaboration with fellow teachers at my school gives me access to new ideas and knowledge.	75
Section A, No. 9e. The collaboration I have with fellow teachers at my school, ultimately improves my students' learning.	84
Section A, No. 9l. The collaboration I have with fellow teachers at my school improves my teaching	89

[a]Communalities = .56, 70, and 80. Eigenvalue = 2.06, exceeding the parallel analysis eignvalue. Coefficient α = .88.

The factor loadings and the communalities were most favorable for the two collaboration factors and less favorable for the scales on school support and teacher participation in science. These results were mirrored in the alpha coefficients. We believe that all the results suggest that the items can be considered as scales with the caveat that the total scores for the items of the collaboration scales are more reliable than the total scores for the items of the other two scales. Thus, the results provide content validity evidence and construct validity evidence, although less strong for two of the scales than for the other two.

Implementation scale. As part of our test-retest study, we collected implementation data

Table IV-6
Principal Component Analysis Results for the Teacher Participation in Science Activities Scale[a]

Item	Loading
Section B, No. 6a. To what extent do you read science magazines, science journals, or science books outside of the classroom?	55
Section B, No. 6b. To what extent do you attend science teaching conferences or meetings outside of the classroom?	70
Section B, No. 6c. To what extent do you hold leadership positions in science teaching organizations?	53
Section C, No. 12. How many hours have you spent in science-teacher professional development classes (other than FAST) over the past five years?	61

[a]Communalities = .30, .49, .28, and .37. Eigenvalue = 1.45, exceeding the parallel analysis eignvalue. Coefficient α = .71.

Table IV-7
Principal Component Analysis Results for the School Support For Inquiry Science Scale[a]

Item	Loading
Section A, No. 9g. My school has adequate funding for FAST books and materials.	68
Section A, No. 9h. My school ensures that I have sufficient opportunities for professional development in teaching science.	72
Section A, No. 9i. My principal supports teaching FAST in my school.	82

[a]Communalities = .45, 51, and 67. Eigenvalue = 1.63, exceeding the parallel analysis eignvalue. Coefficient α = .81.

from 156 teachers who responded to the first administration of the test-retest questionnaire. We conducted factor analyses of the data collected from these 156 teachers, combined with the data for the nine teachers who did not respond to the test-retest instrument but had responded to the first version of the ISTQ. The total *N* for the factor analyses = 165. This number is preferable to using only the 79 respondents from the original ISTQ data collection because of the small item-respondent ratio (about 1:3) for that data collection.

Our initial factor analyses showed that some of the items did not load on the factors; therefore, we conducted additional factor analyses until we narrowed the set to those that loaded. This resulted in three factors, as shown in Table IV-8. We have labeled Factor 1 as *Teacher-Student Interaction*, Factor 2 as *Connecting Science to the World Outside School*, and Factor 3 as *Introducing the Investigation*.

As seen in Table IV-8, the eigenvalues for Factors 2 and 3 are small, and the parallel analysis result supports a one-factor solution only. The coefficient α shows adequate internal consistency reliability. We report all three factors, despite the results of the parallel analysis and the small number of variables (three) for each of the Factors 2 and 3, because all the factors are interpretable and because we wish to examine the relationship of each these adherence factors with the results for other variables in our study.

Test-Retest Reliability

Factor analysis results and internal consistency analysis results can support data reliability, but test-retest analyses provide more convincing evidence. We analyzed the test-retest results from the 111 respondents to the 22 implementation items on the log. Coefficient α for the first

Table IV-8

Common Factor Analysis Results for the Inquiry Science Implementation Scale[a]

Item ***When you teach science, how frequently do you:***	Loading, by factor		
	F. 1	F. 2	F. 3
c. review relevant concepts and skills that were learned in previous lessons? (.41)	26	1	46
d. introduce new vocabulary words? (.62)	-6	8	78
e. ask students to identify and define words? .(50)	1	-6	73
h. discuss how everyday situations directly relate to experiments that students are currently, or will be conducting? (.51)	8	63	7
j. monitor small group progress during experiments? (.47)	76	-14	0
k. encourage students to collaborate within their groups? (.45)	68	-6	4
l. circulate and interact with students while they are conducting experiments? (.40)	61	5	-3
m. discuss variations in data collected by students following their experiments? (.53)	66	13	-2
o. have students share their data or findings with the class? (.39)	63	-3	2
p. challenge students to consider the effects of errors on groups' results? (.56)	63	20	-4
q. compare and contrast students' explanations of findings? (.54)	60	11	11
s. connect new information with students' personal lives (interests, home environment, community, culture, etc.)? (.68)	-4	88	-7
t. connect current events and other subjects with current science concepts, skills, and investigations? (.66)	1	78	5
u. use questioning strategies to respond to students' questions about experiments? (.34)	49	14	0
v. have students ask questions about the scientific phenomena addressed during experiments? (.45)	44	25	9

[a]N = 165. Communalities are shown in parentheses after item text. Factors are defined by loadings ≥ 40. The label for Factor 1 (defined by nine items) is, *Teacher-Student Interaction*; for Factor 2 (defined by three items) it is, *Connecting Science to the World Outside School*, and for Factor 3 (defined by three items) it is, *Introducing the Investigation*. Eigenvalues = 5.8, .92, and .80, accounting for 100% of the variance but with only the eigenvalue for Factor 1 exceeding the parallel analysis results. Coefficient α for Factor 1 = .87; for Factor 2, α = .84, and for Factor 3, α = .74.

administration of the test-retest questionnaire = .87; for the second administration, α =.89. The Pearson correlation between the total scores for the two instruments = .76. The results of a generalizability theory analysis showed that the proportion of variance due to items = 7%, the variance due to respondents = 10%, and the proportion of variance due to occasion = 0%. These results show that the Inquiry Science Implementation Scale data collected from our sample of FAST teachers are highly reliable.

Concurrent Validity Analyses

We conducted three concurrent validity analyses—one among context scales, a second between the ISTQ Implementation Scale and implementation results on the teacher log, and a third of total scores on the three ISTQ Implementation factors with the results of two other implementation measures that we developed and tried out in the study.

Correlations Among Context Scales

Data are valid in part to the extent to which the relationships among results for separate subscales on an instrument are as expected. To examine this issue, we calculated the Pearson correlations among the two collaboration scales, the Teacher Participation in Science Activities Scale, and the School Support for Inquiry Science Scale. Organizational learning research has shown that schools that excel in these variables tend to have a high capacity for learning and are more likely to be high-functioning schools.

The results are shown in Table IV-9. The correlations for three of the scales confirmed our expectations and show concurrent validity. The results for the two Collaboration Scales show a substantial correlation (.51) with each other and modest correlations (.31 and .32) with the School Support for Inquiry Science Scale. The result for the Teacher Participation in Science Activities Scale shows virtually no correlation with the collaboration or support scales. The participation subscale includes items addressing the extent to which the teacher reads science publications, attends science conferences or meetings, and holds leadership in science teaching

Table IV-9
*Correlations (*N *= 79) Among Four Context Scales*[a]

Scale	(1)	(2)	(3)	(4)
(1) Collaboration Frequency	—	—	—	—
(2) Collaboration Benefits	.51[b]	—	—	—
(3) School Support for Inquiry Science	.32[b]	.31[b]	—	—
(4) Teacher Participation in Science Activities	.06	.05	.00	—

[a] *N* = 79; confidence interval = .22.
[b] Significant at .005.

organizations, as well as an item about the number of hours they took science PD in the last five years. These activities are less controlled by the school than teacher collaboration activities, which are heavily influenced by school leadership. The School Support Scale is measured by three items about whether the school has enough funding for FAST, enough opportunities for science PD, and enough money for teacher PD.

The 95% confidence interval for correlations among 79 respondents is .22; therefore, the findings of this and other correlational analyses of scale results presented here should be considered suggestive, not conclusive.

Correlation of ISTQ Implementation Items with Teacher Log Results

To prepare to examine the concurrent validity of the ISTQ's Inquiry Science Implementation Scale, we totaled the ratings across the five log implementation items (α = .72). We then calculated the correlation between the total scores for these five items with the total score for the 26-item implementation scale. The correlation was .66 (significant at the .01 level; confidence interval = .25)—substantial evidence of the concurrent validity of our measurement of implementation.

Correlation of ISTQ Implementation Items with Observation Results and Teacher Quality Rank

In our third concurrent validity analysis, we correlated the total scores for each of the three Inquiry Science Implementation Scale factors (Teacher-Student Interaction, Connecting Science to the World Outside School, and Introducing the Investigation) with the results from two other measures of implementation—the ISOCS and the ISQQ. (Preparation of ISQQ validity data is described in the next section of this report.) The first of these two methods is a measure of adherence by observers. A reasonably strong correlation of the results on factors of the self-report Implementation Scale with ISOCS and ISQQ results would suggest ISTQ Implementation Scale concurrent validity. The ISQQ is a measure of the quality aspect of implementation; a reasonable correlation of results on the Implementation Scale—an adherence measure—with the results on the ISQQ quality measure also suggests concurrent validity. A limitation of these analyses is that we have ISOCS results and ISQQ results for only nine teachers; therefore, the correlations are only suggestive.

The Pearson correlations of the three Implementation Scale factors with the ISOCS and ISQQ results are shown in Table IV-10. As might be expected, the .50 correlation of the ISTQ Teacher-Student Interaction with the ISOCS teacher-student interaction results (i.e., the total of Codes D6 and D8 expressed as a percentage of all ISOCS codes for the teacher) is the highest shown between any of the ISTQ Implementation Scale factor scores and the results on the other instruments. This a strong correlation for this kind of analysis. However, as seen in Figure IV-2,

Table IV-10

Correlations Among Three ISTQ Implementation Scale Factors and Other Implementation Results

Scale	(1)	(2)	(3)	(4)	
(1) Teacher-Student Interaction (ISTQ Implementation Scale)[a]	1.0	—	—	—	—
(2) Connecting to Outside (ISTQ Implementation Scale)[a]	.79[a]	1.0	—	—	—
(3) Introducing Investigation (ISTQ Implementation Scale)	.56[a]	.45[a]	1.0	—	—
(4) Teacher-Student Interaction (ISOCS, D6 + D8)	.50[b]	.42[b]	.21[b]	1.0	—
(5) Teacher quality rank (ISQQ [Ch. IV])	.39[b]	.02[b]	-.02[b]	.36[b]	1.0

[a] *N* = 79; significant at .0001.

[b] *N* = 9.

an outlier is affecting the correlation positively. The .42 correlation of the Connecting Science to the World Outside School factor total score with the ISOCS results is also quite respectable. The correlation of the Implementation Scale Teacher-Student Interaction factor results with the ISQQ results (.39) is somewhat smaller; it suggests a relationship between teacher-student interaction and the extent to which the teacher uses questioning appropriately, but, as seen in Figure IV-3, an outlier probably accounts for some of the correlation. The lack of relationship of the other two factors with questioning quality is not surprising, because the focus of the ISQQ was strictly on the use of questioning strategies, which is reflected in the Teacher-Student Interaction items but not in the items of the other two factors. Altogether, these results for concurrent validity evidence for the ISTQ Implementation Scale are mixed.

Criterion-Related Validity Analyses

Our final validity analysis for the ISTQ examined the relationship between the three Implementation Scale factors and results on the ISSA, which was administered to the students of 10 teachers. We examined the relationships of the total scores for each of the three Implementation Scale factors with the total scores for the multiple choice items of the student assessment and the extended-response (i.e., written-response) items of the assessment. The correlations of the three factors with pretest and posttest results, shown in Table IV-11, provide some evidence of a relationship between the Teacher-Student Interaction factor with the assessment results and

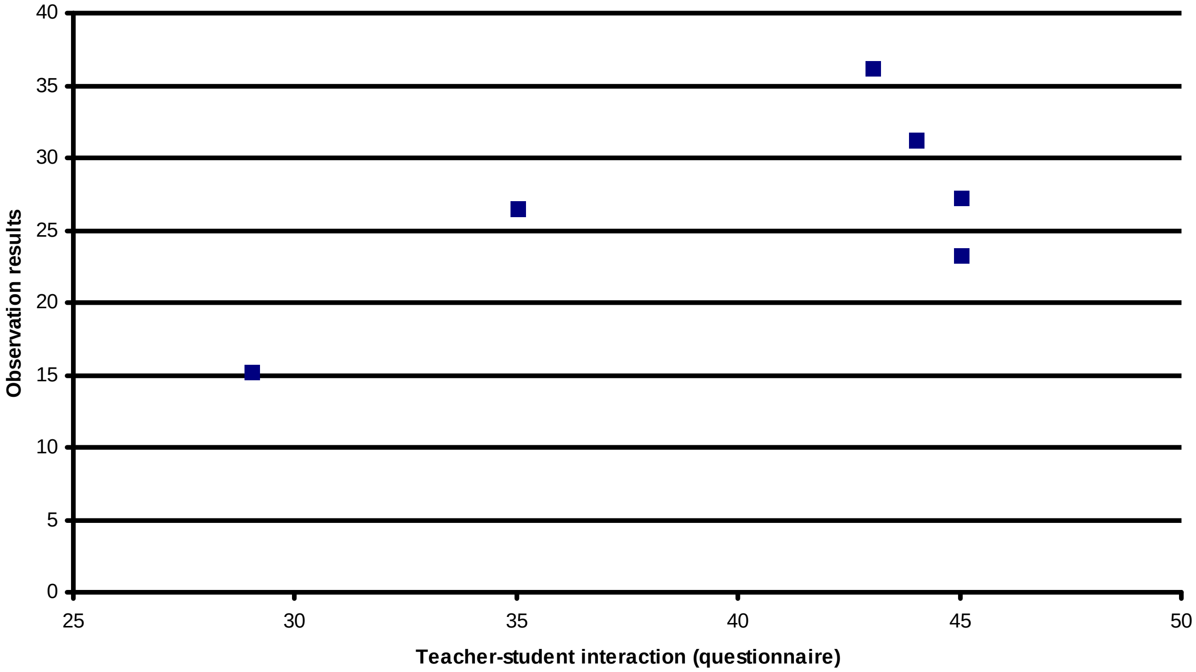


Figure IV-2. Scatterplot of ISTQ Teacher-Student Interaction factor score by ISOCS observation scores (r = .50) for the six teachers for whom both types of scores were available.

some relationship of the Introducing the Investigation factor with the assessment results, but no relationship of the Connecting to the Outside World factor with the assessment results. The scatterplot given in Figure IV-4 for the extended-response pretest scores and the Teacher-Student Interaction factor results, which show the strongest correlation in Table IV-11, suggest some evidence of outliers, although not as much as in Figures IV-2 and IV-3. These findings provide tentative evidence, because of the small *N* and because the statistical foundation for the second and third Implementation Scale factors is equivocal, as explained above. Nevertheless, the results favor the validity of the ISTQ Implementation Scale data.

INQUIRY SCIENCE QUESTIONING QUALITY METHOD VALIDITY STUDY

Data Collection

Selection of a Sample of Teachers and Observation Segments to Judge

The first step in our validity study of the ISQQ data was to select a sample of videotaped FAST student investigations (which had been transferred to DVD-ROMs) for the judges to examine teachers' use of questioning during inquiry science classes. We chose the sample that was used for validating the ISOCS, as shown in Table IV-1. Segments of DVD-ROMs were selected for viewing and judging in paired comparisons. A FAST scientist/educator with

Table IV-11

*Correlations of Teachers' Total Scores for Three ISTQ Implementation Scale Factors With the Teachers' Classroom's Inquiry Science Student Assessment Multiple-Choice Mean Total Scores and Extended-Response Item Mean Total Scores (*N *= 10)*

Scale	Multiple.-choice pretest	Multiple-choice posttest	Extended-response pretest	Extended-response posttest
(1) Teacher-Student Interaction (ISTQ Implementation Scale)	.28	.37	.43	.32
(2) Connecting to Outside (ISTQ Implementation Scale)	-.06	.04	.11	-.02
(3) Introducing Investigation (ISTQ Implementation Scale)	.37	.28	.39	.20

expertise in videotape editing reviewed all the sampled DVDs and sampled approximately 15 minutes of the three phases of FAST student investigations (the Introduction Phase, the Investigation—conducting the experiment—Phase, and the Interpretation Phase), for a total of approximately 45 minutes per teacher. The goal was to sample segments of 15 contiguous minutes per phase, although in a few instances up to three segments were sampled per phase, particularly in the Investigation Phase such as when equipment gathering and clean up interrupted the teacher-student interaction. The sampled segments were copied on laptop hard drives for the judges' use later in the study. Samples for judge training also were identified and copied to laptop computers for group or individual viewing during the ensuing training.

Judge Recruitment

The next step in the validity study was to recruit five judges—a number deemed more than sufficient for obtaining reliable results and feasible within the fiscal resources of the project. The judges were male FAST experts from five states who had taught FAST and served as FAST trainers.[12] All agreed to participate in the three-day study immediately following a three-day FAST training workshop on another topic. They were compensated for airfare, local travel, and lodging and given a taxable stipend of $1,000 each.

Administration of the ISQQ

The ISQQ was administered over a three-day period with the entire group of judges in a University of Hawai'i at Mānoa Campus Center conference room on Day 1, individually in the

[12] Historically, all teachers are required to be trained in a two-week workshop before their schools can purchase FAST materials. FAST teacher trainers are experienced FAST teachers who receive additional training.

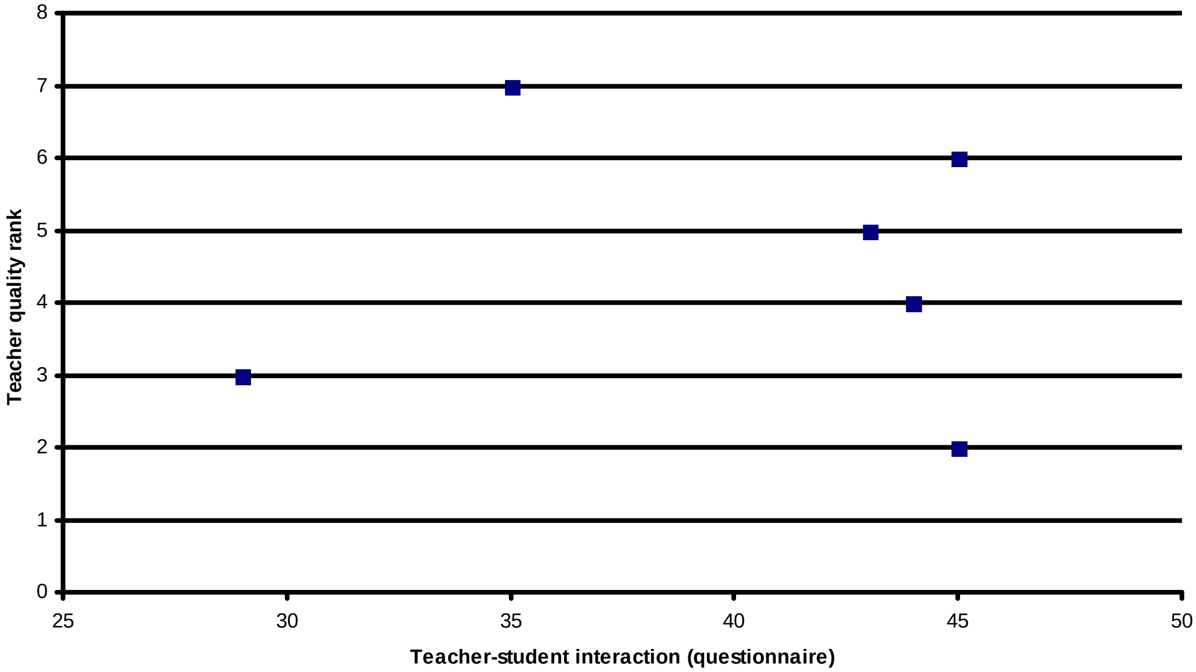


Figure IV-3. Scatterplot of ISTQ Teacher-Student Interaction factor score by ISQQ teacher quality rank (*r* = .39) for the six teachers for whom both sets of scores were available.

judges' hotel rooms on Day 2 and the morning of Day 3, and at another University of Hawai'i at Mānoa conference room on the afternoon of Day 3. It consisted of four steps. First, on Day 1, the judges were introduced to the ISQQ. They reviewed and, with slight revisions, validated the statement of quality questioning criteria that had been prepared during the development of the ISQQ. Second, they were trained in how to use the statement to make and record preference votes among pairs of teachers. Third, on Day 2 and the morning of Day 3, they independently observed the videotape segments for each teacher. Fourth, on the afternoon of Day 3, the judges reconvened and made paired-comparison judgments. Throughout the meetings of Days 1 and 3, the group facilitator (the project Principal Investigator) endeavored to lead the group in such a manner so as to ensure that all participants had opportunities to express their views and that the FAST developers contributed without dominating the discussion. Each of these four steps is described in depth in this section.

Introduction and Validation of the Quality Criterion Statement

Day 1 began with an introduction to the study and the training of the judges. The conference room was equipped with a laptop computer, computer projector, and a screen for group viewing, as well as individual laptops for the judges' use. The participants received folders describing the study, and the study administrators used the training and administration manual. A

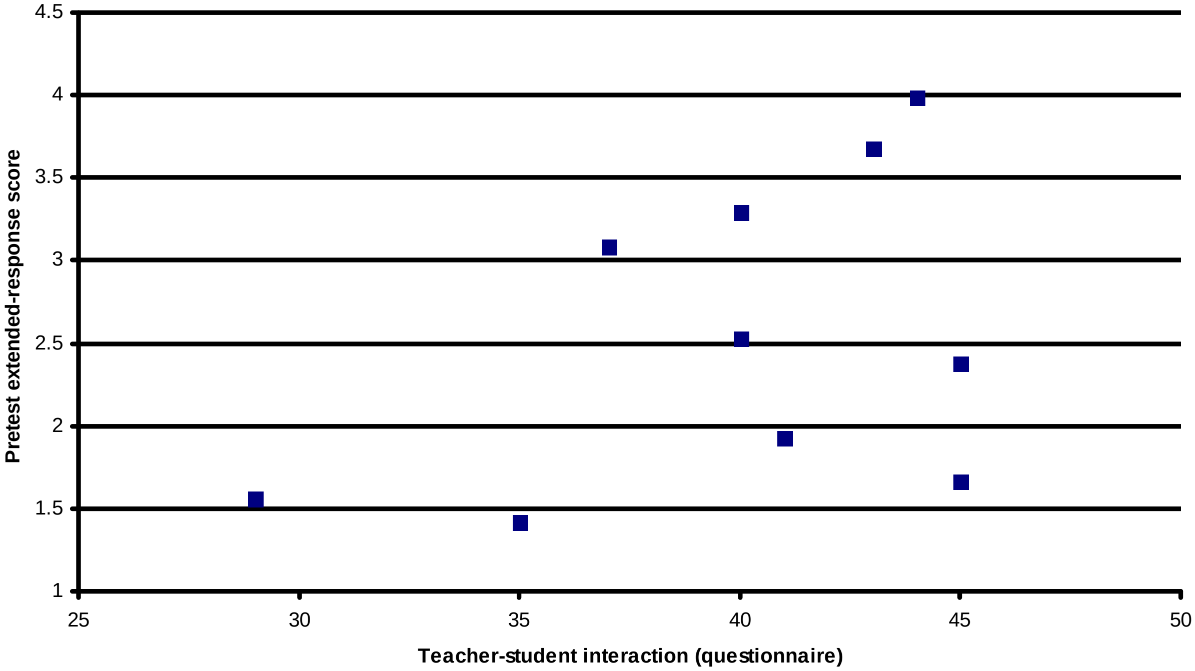


Figure IV-4. Scatterplot of ISTQ Teacher-Student Interaction factor total score by mean ISSA pretest extended-response score (*r* = .43) for the 10 teachers for whom both types of scores were available.

flip chart was used to record comments when appropriate. The project Principal Investigator served as the primary trainer; the project manager/project researcher and the project co-Principal Investigator, who was one of the two lead FAST developers, participated in the discussion among the judges.

The workshop began with a description of the purpose of the study, of the overall NSF project, and of the place of the study within the overall project. The FAST classroom observations that had been conducted were briefly described, and the judges were told that they were to try out a method for judging teaching quality in inquiry science classrooms. The facilitator explained that they were selected because of their expertise in FAST and described the steps that would occur during the remainder of the workshop. He also explained that this was the first time that the team had used this method and that the judges could help refine the method; this was stated in part because this was the case and in part to help the judges feel at ease and participate fully. The judges read the statement of quality questioning criteria silently; then the facilitator presented it on a computer slide and asked the group about revisions, omissions, or additions that might be made to the statement. The group briefly discussed the statement, revised it slightly, and agreed that it was accurate and appropriate as a description of quality questioning.

Judge Training

The judges were trained in how to apply the quality criteria to videotapes of teachers in two steps. First, as a group they viewed a 15-minute recording, projected on a screen, of a teacher who exhibited what we estimated to be mid-level questioning quality. They were instructed to look for behaviors and events that reflected quality or a lack thereof. Then each member of the group was asked to present their opinions, one at a time and without interruption, about the quality of questioning that the teacher displayed. Next, the group discussed each other's opinions. One member judged the quality somewhat differently from others because he was weighting some aspects differently; the group discussed this difference and the outlying judge agreed to modify his approach.

The second step of the training was to have the judges view another15-minute video segment individually on laptops with headphones. They took notes and viewed the tapes without discussion. Then the judges again presented their opinions about the level of quality exhibited on the video. The differences among the judges' opinions about levels of quality varied little.

Our preference in this phase of the training was to have the judges view at least three video segments individually and to compare each teacher with each other teacher. However, because there was a limited number of segments of good quality (not including the tapes that were to be judged later), we were unable to view more tapes.

When the viewing was complete, the meeting concluded at about midday with instructions for the tasks for the second phase of the process: They were told that

1. their task until noon of Day 3 was to view three 15-minutes samples for each of nine teachers.
2. they were to work at places of their choosing except where laptops might be damaged.
3. they were not to discuss any of their work with each other.
4. they should take notes about the extent to which each teacher addressed the quality criteria.
5. their notes should address all aspects of the criteria.
6. they should write summary statements for each teacher and comparison.
7. they should review the teachers as much as necessary to make global judgments of quality.
8. they would reconvene after lunch on Day 3 to formally make judgments about quality using a method that would be described at the time.

The judges were shown how to access the tapes on their laptops and were given the necessary additional equipment and supplies (headset, cords, tablets and pens, and contact phone numbers for asking any unanswered follow-up questions).

Making the Paired Comparison Judgments

In the early afternoon of Day 3, the judges reconvened in a University conference room for the final session. They brought the notes that they made when viewing the videotape segments and were provided with judgment recording forms. The paired-comparison method was then described in detail. It was explained that the method can be used to compare a set of “objects” on any attribute and that it produces an interval-level scale of the objects. It was contrasted with ratings, and an example of using the method was presented. The judges were told that, referring to their notes, they would compare each teacher with each other teacher and judge (a) which member of each pair showed greater quality than the other member of the pair and (b) on a scale of 1 to 7, the similarity of their quality. (The similarity results are not reported here.) Questions were fielded. Finally, the judges made the paired comparison judgments (using forms on which the teachers were randomly sorted in a different order on each form), which took from about 15 minutes to one-half hour.

Judges’ Feedback About the Process

At the conclusion of the meeting, the judges were asked for their feedback about the process. They reported several conclusions:

1. Viewing the two training samples was sufficient to feel comfortable about assessing quality.
2. Viewing the videotaped segments took from one to two hours per teacher.
3. Entering the notes into computer files while viewing the video segments did not complicate note-taking; the judges alternated between viewing and videotaping.
4. One judge stated that he found it difficult to summarize quality across the segments for the three FAST investigation phases; another found that having three segments ensured that he had a good sense of whether the teacher used good questioning strategies.
5. The judges tended to apply some additional criteria such as the extent to which the teacher waited long enough for answers to questions and whether the teachers missed questioning opportunities. One judge said that he had to continue to return to the statement of quality criteria because he had additional criteria of his own in mind when viewing the videotape segments. Another offered additional statements to include in the quality statement. A third judge reported that he found it difficult to focus only on teacher questioning. For example, at first he tended to look at the children’s behavior. One remedy was to listen but not to watch. Another tended to look at the students to see if they were engaged.
6. A judge stated that he thought it would have helped if the viewing had been organized by FAST investigation. Another stated that viewing different lessons by different teachers did not complicate the judgments of quality.
7. It was suggested that the criteria be identified with labels or keywords to help the judges keep the criteria distinct from each other and in mind while judging.

8. One judge stated that he saw many characteristics of good teaching in all the teachers and another stated that he saw a lot of bad teaching.
9. A judge stated that having more than nine teachers to view and assess would have been onerous.

Validity Analyses

Preparing and Scaling the Data

The first step in the analyses was to prepare the preference datasets. The judges' preference responses on the paper questionnaires were recorded on an electronic spreadsheet with a teacher-by-teacher matrix for each judge. The cell entries showed the number of the teacher (row or column) who was preferred over the other. All cells in the square matrixes were filled except for the diagonals, which were left blank.

The second step was to total the preference votes and rank the totals. The preference data in each of the judges teacher-by-teacher matrixes, which were prepared in the first step, were transformed to ones and zeros, with *1* entered if the teacher in the column heading was preferred over the teacher in the row heading and *0* entered if the teacher in the row heading was preferred over the teacher in the column heading. The diagonals were assigned the value of *.5*. The cells were then totaled across the five judges, resulting in one matrix. The columns of this matrix were totaled, and the results were ranked. The ranks formed a set of scale data that we call Analysis Dataset No. 1. These are shown in Table IV-12.

The second set of scale data that was analyzed was formed using the Thurstone Case 5 paired comparison scaling method (Dunn-Rankin, Knezek, Wallace, & Zhang, 2004; Edwards, 1957), which produces a scale with unequal distances among the scaled objects. Using the Thurstone method, the cell totals in Analysis Dataset No. 1 were converted to proportions of the total possible number of votes (five); then the columns were summed and were sorted on the sums. The cells were then converted to normal deviates (.5 = 0), and the columns were summed. The distance between column sums was calculated; these distances resulted in scale values. For ease of interpretation, we transformed these values to a scale with a minimum value of 0.0. The scale is shown in Table IV-13, which we call Analysis Dataset No. 2.

Finally, we prepared Analysis Dataset No. 3, which was a 5-judge by 36-between-teacher-comparisons matrix for Guttman scale analysis, with *1*s and *2*s in the cells.

Reliability of the Preference Data

Data cannot be valid unless they are reliable. We conducted five reliability analyses of the results of the paired comparisons, each coming from a different measurement tradition. The coefficients we produced included Kendall's coefficient of concordance (*W*), Thurstone's absolute average discrepancy coefficient (Edwards, 1957; Gulliksen & Tukey, 1958), Guttman's

Table IV-12

Results of Paired Comparisons (Analysis Dataset No. 1)

Teacher	Teacher								
	2	3	5	7	13	15	16	20	21
2	2.5	2	1	5	4	1	3	1	3
3	3	2.5	1	4	3	2	3	1	4
5	4	4	2.5	5	5	4	4	1	5
7	0	1	0	2.5	1	0	0	0	1
13	1	2	0	4	2.5	1	1	1	3
15	4	3	1	5	4	2.5	2	1	4
16	2	2	1	5	4	3	2.5	1	4
20	4	4	4	5	4	4	4	2.5	5
21	2	1	0	4	2	1	1	0	2.5
Total	22.5	21.5	10.5	39.5	29.5	18.5	20.5	8.5	31.5
Rank	4	5	8	1	3	7	6	9	2

[a] Each cell shows the total number of judges' preferences for the teacher in the column over the teacher in the row.

coefficient of reproducibility (Edwards, 1957), and the intraclass correlation coefficient (ICC), Model 2 (Shrout & Fleiss, 1979). For *W* and the ICC, we used the total row of Analysis Dataset No. 1; for the Thurstone analysis we used Analysis Dataset No. 2; and for the Guttman analysis, we used Analysis Dataset No. 3. We also calculated the average percent agreement. The results (shown in Table IV-14) are:

1) Kendall's *W*, corrected for ties, = .55. This is a measure of the degree of agreement among judges. Howell (2007) suggested translating *W* into Spearman's rho, because the latter is more interpretable. Our calculations shows that rho = .44, a value indicating modest reliability.
2) The Thurstone's absolute average discrepancy method produces a coefficient (.022) based on comparing empirical proportions with theoretically expected proportions; the lower the value, the better the result. The coefficient that we found (.022) is comparable to values reported by Edwards (1957) and suggests fairly high reliability.

Table IV-13

Thurstone Case 5 scale scores (Analysis Dataset No. 2)

Teacher	Thurstone scale score
20	0.00
5	0.05
15	0.69
16	0.82
3	0.96
2	0.95
13	1.52
21	1.71
7	2.43

3) Guttman's coefficient of reproducibility = .80. This value indicates the per cent accuracy with which responses to the paired comparisons can be reproduced from ranks (Edwards, 1957). Our result indicates good reproducibility.

4) The ICC, Model 2 (.48) is a measure of association among raters that takes into consideration the proportion of variance that raters have in common. According to Barrett (2001), Fleiss's (1981) and Cicchetti and Sparrrow's (1981) interpretations of a coefficient of this magnitude are that it indicates a fair level of reliability.

5) The average percent agreement = 63. This was calculated as the average agreement on teacher preference between each judge paired with each other judge. We interpret this as a fair level of agreement; a minimum of 80% would have been preferable.

These results clearly show favorable, albeit not uniformly high, levels of reliability. To examine further why the results were somewhat less consistent than desirable, we calculated the Spearman's rho correlations among the five judges' ranks of the nine observed teachers. The results ranged from .14 to .63, with the correlations for two of the judges with the remaining three judges clearly standing out as lower than the correlations of the three judges among each other.

Together, these results show discrepancies of two of the judges' results with the others. The most obvious reasons for the discrepancies are insufficient attention of the two judges to the

Table III-14
Results of Analyses of the Reliability of the Thurstone Case 5 Scale Values

Coefficient	Value
Kendall's coefficient of concordance (*W*) (rho = .44)	.55
Thurstone's absolute average discrepancy coefficient	.02
Guttman's coefficient of reproducibility	.80
Intraclass correlation coefficient (ICC)	.48
Average percent agreement	63

questioning quality criteria or insufficient differences among the teachers to reliably differentiate among them. The Thurstone Case 5 scale values, shown in Table IV-13, are potential evidence for the second reason. They clearly show small differences among the two teachers at the bottom of the scale (Nos. 20 and 5) and among three of the teachers in the mid-range of the scale (Nos. 16, 3, and 2). The finding about the small differences in the middle is consistent with evaluations of many types; our experience has shown that identifying the highest and lowest performing evaluands, whether they be persons, programs, or organizations, is usually a straightforward task that yields strongly defensible conclusions but that distinguishing reliably among evaluands in the middle is often difficult.

The paired comparison method is designed to ensure that distinctions are made between closely performing objects and that ties, such as might occur with ratings, are avoided. However, the differences among the quality of the questioning strategies might be too small in the sample of inquiry science teachers that the judges examined. This conclusion is supported by an analysis of the *circular triads* among the paired comparisons (Dunn-Rankin et al., 2004). Circular triads occur among paired comparisons when judges make decisions inconsistently—by indicating, for example, that Object 1 is preferred over Object 2 and Object 2 is preferred over Object 3 but that Object 3 is preferred over Object 1. Using Dunn-Rankin et al.'s TRICIR software program, we found a total of 16 circular triads in the judges' preference data. The circular triads were found for comparisons for each of two teachers in 11 of the circular triads. (Eight of the 16 circular triads were made by the judge whose Spearman's rho correlations with the remaining judges were the lowest of all five judges.) Clearly, it tended to be difficult for some the judges to be consistent in their comparisons among some of the teachers.

Content Validity

The content aspect of validity addresses "content relevance, representativeness, and technical quality" (Messick, 1995, p. 745). We believe that our description of the development of the method of the quality-judging procedures supports the content-related validity of the procedure.

The judges' feedback at the conclusion of the workshop, however, provides somewhat mixed evidence about validity. Evidence supporting validity includes the comment about the adequacy of the number of training samples, the comment about the ease of taking notes while making judgments, and the comment about the appropriateness of providing three videotaped segments of the work of each teacher. Evidence not supporting validity is found in the comment that it was difficult to make holistic judgments about quality. Other evidence that particularly does not support content validity is found in the comments by multiple judges about their tendency to add quality criteria of their own to those specified in the statement that the judges were instructed to use. These comments suggest that the judges' conceptualization of the task tended to be insufficiently well bounded. This might help explain the mixed findings about reliability.

Concurrent Validity

Concurrent validity evidence in the form of a correlation of the ISQQ results with the results for an alternative method for assessing questioning in FAST classrooms will help assure us that the data are valid. Preferably, the alternative method should assess questioning analytically, in contrast to the holistic approach of the ISQQ. For the purpose of validating the ISQQ data, we correlated the Thurstone Case 5 scale scores with ISOCS results for Codes C3 and D6/D8. The Spearman's rho correlation of the ISQQ Thurstone Case 5 scale scores with the percentage that student comments constituted of all teacher codes = .52, and the Spearman's rho correlation of the Case 5 scores with the percentage that the teachers' used follow-up statements and probing questions = .45. We believe that these correlations provide solid evidence of a substantial relationship between the two sets of results, thus supporting the validity of the data collected with the ISQQ.

INQUIRY SCIENCE STUDENT ASSESSMENT VALIDITY STUDY

Data Collection

A total of 428 students were administered both the pre-test assessment suite and the post-test assessment suite. Of those, 365 students had complete data sets. The validity analyses were completed on these students. The students who took the assessment suite were 7th or 8th grade students in schools located in Hawai‘i. All students were taught the FAST 1 materials.

Validity Analyses

We calculated coefficient alphas as evidence of content-related validity and correlations as evidence of concurrent validity. As seen in this section, the coefficient alphas were high and the concurrent validity correlations for the most part ranged from moderate to high.

Student Achievement Test

Students who took both the pretest and posttest showed improvement from pre- to post- at a statistically significant level (M_{pre} = 10.9, st. dev. = 4.7; M_{post} = 12.3, st. dev. = 4.9, t = 6.28, $p \leq .001$), suggesting that the instrument is sensitive to student learning in the FAST classroom. This is evidence for the content validity of the scores. The scores were lower in the Hawai‘i sample than in our earlier pilot-test sample, perhaps affecting the pre-post difference scores and some of the analyses of correlations with the scores on some of our attitudinal scales.

Student Performance Assessment

A total of 272 students completed the Rocky River performance assessment. Of those students, 134 completed all surveys, pre- and posttests as well as the performance assessment. Students who completed all the instruments achieved higher on the posttest than the entire group of students (respectively M_{post} = 12.3, st. dev. = 4.9 vs. M_{post}= 13.7, st. dev. = 4.5). The performance assessment total score and the posttest multiple choice test were correlated with each other (r = .34, p < .01). Since Shavelson and Ruiz-Primo (1996) reported correlations between performance assessments and multiple-choice tests of .45, we are confident in the concurrent validity of the performance assessment scores.

Students started the performance assessment in groups and then completed the assessment individually. Each student received a performance assessment group score reflecting his or her work in the group as well as a performance assessment individual score reflecting his or her individual culminating effort. The relationship between student performance assessment individual scores (PAIS) and the multiple choice posttest (MCPOST) scores was stronger than the relationship between the performance assessment group scores (PAGS) and the MCPOST scores ($r_{\text{PAIS•MCpost}}$ = .36, p < .01 vs. $r_{\text{PAGS•MCpost}}$= .21, p < .05). Student performance assessment individual scores were more closely related to the students’ total performance assessment scores than they were to the student performance assessment group scores ($r_{\text{PAIS•PA}}$ = .92, p < .01 vs. $r_{\text{PAGS•PA}}$ = .69, p < .01). Thus, having the students begin the assessment in groups for the purpose of reflect the FAST teaching methods, and then complete the assessment individually, did not compromise the assessment of individual students.

We designed the Rocky River Performance Assessment to provide a score for each of the four components of Duschl's Evidence-Explanation Continuum plus science communication (i.e., deciding what data are needed, data into evidence, evidence into patterns and models, patterns into explanations, and science communication). We found that the correlations between each of the five components and the performance assessment total score ranged from .61 to .87 (p < .01) and that the correlations of the five components with the posttest ranged from .20 to .27, (p < .05). Thus, the five components were reflected in the performance assessment score, suggesting content and construct validity, and correlated somewhat with the results from another achievement measure, suggesting concurrent validity.

Nature of Science Scale

For the Nature of Science Scale, we used data from seven of the items in the survey (α = .81). We interpreted that high scores on the nature of science survey indicated that students believed they had greater control in conducting scientific practices than did students who showed lower scores. Although we administered a series of 10 items to the students, three of the items did not behave as predicted and were dropped from the analysis. We found, as expected, that the pre- and post- Nature of Science Scale scores were correlated with each other (r = .67, p < .01). We also found that the Nature of Science Scale scores were positively correlated with the posttest multiple-choice score *(r* = .52, p < .01)—evidence of concurrent validity. We found that high performing students had higher NOS scores than low performing students, suggesting content validity, and did not find differences between pre- and posttests within groups.

Student Science Investigation Self-Efficacy Scale

For the 328 students with complete data sets including the science investigation self-efficacy, we found that all 12 items worked well together as expected (α = .90). The correlation between the pre- and post- surveys was statistically significant (r = .59, p < .01), and the concurrent-validity correlation between posttest achievement scores and post Self-Efficacy Scale survey scores was positive and significant (r = .35, p < .01), as expected. We also found differences between high and low performing students on their scores, as expected.

Motivation Scale

We present the results of the motivation survey in three parts—epistemic beliefs, confidence, and mastery learning.

Epistemic Beliefs Scale

In our survey, there were four items that addressed students' beliefs about the nature of knowing. A high score on this survey implies that the student believes that knowing in science is fixed and that the ability of learning this knowledge is predetermined at birth (i.e., you are born smart in science). The relationship between student epistemic beliefs between pre- and post-surveys was statistically significant (*r* = .48, *p* < .01). We also examined concurrent-validity relationships: As expected, the relationship between achievement and epistemic beliefs was negative (*r* = -.35, *p* < .01). Students with lower achievement believe that no matter how hard they try, they cannot learn, while high achieving students believe that if they try harder, they can learn more. Additionally, we found that student epistemic scores were negatively correlated with student self-efficacy scores at a statistically significant level (*r* = -.28, *p* < .01). The more that students believe intelligence is fixed, the less they believe that they have control over their learning from science investigations.

Confidence Scale

The six items that were used to measure students' confidence in their learning (e.g., I can learn science) worked well together (α = .79). The correlation between pre- and post- surveys was positive and statistically significant (*r* = .48, *p* < .01). The correlation of post-survey confidence scores with posttest achievement, which is evidence of concurrent validity, was also positive and also statistically significant (*r* = .26, *p* < .01). This relationship is weaker than expected, however. The post survey confidence scores vs. student self-efficacy scores were related and strong (*r* = .58, *p* < .01). The more the students indicated that they believed they were in charge of their own learning, the greater their confidence in their success in science. The post survey confidence scores vs. student epistemic beliefs scores were negatively related, as expected (*r* = -.32, *p* < .01). The more students showed that they believed that their intelligence is fixed, the less likely they were to have confidence in their learning.

Mastery Scale

Students' beliefs about mastery learning were assessed using five items (α = .80). In general, mastery oriented students are students who believe that intelligence is malleable and can grow over time. These students tend to be higher performing students. In our survey, we found a positive and statistically significant correlation between pre- and post-surveys (*r* = .56, *p* < .01). We also found positive but small correlations between Mastery Scale post- scores and posttest achievement scores (*r* = .13, *p* < .05) and between Mastery Scale scores and achievement test gain scores (*r* = .15, *p* < .01). However, other indicators suggest the validity of the Mastery Scale scores. Post- Mastery Scale scores were found to be positively related to student Science Self-Efficacy Scale scores (*r* = .33, *p* < .01) and post- Mastery Scale scores were negatively related to Epistemic Belief Scale scores (*r* = -.32, *p* < .01).

Value of Science Scale

Students rated the value of science (i.e., I find science interesting) using a five-point Likert scale. The value of science was measured using seven items (α = .88). In general, student science achievement is positively related to value of science. The more that the students find science interesting and likeable, the more likely they are to perform well. Indeed in this study, we found that students post- achievement scores were positively correlated with Value of Science scores (*r* = .125, p < .05). We also found that student Self-Efficacy Scale scores were positively related to students' Value of Science Scale scores (*r* = .368, *p* < .001).

Science Anxiety

Students rated their anxiety to science (e.g., "I get really uptight during science tests") using a four-point Likert scale with seven items (α = .83). In general, the more anxious a student is about performance in science, the less that student achieves. We found this relationship: Student anxiety towards science was negatively correlated with posttest achievement scores (*r* = -.38, *p* < .01), negatively correlated with Value of Science Scale scores (*r* = -.41, *p* < .01), and negatively correlated with student Self-Efficacy Scale scores (*r* = -.50, *p* < .001).

REFERENCES

Abd-El-Khalick, F., Bell, R. L., & Lederman, N. G. (1998).The nature of science and instructional practice: Making the unnatural natural. *Science Education*, *82*, 417–437.

Anderson, T., & Christiansen, J.-A. (2004). Online conferences for professional development. In C. Vrasidas & G. V. Glass (Eds.), *Online professional development for teachers* (pp. 13–29). Greenwich, CT: Information Age.

Arons, A. B. (1989). *What science do we teach. Curriculum development for the year of 2000.* Colorado Springs: Biological Science Curriculum Study.

Ayala, C. C. (2005a, April). *Developing student outcome measures frameworks.* Paper presented at the meeting of the American Educational Research Association, Montreal.

Ayala, C. C. (2005b, October). *Development and validation of an inquiry science student achievement and additudinal suite.* Paper presented at the meeting of the American Evaluation Association, Toronto.

Ball, D. L., Camburn, E., Correnti, R., Phelps, G., & Wallace, R. (1999). *New tools for research on instruction and instructional policy: A Web-based teacher log*. Seattle: University of Washington, Center for the Study of Teaching and Policy.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Barab, S. A., MaKinster, J. G., Moore, J. A., & Cunningham, D. J. (2001). Designing and building an on-line community: The struggle to support sociability in the Inquiry Learning Forum. *Educational Technology Research & Development, 49*(4), 71–96.

Barrett, P. (2001, March). *Assessing the reliability of rating data- revised*. Retrieved January 3, 2006, from http://www.pbbarrett.net/techpapers/rater.pdf

Berends, M., Kirby, S. N., Naftel, S., & McKelvey, C. (2001). *Implementation and performance in New American Schools: Three years into scale-up*. Santa Monica, CA: RAND.

Berman, P., & McLaughlin, M. W. (1978). *Federal programs supporting educational change: Implementing and sustaining innovations*. Santa Monica, CA: RAND.

Bickman, L. (1985). Improving established statewide programs: A component theory of evaluation. *Evaluation Review, 9*, 189–208.

Blakely , C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. M., Roitman, D. B., et al. (1987). The fidelity-adaptation debate: Implications for the implementation of public section social programs. *American Journal of Community Psychology, 15*, 253–268.

Blank, R. (1993). Developing a system of education indicators: Selecting, implementing, and reporting indicators. *Educational Evaluation and Policy Analysis*, *15*, 65–80.

Blank, R., Porter, A., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics & science. Results from Survey of Enacted Curriculum Project.* Washington, DC: Council of Chief State School Officers.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. W. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research, 2*, 75–87.

Bosker, R. J., & Scheerens, J. (1994). Alternative models of school effectiveness put to the test. Conceptual and methodological advances in educational effectiveness research. *International Journal of Educational Research, 21*, 159–180.

Brandon, P. R. (1993, November). *Studying the implementation of a medical-school problem-based learning curriculum: Lessons learned about the component-evaluation approach*. Paper presented at the meeting of the American Evaluation Association, Dallas.

Brandon, P. R., & Heck, R. H. (1998). The use of teacher expertise in decision making during school-conducted needs assessments: A multilevel perspective. *Evaluation and Programming Planning, 21*, 321–331.

Brandon, P. R., & Taum, A. K. H. (2005a, April). *Instrument development for a study comparing two versions of inquiry science professional development*. Paper presented at the meeting of the American Educational Research Association, Montreal.

Brandon, P. R., & Taum, A. K. H. (2005b, October). *Development and validation of the Inquiry Science Teacher Log and the Inquiry Science Teacher Questionnaire.* Paper presented at the meeting of the American Evaluation Association, Toronto.

Brandon, P. R., Taum, A. K. H., Young, D. B., & Pottenger, F. M., Speitel, T. W., Gray, M., et al. (2006b, July). *The implementation and outcomes of the Foundational Approaches in Science Teaching program: Results of a pilot study*. Paper presented at the meeting of the Pacific Circle Consortium, Mexico City.

Brandon, P. R., Taum, A. K. H., Young, D. B., Pottenger, F. M., Speitel, T. W., & Gray, M., et al. (2006b, November). *Findings about the implementation and outcomes of inquiry-based science in middle-school classrooms.* Paper presented at the meeting of the American Evaluation Association, Portland, OR.

Britner, S. L. (2002). *Science self-efficacy of African American middle school students: Relationship to motivation self-beliefs, achievement, gender, and gender orientation.* Unpublished doctoral dissertation, Emory University.

Britner, S. L., & Pajares, F. (2001). Self efficacy beliefs, motivation, race and gender in middle school science. *Journal of Women and Minorities in Science and Engineering, 7*, 271–285.

Brophy, J., & Evertson, C. (1976). *Learning from teaching: A developmental perspective*. Boston: Allyn and Bacon.

Brown, K. L., McDonald, S.-K., & Schneider, B. (2006). *Just the Facts: Results from IERI scale-up research.* Chicago: University of Chicago, National Opinion Research Center, Data Research and Development Center.

Brunvand, S., Fishman, B., & Marx, R. (2003, April). *Moving professional development on-line: Meeting the needs of all teachers*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Building Science and Engineering Talent. (2004). *What it takes: Pre-K-12 design principles to broaden participation in science, technology, engineering and mathematics.* San Diego, CA: Author.

Bybee, R. (1993). *Reforming science education: Social perspectives and personal reflections.* New York: Teachers College.

Callahan, W. P., & Switzer, T. J. (1999). *Technology as facilitator of quality education: A model*. Retrieved February 18, 2005, from http://www.intime.uni.edu/model/modelarticle.html

Camburn, E., & Barnes, C. A. (2004). *Assessing the validity of a language arts instruction log through triangulation*. Retrieved March 3, 2005, from www.sii.soe.umich.edu/documents/esj_log_validity_10Mar04.pdf

Carey, N., & Shavelson, R. (1989). Outcomes, achievement, participation, and attitudes. In R. Shavelson, L. McDonnell, & J. Oakes (Eds.), *Indicators for monitoring mathematics and science education: A sourcebook* (pp. 147–181). Santa Monica, CA: RAND.

Carey, N. (1989). Instruction. In R. Shavelson, L. McDonnell, & J. Oakes (Eds.), *Indicators for monitoring mathematics and science education: A sourcebook* (pp. 123–146). Santa Monica, CA: RAND.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive-behavior. *American Journal of Mental Deficiency*, 86, 127–137.

Cohen, D. K., & Hill, H., C. (1998). *State policy and classroom performance: Mathematics reform in California* (CPRE Policy Brief No. RB-23). Philadelphia, PA: Consortium for Policy Research in Education.

Coker, H., Lorentz, J., & Coker, J. (1980, April). *Teacher behavior and student outcomes in the Georgia Study*. Paper presented at the meeting of the American Educational Research Association, Boston.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity*. Washington, DC: U. S. Government Printing Office.

Council of Chief State School Officers. (2000). *Using data on enacted curriculum in mathematics and science: Sample results of a study of classroom practices.* Washington, DC: Author.Creemers, B. (1993). *Toward a theory on educational effectiveness*. Paper presented at the meeting of the International Congress for School Effectiveness and Improvement, Norrkoping, Sweden. (ERIC Document Reproduction Service No. ED361828)

Creemers, B., & Reezigt, G. (1996). School level conditions affecting the effectiveness of instruction. *School Effectiveness and School Improvement, 7*, 197–228.

Creemers, B., & Scheerens, J. (1994). Developments in the educational effectiveness research programme. *International Journal of Educational Research*, *21*, 125–140.

Curriculum Research & Development Group (2000). *FAST: A summary of evaluations*. Honolulu: Author.

Curriculum Research & Development Group. (1996). *Alignment of Foundational Approaches in Science Teaching (FAST) with the national science education standards grades 5—8.* Honolulu: Author.

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*, 23–45.

Darling-Hammond, L., & Hudson, L. (1989). Teachers and teaching. In R. Shavelson, L. McDonnell, & J. Oakes (Eds.), *Indicators for monitoring mathematics and science education: A sourcebook* (pp. 66–95). Santa Monica, CA: RAND.

Darling-Hammond, L., & McLaughlin, M. W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan, 76*, 597–605.

Datnow, A. (1998). *The gender politics of educational change*. London: Falmer.

Datnow, A., & Stringfield, S. (2000).Working together for reliable school reform. *Journal of Education for Students Placed At Risk, 5*, 183–204.

David, H. A. (1963). *The method of paired comparisons.* New York: Hafner.

de Jong, T., & Ferguson-Hesser, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist*, *31*, 105–113.

Dekkers, J. (1978). The effects of junior inquiry science programs on student cognitive and activity preferences in science. *Research in Science Education, 8*, 71–78.

Desimone, L., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*, 1–22.

Desimone, M. L., Porter, A. C., Garet, M., Yoon, S. K., & Birman, B. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*, 81–112.

Dunn-Rankin, P., Knezek, G., A. Wallace, S., & Zhang, S. (2004). *Scaling methods* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Duschl, R. (1990). *Restructuring science education*. New York: Teachers College.

Duschl, R.A. (2003). Assessment of inquiry. In M. Atkin & J.E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41–59). Arlington, VA: National Science Teachers Association.

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, 18*, 237–256.

Dweck, C., Davidson, W., Nelson, S., & Enna, B. (1978). Sex difference in learned helplessness: (II) The contingencies of evaluative feedback in the classroom and (III) An experimental analysis. *Developmental Psychology. 14*, 268–276.

Dweck, C., & Henderson V. (1989). *Predicting individual differences in school anxiety in early adolescence* (ERIC Document Reproduction Service No. ED310884)

Edwards, A. L. (1957). *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts.

Evans, W. (1986). An investigation of curriculum implementation factors. *Education, 106*, 447–453.

Evertson, C. M., & Green, J. L. (1986). Observation as inquiry and method. In Wittrock, M. C. (Ed.), *Handbook of research on teaching* (pp. 162–213). New York: Macmillan.

Firestone, W. A., & Corbett, H. D. (1988). Planned organizational change. In N. J. Boyan (Ed.), *Handbook of research on educational administration* (pp. 321–340). New York: Longman.

Fishman, B. J., Marx, R. W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and Teacher Education*, *19*, 643–658.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Fullan, M. G. (2001). *The new meaning of educational change* (3rd ed.). New York: Teachers College.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. *Review of Educational Research, 47*, 355–397.

Gall, M. (1970). The use of questions in teaching. *Review of Educational Research, 40*, 707–721.

Gall, M. (1984). Synthesis of research on teachers' questioning. *Educational Leadership, 42*(3), 40–47.

Galley, M. (2004, May 19). Studies suggest science education neglected. *Education Week*, *23*, 12.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*, 915–945.

Gray, M. E., Nguyen, T. T. T., & Speitel, T. W. (2005, April). *Developing and implementing an alternative version of FAST professional development.* Paper presented at the meeting of the American Educational Research Association, Montreal.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review, 18*, 37–50.

Gulliksen, H., & Tukey, J. W. (1958). Reliability for the law of comparative judgment. *Psychometrika, 23*, 95–110.

Guskey, T. R., & Huberman, M. (Eds.). (1995). *Professional development in education: New paradigms and practices.* New York: Teachers College.

Hall, G. E., & Loucks, S. F. (1978). *Innovation configurations: Analyzing the adaptation of innovations.* Paper presented at the meeting of the American Educational Research Association, Toronto.

Hamilton, R., & Brady, M. P. (1991). Individual and classwide patterns of teachers' questioning in mainstreamed social studies and science classes. *Teaching and Teacher Education, 7*, 253–262.

Hannafin, M. J., Hill, J. R., Oliver, K., Glazer, E., & Sharma, P. (2003). Cognitive and learning strategies in web-based environments. In M. G. Moore & W. G. Anderson (Eds.), *Handbook of distance education* (pp. 245–260). Mahwah, NJ: Lawrence Erlbaum.

Hara, N., & Kling, R. (1999). Students' frustrations with a web-based distance education course. *First Monday*, *4*(12), retrieved May 10, 2007 from http://www.firstmonday.org/issues/issue4_12/hara/index.html

Harasim, L., Hiltz, S. R., Teles, L., &. Turoff, M. (1995). *Learning networks: A field guide to teaching and learning online*. Cambridge, MA: Massachusetts Institute of Technology.

Harlen, W. (2004). The development of assessment for learning: Learning from the case of science and mathematics. *Hodder Arnold Journals*, *21*, 390–408.

Harlen, W., & Doubler, S. J. Can teachers learn through enquiry on-line? Studying professional development in science delivered on-line and on-campus. *International Journal of Science Education, 26*, 1247–1267.

Hawai'i Science Curriculum Council. (1967). *Status report on grade 7–9 science project.* Honolulu: Curriculum Research & Development Group.

Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession* (pp. 127–150). San Francisco: Jossey-Bass.

Haydel, A., & Roser, R. (2002). On motivation, ability and the perceived situation in science test performance: A person centered approach with high school students. *Educational Assessment*, *82*, 163–190.

Heath, R. W., & Brandon, P. R. (1982). An alternative approach to the evaluation of educational and social programs. *Educational Evaluation and Policy Analysis*, *4*, 477–486.

Heck, R. H. (1993). School context, principal leadership, and achievement: The case of secondary schools in Singapore. *Urban Review, 25*, 151–166.

Heck, R. H., & Brandon, P. R. (1995). Teacher empowerment and the implementation of school-based reform. *Empowerment in Organizations, 3*(4), 10–19.

Heck, R. H., Brandon, P. R., & Wang, J. (2001). Implementing site-managed educational changes: Examining levels of implementation and effect. *Educational Policy, 15*, 302–322.

Heck, R., & Mayor, R. (1993). School characteristics, school academic indicators and student outcomes: Implications for policies to improve schools. *Journal of Education Policy, 8*, 143–154.

Heck, R., Larsen, T., & Marcoulides, G. (1990). Instructional leadership and school achievement: Validation of a causal model. *Educational Administration Quarterly, 26*, 94–125.

Hilberg, R. S., Doherty, R. W., Epaloose, G., & Tharp, R. G. (2004). The Standards Performance Continuum: A performance-based measure of the standards for effective pedagogy. In H. C. Waxman, R. G. Tharp, & R. S. Hilberg (Eds.), *Observational research in U. S. classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 48–71). Cambridge, England: Cambridge University Press.

Honey, M., & Moeller, B. (1990). *Teachers' beliefs and technology integration: Different values, different understandings*. (CTE Technical Report Issue No. 6). New York: Center for Children and Technology. (ERIC Document Reproduction Service No. ED326203)

Hord, S. M., Rutherford, W. L., Huling-Austin, L., & Hall, G. E. (1987). *Taking charge or change*. Alexandria, VA: Association for Supervision and Curriculum Development.

Howell, D. C. (2007). *Statistical methods for psychology*. Belmont, CA: Thomson/Wadsworth.

Huai, N., Braden, J., White, J., & Elliott, S.N. (2003). *Effect of an Internet-based multimedia teacher development program in enhancing teachers' assessment literacy*. (WCER Working Paper No. 2003-9). Madison, WI: Wisconsin Center for Education Research.

Huberman, M. (1989).The professional life cycle of teachers. *Teachers College Record, 91*, 30–57.

Inquiry Synthesis Project. (2004). *Technical Report 2: Conceptualizing Inquiry Science Instruction.* Boston: Center for Science Education, Education Development Center.

Jones, T. H., & Paolucci, R., (1998). The learning effectiveness of educational technology: A call for further research. *Educational Technology Review, 9*, 10–14.

Joyce, B., & Showers, B. (1987). Low cost arrangement for peer coaching. *Journal of Staff Development, 8*(1), 22–24.

Joyce, B., & Showers, B. (1995). *Student achievement through staff development: Fundamentals of school renewal* (2nd ed.). White Plans, NY: Longman.

Joyce, B., Showers, B, & Rolheiser-Bennett. C. (1987). Staff development and student learning: A synthesis of research on models of teaching. *Educational Learning, 45*(2), 11–23.

Kabilan, M. K. (2004). On-line professional development: A literature analysis of teacher competency. *Journal of Computing in Teacher Education, 21*(2), 51-58.

Kane, M. T. (2006). Validation. In R. L. Brennan, *Educational measurement* (4th ed.) (pp. 17–64). New York: American Council on Education/Praeger.

Kennedy, M. (1998). *The relevance of content in in-service teacher education*. Paper presented at the meeting of the American Educational Research Association, San Diego.

Kirby, S. N., Berends, M., & Naftel, S. (2001). Implementation in a longitudinal sample of New American Schools: Four years into scale-up. Santa Monica, CA: RAND.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86.

Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., Robyn, A., & Burroughs, D. (2000). *Teaching practices and student achievement: Report for first-year findings from the 'mosaic' study of systemic initiatives in mathematics and science*. Santa Monica, CA: RAND.

LeBlanc, L., & Turnbull, B. (2001). *The Longitudinal Evaluation of School Change and Performance (LESCP) in Title I schools: Final report. Executive summary.* Washington, DC: U.S. Department of Education, Planning and Evaluation Services. (ERIC Document Reproduction Service No. ED457305)

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science, *Journal of Research on Science Teaching, 39*, 497-512.

Lee, O., Hart, J. E., Cuevas, P., & Enders, C. (2004). Professional development in inquiry-based science for elementary teachers of diverse student groups. *Journal of Research in Science Teaching, 41*, 1021– 1043.

Lee, V. E., & Smith, J. B. (1997). High school size: Which works best and for whom? *Educational Evaluation and Policy Analysis*, *19*, 205–227.

Li, M. & Shavelson, R. J. (2001, April). *Examining the links between science achievement and assessment.* Paper presented at the meeting of the American Educational Research Association, Seattle.

Little, J. W. (1981). *The power of organizational setting*. Washington, DC: National Institute of Education.

Loucks-Horsley, S., Hewson, P. W., Love, N., & Stiles, K. E. (1998). *Designing professional development for teachers of science and mathematics*. Thousand Oaks, CA: Corwin Press.

Lynch, S. J. (2007, April). *A model for fidelity of implementation in a study of a science curriculum unit: Evaluation based on program theory.* Paper presented at the meeting of the American Educational Research Association, Chicago.

Lynch, S. J., & O'Donnell, C. L. (2005, April). *"Fidelity of Implementation" in implementation and scale-up research designs: Applications from four studies of innovative science curriculum materials and diverse populations.* Paper presented at the meeting of the American Educational Research Association, Montreal.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., & Soloway, E. (1998). New technologies for teacher professional development. *Teaching and Teacher Education, 14*, 33–52.

Matthews, D. (1998). Transforming higher education: implications for state higher education finance policy. *Educom Review, 33*(5), 48–50, 52–54, 56–57.

Mayer, D., Mullens, J., Moore, M., & Ralph, J. (2000). *Monitoring school quality: An indicators report* (NCES Report No. 2001-030). Washington, DC: National Center for Education Statistics.

McDermott, L. C. (1990). A perspective on teacher preparation in physics and other sciences: The need for special science courses for teachers. *The American Journal of Physics, 58*, 734–742.

McLaughlin, M. W., & Marsh, D. D. (1990). Staff development and school change. In A. Lieberman (Ed.), *Schools as collaborative cultures: Creating the future now* (pp. 213–232). New York: Falmer.

Meece, J., Wigfield, A., & Eccles, J. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment and performance in mathematics. *Journal of Educational Psychology, 82*, 60–70.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11*, 247– 266.

Morine-Dershimer, G. (1985). *Talking, listening, and learning in the elementary classroom*. New York: Longman.

Mortimore, P., Sammons, P., Stoll, L., & Lewis, D. (1989). A study of effective junior schools. *International Journal of Educational Research, 13,* 753–768**.**

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years*. Sommerset, England: Open Books.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*, 315–340.

Mullens, J. & Kasprzyk, D. (1996). Using qualitative methods to validate quantitative survey instruments. In *Proceedings of the Section on Survey Research Methods* (pp. 638–643). Alexandria, VA: American Statistical Association.

Murname, R. (1981). Interpreting the evidence on school effectiveness. *Teachers College Record, 83*, 19–35.

Muthen, H., Huang, L. C., Jo, B., Khoo, S. T., Goff, G. N., Novak, J. R., et al. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis*, *17*, 371–403.

National Center for Education Statistics. (2003). *Schools and Staffing Survey 2003-04 Teacher Questionnaire*. Washington, D.C.: U.S. Department of Education.

National Center for Education Statistics. (2003). *Digest of education statistics tables and figures, 2003*. Retrieved October 12, 2005 from National center for Edui Statistics Web site: http://nces.ed.gov/programs/digest/d03/tables/dt069.asp

National Center for Education Statistics. (n.d.). *Schools and Staffing Survey (SASS)*. Retrieved March 3, 2007 from http://nces.ed.gov/surveys/sass/

National Center of Education Statistics. (2007). *Fast response survey system.* Retrieved March 3, 2007 from http://nces.ed.gov/surveys/frss/

National IOTA Program (1970). *Assessment of teaching competence for improvement of instruction.* Author: Tempe, AZ. (ERIC Document Reproduction Service No. ED102687)

National Research Council. (1996). *National science education standards*. Washington D.C.: National Academy Press.

National Science Foundation (2004). *Interagency education research initiative* (NSF program solicitation 05-553). Retrieved January 26, 2007 from http://www.nsf.gov/pubs/2004/ nsf04553/nsf04553.htm

National Staff Development Council. (1999). *What works in the middle: Results-based staff development*. Oxford, OH: Author.

Newmann, R., & Wehlage, G. (1995). *Successful school restructuring.* Madison, WI: Center on Organization and Restructuring of Schools.

Nguyen, T.T. T., Speitel, T. W., & Gray, M. E. (2007). *Development of FASTeR: A multimedia DVD-ROM for science teacher education.* Manuscript submitted for publication.

Northwest Regional Educational Laboratory. (1998). *Catalog of school reform models.* Portland, OR: Author.

Oakes, J., & Carey, N. (1989). Curriculum. In R. Shavelson, L. McDonnell, & J. Oakes (Eds.), *Indicators for monitoring mathematics and science education: A sourcebook* (pp. 96–122). Santa Monica.CA : RAND.

Oakes, J. (1989a). School context and organization. In R. Shavelson, L. McDonnell, & J. Oakes (Eds.), *Indicators for monitoring mathematics and science education: A sourcebook* (pp. 40–65). Santa Monica, CA: RAND.

Oakes, J. (1989b). What educational indicators? The case for assessing the school context. *Educational Evaluation and Policy Analysis*, *11*, 181–199.

O'Connor, B. P. (2000). SPS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32*, 396–402.

O'Donnell, C. (2007). *Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research.* Manuscript submitted for publication.

Office of Educational Research and Improvement. (1990). *Science education programs that work.* Washington, DC: U.S. Department of Education.

Office of Educational Research and Improvement. (1994). *Science and mathematics education programs that work.* Washington, DC: U.S. Department of Education.

Ogawa, R., & Bossert, S. (1995). Leadership as an organizational quality. *Educational Administration Quarterly, 31*, 224–243.

Pajares, F., & Urdan, T. (1996). Exploratory factor analysis of the Mathematics Anxiety Scale. *Measurement and Evaluation in Counseling and Development, 29*, 35–47.

Patton, T.R. (1978). *Utilization-focused evaluation*. Beverly Hills, CA: Sage.

Pauls, J., Young, D. B., & Lapitkova, V. (1999). Laboratory for learning. *The Science Teacher, 66*, 27–29.

Pinto, J. K., & Prescott, J. E. (1990). Planning and tactical factors in the project implementation process. *Journal of Management Studies, 27*, 305–327.

Pintrich, P. R. (1999). *Motivational beliefs as resources for and constrains on conceptual change*. New York: Pergamon Press.

Pintrich, P. R., Marx, R., & Boyle, R. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research, 63*, 167–199.

Polin, L. (2000, April). *Affordances of a VR world as a place for learning: Discourse patterns and contextualization cues framing learning experiences for adult learners in a real-time, text-based, virtual reality setting.* Paper presented at the meeting of the American Educational Research Association, San Diego.

Porter, A. (1993). School delivery standards. *Educational Researcher, 22*(5), 24–30

Porter, A., Kirst, M., Osthoff, E., Smithson, J., & Schneider, S. (1993). *Reform up close: An analysis of high school mathematics and science classrooms*. Washington, DC: The National Center for Improving Science Education. (ERIC Document Reproduction Service No. ED364429)

Pottenger, F. M. (2005). *Inquiry in the Foundational Approaches in Science Teaching program.* Honolulu: University of Hawai‘i at Mānoa, Curriculum Research & Development Group.

Pottenger, F. M., & Young, D. B. (1992a). *Instructional guide: FAST, Foundational Approaches in Science Teaching* (2nd ed.). Honolulu: University of Hawai‘i at Mānoa, Curriculum Research & Development Group.

Pottenger, F. M., & Young, D. B. (1992b). *The local environment: FAST 1, Foundational Approaches in Science Teaching* (2nd ed.). Honolulu: University of Hawai‘i at Mānoa, Curriculum Research & Development Group.

Pottenger, F. M., & Young,D. B. (1992c). *The local environment: FAST 1, Foundational Approaches in Science Teaching, teacher's guide* (2nd ed.). University of Hawai‘i at Mānoa, Curriculum Research and Development Group.

Pratton, J., & Hales, L. W. (1986). The effects of active participation on student learning. *Journal of Educational Research*, 79, 210–215.

Price, R. V. (1996). Technology doesn't teach. People do. *Techtrends, 41*, 17-18.

Redfield, D. L., &. Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research, 51*, 236–245.

Reezigt, G., Guldemon, H., & Creemers, B. (1999). Empirical validity for a comprehensive model on educational effectiveness. *School Effectiveness and School Improvement, 10*, 193–216.

Roberts-Gray, C. (1985). Managing the implementation of innovations. *Evaluation and Program Planning, 8*, 261–269.

Rogg, S., & Kahle, J. B. (1997). *Middle level standards-based inventory.* Oxford, OH: University of Ohio.

Rowan, B., Harrison, D. M., & Hayes, A. (2003.) *Using instructional logs to study elementary school mathematics: A close look at curriculum and teaching in the early grades*. Ann Arbor, MI: University of Michigan.

Ruiz-Primo, M. A. (2005, April). *A multi-method and multi-source approach for studying fidelity of implementation.* Paper presented at the meeting of the American Educational Research Association, Montreal..

Ruiz-Primo, M. A., & Shavelson, R. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching, 33,* 1045–1063.

Samson, G. E., Sirykowski, B., Weinstein, T., &. Walberg., H. J. (1987). The effects of teacher questioning levels on student achievement: A quantitative synthesis. *Journal of Educational Research, 80*, 290–295.

Scheerens, J., & Creemers, B. (1989). Conceptualizing school effectiveness. *International Journal of Educational Research, 13,* 691–706.

Scheerens, J., & Creemers, B. (1996). School effectiveness in the Netherlands: The modest influence of a research programme. *School Effectiveness and School Improvement, 7*, 181–195.

Scheerens, J., Vermeulen, A., & Pelgrum, W. (1989). Generalizability of instructional and school effectiveness indicators across nations. *International Journal of Educational Research, 13,* 789–799.

Scheirer, M. A., & Rezmovic, E. L. (1983). Measuring the degree of program implementation: A methodological review. *Evaluation Review, 7*, 599–633.

Schlager, M. S. & Schank, P. K. (1997, December). *Tapped in: A new on-line teacher community concept for the next generation of internet technology*. Paper presented at the meeting of the Second International Conference on Computer Support for Collaboration Learning, Toronto.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century. Ninetieth yearbook of the National Society for the Study of Education, Part II* (pp. 19–64). Chicago: National Society for the Study of Education

Shavelson, R. (1995). On the romance of science curriculum and assessment- reform in the United States. In D.K. Sharpes & A-L Leino (Eds.), *The dynamic concept of curriculum: Invited papers to honor the memory of Paul Hellgren* (pp. 57-76) (Research Bulletin 90). Helsinki, Finland: University of Helsinki, Department of Education.

Shavelson, R., McDonnell, L., & Oakes, J. (1989). The design of educational indicator systems: An overview. In R. Shavelson, L. McDonnell, and J. Oakes (Eds.), *Indicators for monitoring mathematics and science education: A sourcebook* (pp. 1–39). Santa Monica, CA: RAND.

Shavelson, R., Solano-Flores, G., & Ruiz-Primo, M. A. (1998.) Toward a science performance assessment technology. *Evaluation and Program Planning, 21*, 171–184.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Slater, R., & Teddlie, C. (1992). Toward a theory of school effectiveness and leadership. *School Effectiveness and School Improvement*, *3*, 242–257..

Smylie, M. (1992). Teacher participation in school decision making: Assessing willingness to participate. *Educational Evaluation and Policy Analysis, 15*, 53–67.

Soar, R. S. (1973). *Follow through classroom process measurement and pupil growth (1970-1971, final report)*. Gainesville: University of Florida, Institute for Development of Human Resources. (ERIC Document Reproduction Services Np. ED 106 297)

Speitel, T. W., & Nguyen, T. T. T. (2001). *School Web of Instructional Media Website*. Honolulu: University of Hawai'i at Mānoa, Curriculum Research & Development Group. Retrieved March 7, 2005 from http://www.hawaii.edu/swim/

Stallings, J., & Kaskowitz, D. (1974). *Follow-through classroom evaluation, 1972-1973: A study of implementation*. Menlo Park, CA: Stanford University.

Stanford Education Assessment Laboratory and Curriculum Research & Development Group (2005). *Embedding assessments in the FAST curriculum: The romance between curriculum and assessment. Final report for* Palo Alto, CA: Authors

Study of Instructional Improvement. (2001). *Teacher questionnaire, 2000–2001* . Ann Arobor, MI: University of Michigan, Survey Services Lab.

Supovitz, J. A. & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*, 963–980.

Supovitz, J. A. (2001). Translating teaching practice into improved student achievement. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (pp. 81–98). Chicago, Illinois: National Society for the Study of Education.

Supovitz, J. A., Mayer, D. P., & Kahle, J. B. (2000). Promoting inquiry-based instructional practice; the longitudinal impact of professional development in the context of the system. *Educational Policy, 14*, 331–356.

Tamir, P. & Yamamoto, K. (1977). The effects of the junior high FAST program on student achievement and preferences in high school biology. *Studies in Educational Evaluation, 3*, 7–17.

Tamir, P. (1983). Inquiry and the science teacher. *Science Teacher Education, 67*, 657–672.

Tamir, P., & Yamamoto, K. (1977). The effects of the junior high FAST program on student achievement and preferences in high school biology. *Studies in Educational Evaluation, 3*, 7–17.

Taum, A. H. K. (2004). *Foundational Approaches to Science Teaching and the five standards for effective pedagogy*. Unpublished manuscript.

Taum, A. H. K. & Brandon, P. B. (2005a, April). *Coding teachers in inquiry science classrooms using the Inquiry Science Observation Guide*. Paper presented at the meeting of the American Educational Research Association, Montreal, Canada.

Taum, A. K. H., & Brandon, P. R. (2005b, October). *The development of the Inquiry Science Observation Code Sheet*. Paper presented at the meeting of the American Evaluation Association, Toronto.

Taum, A. K. H., & Brandon, P. R. (2006, April). *The iterative process of developing an inquiry science classroom observation protocol.* Paper presented at the meeting of the American Educational Research Association. San Francisco.

Teddlie, C. & Reynolds, D. (2001). Countering the critics: Responses to recent criticisms of school effectiveness research. *School Effectiveness and School Improvement*, *12*, 41–82.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: John Wiley & Sons.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: Reflections on a century of public school reform*. Cambridge, MA: Harvard University.

U.S. Department of Education Mathematics and Science Education Expert Panel. (2001). *Exemplary & promising science programs 2001.* Washington, DC: U.S. Department of Education Office of Educational Research and Improvement.

Van Dusen, G. C. (2000). *Digital dilemma: Issues of access, cost, and quality in media enhanced and distance education*. San Francisco: Jossey Bass.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In. R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 41–71). Boston: Kluwer.

Vrasidas, C., & Glass, G. V. (2004) Teacher professional development: Issues and trends. In C. Vrasidas & G.V. Glass (Eds.), *Online professional development for teachers* (pp.1–12). Greenwich, CT: Information Age Publishing.

Wahlberg, H., & Shanahan, T. (1983). High school effects on individual students. *Educational Researcher, 12*(17), 4–9.

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20*, 137–156.

Watson, C. R. (1999). *Best practices from America's middle schools.* Larchmont, NY: Eye on Education.

Weil, M., & Murphy, J. (1982). Instructional processes. In H. H. Mitzel (Ed.), *Encyclopedia of educational research* (Vol. 4, 5th ed.). New York: Macmillian.

Weiss, I. R., Montgomery, D. L., Ridgway, C. J., & Bond, S. L. (1998). *Local systemic change through teacher enhancement: Year three cross-site report.* Chapel Hill, NC: Horizon research.

WGBH Educational Foundation. (2005). *Teachers' Domain*. Retrieved March 2, 2005 from http://www.teachersdomain.org

Wholey, J. S. (1994). Assessing the feasibility and likely usefulness of evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 15–39). San Francisco: Jossey-Bass.

Willms, J., & Kerckhoff, A. (1995). The challenge of development new educational indicators. *Educational Evaluation and Policy Analysis*, *17*, 113–131.

Wu, H. K., & Hsieh, C. E. (2006). Developing sixth grader's inquiry skills to construct explanations in inquiry-based learning environments. *International Journal of Science Education*, *28*, 1290–1313.

Yamamoto, K. K. (1996). *Against all odds: Tales of survival and growth of the Foundational Approaches in Science Teaching (FAST) project*. Unpublished doctoral dissertation, Stanford University.

Young, D. B. (1982). Local science program makes good: The evaluation of FAST. *Human Sciences, Technology, and Education, 1*, 23–28.

Young, D. B. (1993). Science achievement and thinking skills. *Pacific-Asian Education, 5*, 35–49.

Young, D. B. (1996). School change and improvement: The CRDG experience. *Educational Perspectives*, *30*(2), 23–32.

Zigarmi, P., Betz, L., & Jennings, D. (1977). Teachers' preferences in and perceptions of inservice. *Educational Leadership*, *34,* 545–551.