Development and Validation of the Inquiry Science Teacher Log and the Inquiry Science Teacher Questionnaire[1]

Paul R. Brandon and Alice K. H. Taum
University of Hawai'i at Mānoa

Presented at the annual meeting of the American Evaluation Association, Toronto, October 2005

This paper describes the development and validation to date of a teacher log and a questionnaire for a National Science Foundation-funded study of inquiry-based middle-school science (here called *inquiry science*) at the University of Hawai'i at Mānoa. The log, known as the Inquiry Science Teacher Log (ISTL), is for collecting data on the implementation of inquiry science lessons as they are taught, and the questionnaire, known as the Inquiry Science Teacher Questionnaire (ISTQ), is for collecting data annually on program implementation and on the teacher and contextual characteristics that might affect implementation and student achievement. Teacher observation and interview instruments and student instruments are described in other papers in this session. Data on school characteristics are collected from Web sites.

Instruments assessing program implementation should be based on the program being studied. We begin with a description of this program. The instruments also should address program-implementation theory and the aspects of implementation that researchers and evaluators have previously studied. We describe the literature on program implementation and how our instruments address this literature; then we describe how we identified the variables that the instruments address and describe the development of the instruments, with validation results.

The Program for Which the Instruments Were Developed

The ISTL and the ISTQ are designed to assess the implementation of the Foundational Approaches in Science Teaching (FAST), an interdisciplinary middle-school, inquiry science program developed at Curriculum Research & Development Group (CRDG), University of Hawai'i at Mānoa. It consists of three inquiry courses entitled, "The Local Environment," "Matter and Energy in the Biosphere," and "Change Over Time." FAST is aligned with the National Science Education Standards (CRDG, 1996; Rogg & Kahle, 1997) and addresses the components and characteristics of effective PD described in the summary of the research presented in this proposal.

There are a multitude of definitions of inquiry science education (Inquiry Synthesis Project, 2004). Most focus on the steps of a simplified process of scientific inquiry, as given in the National Science Education Standards (National Research Council, 2000): 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

[1]The project described here was funded by National Science Foundation Grant No. REC0228158.

clarified through interactions with others.

These features are common to student inquiry and are reflected in the learning that occurs in FAST's version of science education. FAST classrooms model the experience of practicing scientists, with students working in research teams to generate the theoretical content of the program. Usually working in small collaborative groups, students create hypotheses, conduct experiments, organize and analyze data, communicate their findings, and develop consensual conclusions. Class discussion follows each investigation to identify universal principles—for example, about why things sink and float. Students spend 70–80% of their time in laboratory or field studies, allowing time to define categories of events, generate hypotheses, test hypotheses, correct misconceptions, and ultimately come to a consensus on the adequacy of their data and subsequent explanations. The remainder of students' time is devoted to data analysis, small group or class discussions, literature research, and report writing. Concepts are presented to the students in a sequence of tasks in which FAST teachers are "research directors," stimulating and facilitating ever deeper probing into problems. The FAST student research-team approach tolerates temporary student misconceptions, because the contexts of investigation are carefully sequenced so that hypotheses, explanations, and conclusions are reviewed and retested.

Descriptions of the steps of traditional inquiry-learning methods do not adequately depict the full range of inquiry in scientists' practice, however; nor do they fully reflect inquiry as it is taught in FAST PD. Pottenger (2005) presents a full description of the conceptualization of FAST inquiry. For our purposes here, in addition to the inquiry steps in which students engage, another key aspect of the FAST inquiry model is 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]). Socratic inquiry occurs when teachers guide students through the other forms of inquiry.

The Literature on 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, Berman and McLaughlin (1976) and 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 in public health. Educational researchers and evaluators began to contribute more on the topic again after 2000. (In our opinion, Ruiz-Primo [2005] has provided the most comprehensive and thoughtful recent review of the literature on program implementation, with an emphasis on educational program implementation.)

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. It addresses three questions: What are the purposes of studying program implementation? Should the focus be on program components or another division of programs? What aspects of implementation should be examined? These are answered in the following three 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. One 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 irrespective in causal studies 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 features of their programs, 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., Baumann, Stein, & Ireys, 1991; 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, aspect of implementation. Therefore, in our model, contextual variable 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 directly examined participant responsiveness in our study except in the form of measuring student attitudes toward science. We intend to broaden our student surveys in the future by asking for their opinions about inquiry science, because student participation in classroom activities is the key to successful implementation (Lynch & O'Donnell, 2005). We also hope to develop methods for measuring student participation through observation—the most valid and productive means of gathering data on student participation. As Fullan and Pomfret (1977) stated, the interaction among roles (or "instructional transactions," as stated by Ruiz-Primo [2005]) is central to successful program implementation in the classroom. Furthermore, we have not addressed program differentiation. This has to do with a feature of good design that any experiment should address. When we use our implementation procedures in an experimental study, we will address it, although not as a feature of program implementation.

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.

Identifying Variables to Address in the Implementation Instruments

The first step in developing the instruments was to identify the variables that the instruments should address. We examined variables addressing two bodies of constructs that we describe fully in this section—those addressing teaching science with inquiry methods and those addressing the context within which inquiry science is taught. Inquiry science succeeds to the extent that teachers adequately use inquiry methods to teach science and to the extent that these methods positively affect student achievement. The context variables were specified in our grant proposal, which discussed the need to consider a number of variables that might affect program implementation.

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

To identify variables addressing inquiry science teaching, we reviewed a monograph describing inquiry and the manifestation of inquiry within FAST, which was prepared for the project by the FAST senior developer (Pottenger, 2005). We reviewed 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). From these documents, we identified the variables that all the FAST student investigations had in common. We examined an observation protocol and teacher log developed at the Stanford Educational Assessment Laboratory, with whom Curriculum Research & Development Group has collaborated on another NSF project about inquiry science in the form of the FAST program (No. ESI 0095520). Finally, we examined 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 inquiry-science studies. Our reviews of these two instruments proved helpful as we narrowed our conceptualization of our instruments, but we did not base any of our instruments on them.

Preparing a Draft of the List of Variables That Address the Context Within Which Inquiry Science Is Taught

To identify the variables addressing the school context within which inquiry science is taught, we examined and, in many cases, outlined all or some of about 55 pertinent books, monographs, and articles that widely reviewed the curriculum-indicator literature or the school effectiveness 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; Rowan et al., 2001; Scheerens & Creemers, 1989, 1996; Scheerens, Vermeulen, & Pelgrum, 1989; Shavelson, McDonnell, & Oakes, 1989; Slater & Teddlie 1992; Teddlie & Reynolds, 2001; Wahlberg & Shanahan, 1983; Wang, 1998; Willms & Kerckhoff, 1995). We prepared a 24-page table showing the school and community

variables affecting program implementation and student achievement that we identified in the literature review, including the variable names, the sources discussing the literature on the variables, and the conclusions about the variables that were drawn from the sources.

Combining and Defining the Two Draft Lists of Variables

The next step was to combine the two lists of variables. We then classified them into five categories, including student, teacher, classroom, school, and district or community. The combined list of variables was refined through 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 lessons ("investigations") and classified them according to the phase of the lesson (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. The resulting list of variables was reviewed several times by FAST developers, followed on each occasion by revisions, deletions, and additions. Variables were (a) revised because of vagueness or inaccuracy, (b) deleted in an effort to make data collection with the observations and log manageable, or (c) added to ensure that data would be collected on the aspects of student inquiry that are essential for student learning. Further revisions in the variables were made during the ensuing months in the process of developing observation procedures for the project.

In consultation with 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), we decided that the project 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 final step in this stage was to develop a "blueprint" showing the match between instruments and variables (but, unlike most blueprints, not the match between instruments and items). We prepared a list of the instruments that were best suited to collect data addressing the variables, including (a) the *Inquiry Science Observation Guide* (ISOG), addressing many key aspects of the implementation of each FAST student investigation; (b) the *Inquiry Science Teacher Log* (ISTL), addressing a few key aspects the implementation of each FAST student investigation; and (c) the *Inquiry Science Teacher Questionnaire* (ISTQ), addressing teacher background, implementation of FAST in the classroom, the classroom context within which teachers implement inquiry science, and the support the school provides teachers to implement inquiry science.

Developing the Inquiry Science Teacher Log

The purpose of the ISTL is to examine the extent to which inquiry science teachers using the FAST program implement the key features of inquiry science. The purpose as originally conceptualized was to collect data on each activity within inquiry-science investigations. However, the planned scope of the log was reduced considerably from its earliest conceptualization. It was decided that detailed reporting would put an excessive burden on respondents, resulting in a low response rate. Furthermore, a detailed log might not leave room for asking questions about aspects

of inquiry that the FAST program developers deemed essential for fully-implemented inquiry science (e.g., about teachers' use of questioning strategies or about their facilitation of student investigation processes).

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 on 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 the agreement within one scale point was 100% on two items of nine, 86% on four of the items, 71% on one item, and 57% on two items. 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 1/5 of the items, and below 70% on 3/10 of the items).

The results of these studies suggest that (a) log items have been shown to be valid and 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 coefficients can be at acceptable levels, (c) validity is the greatest for broad 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.

We decided to ask about both teacher activities and a few student activities in the log. A draft was developed that asked for teachers to respond about given investigations, with a focus on those that we are observing. The items addressed what the program developers believe are the key features of good inquiry science, including (a) disruptions by activities inside or outside the classroom, (b) the number of class periods it takes to complete the investigation, (c) the customization of the investigation by using supplemental materials or any other method, (d) the adequacy of materials and equipment, (e) students' questioning behaviors, (f) the teacher's use of questioning strategies, (g) the teacher's circulation about the classroom, and (h) the teacher's discussions about variations in the data. The log was reviewed on multiple occasions by project staff and the program developers. It was pilot-tested using a quasi-cognitive interview method with a sample of four existing FAST teachers, one on two occasions. Observer data, which we will use to help validate the log, are currently being

collected from a sample of 19 teachers.

Developing the Inquiry Science Teacher Questionnaire

The purposes of the ISTQ are to gather information annually on the breadth and depth of the implementation of inquiry science in the classroom and on contextual variables that might affect instruction. The major variables that the questionnaire addresses are

1) teachers' implementation of 26 key features of inquiry science;
2) the adequacy and availability of science labs, equipment, textbook, and other materials;
3) teacher demographics such as 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 so forth;
4) teacher attitudes toward science;
5) the extent to which the teacher shows interest in science by participating in science activities outside of the classroom, including taking science PD courses;
6) the extent to which the teacher takes responsibility for student learning,
7) teacher collaboration with other teachers in the school and participation in decision making;
8) planning for and customization of FAST in the classroom;
9) the extent to which the teacher provides students with extra support;
10) class and classroom characteristics such as the average class size, ability grouping, and so forth;
11) the FAST investigations that the teacher has taught or plans to teach during the year; and
12) the teacher's perception of some student characteristics such as behavior and perseverance.

The variables were identified through the systematic, iterative process described earlier in this paper. To identify potentially useful items, we reviewed instruments from several national studies or ongoing federal data-collection efforts such as Reform Up Close, the Survey of Enacted Curriculum, the Third International Mathematics and Science Study, the National Longitudinal Educational Survery:1988 the Study of Instructional Improvement, the Schools and Staffing Survey, the Longitudinal Evaluation of School Change and Performance (LESCP) in Title I Schools, and the USDOE's Fast Survey Response System, for potential items. We selected items that addressed the contextual variables from these instruments and revised them to fit the purposes of our questionnaire, and we wrote items addressing key features of inquiry science, as defined by FAST. Some items about implementation from the log were included verbatim on the questionnaire. Multiple items were prepared for variables that could be measured with scales.

We obtained a copy of the cognitive-interview procedures manual developed by AIR for a large-scale national survey on the implementation of standards-based reform and Title I, and we adapted the procedures to 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 cognitive interview procedures were piloted-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, we had revised 83 items, added 4, and deleted 9.

Collecting and Analyzing Questionnaire Validation Data

After developing our instruments, we collected questionnaire validation data from current FAST teachers in Hawai‘i and across the mainland. All FAST teachers are required to participate in two-week institutes before CRDG will sell FAST materials in schools. We prepared a list of a total of 948 teachers in the U. S. who had been trained in FAST 1 since 1997. When we reached the teachers, we asked them if they would be willing to complete the ISTQ and to complete an ISTL each time they finished a FAST investigation. 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 up to five logs. Of the 948 teachers, 183 agreed to participate and 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 seven years previously and were deemed the most difficult to reach. Of the 183 who agreed to participate, 81 eventually completed questionnaires, 71 completed one or more logs, and 66 completed both the questionnaire and the log. The 81 teachers comprised 14% of the 563 teachers who we contacted—a percentage that compares favorably with surveys of this type (e.g., in market research).

We analyzed the demographic characteristics of the 81 teachers and compared them with the characteristics of the population of teachers nationwide that was described by the National Center for Education Statistics (2005). The sample is fairly representative but includes more male teachers and more private school teachers who teach smaller classes than teachers nationwide. (FAST has traditionally attracted teachers from private institutions.) Of the 81 teachers, 62% were female (compared with 79% nationwide), 70% taught in public schools (compared with 89% nationwide). The mean age was 42 (compared with 46 nationwide), and 58% reported having Master’s degrees (compared with 56% nationwide). The mean number of students per class was 22 (compared with 28 for secondary teachers and elementary teachers in departments nationwide). The median years of teaching experience in K–12 schools was 12 (compared with 14 nationwide for any kind of teaching experience), and the mean salary was between $40,000 and $50,000 (compared with $43,262 nationwide).

Validity has to do with the adequacy of inferences made from data (Messick, 1989). Messick (1989, 1995) presented a unified theory of validity, in which construct validity has six aspects, including the content, substantive, structural, generalizability, external, and consequential aspects. “In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement” (Messick, 1995, p. 744). In this section, we describe how we addressed each of these except the generalizability and consequential aspects. The generalizability aspect has to do with “the extent to which score properties and interpretations generalize to and across

population groups, settings, and tasks" (Messick, 1995, p. 745); this topic is best examined with large datasets (unlike our dataset of 81 respondents) or multiple independent datasets. The consequential aspect has to do with the effects of flawed instrument development on the use of data collected with the instrument. We have yet to use our data—that is, to make decisions about its meaning for inquiry science. The aspects that we examine apply most directly to the items comprising subscales (e.g, the implementation subscale) that we have developed and less to the items about demographics or background information. We do not examine all the subscales included in the instrument but believe we address those that are the most critical for a validity review.

The Content and Substantive Aspects of Validity

The content aspect addresses "content relevance, representativeness, and technical quality" (Messick, 1995, p. 745), and the substantive aspect addresses the extent to which the information elicited in responses to items addresses the intent of the items. Supportive validity evidence for the content aspect is found in our description in this paper of the development of the questionnaire and log, in which we (a) gave an overview of the extensive literature reviews and consultation with FAST experts that formed the basis for selecting item content, (b) showed how we helped ensure item quality by borrowing from national surveys when appropriate, (c) described how items were reviewed multiple times by FAST experts and our advisory board, (d) stated that all the items went through multiple revisions, and (e) outlined subscale reliability analyses. Supportive evidence about the substantive aspect is found in our description of the think-aloud protocols for the log and questionnaire, in which we asked pilot-test respondents to describe aloud their mental processes while they answered the items.

We plan to conduct factor analyses on the subscales for which there are enough items to justify using the technique. Items that do not load highly on factors will be examined to determine how well they address the construct that the subscale is intended to measure.

The Structural Aspect

The structural aspect of Messick's unified validity theory addresses two issues. The first issue has to do with the method for accumulating item scores. We represent our subscale results with total scores. An alternative would be to use factor scores, which we intend to explore for those subscales for which there are few enough items to warrant factor analyses of our small dataset.

The second issue has to do with the extent to which the relationships among subscale results are as expected. To examine this issue, we reviewed the relationships among three subscales that previous research has shown are aspects of school capacity for learning. The subscales measure the extent to which teachers collaborate with each other (Cronbach's $\alpha = .91$), the extent to which they participate in professional activities outside the classroom (in our case, in science activities; $\alpha = .76$), and school support for program implementation (in our case, support for FAST; $\alpha = .81$). Organizational learning research has shown that schools that excel in these three variables tend to have a high capacity for learning and are more likely to be high-functioning schools. We included these variables in our study because we believe they might affect program implementation.

The correlations among results on the three subscales addressing these variables are shown in the blue-colored cells in Table 1. They partially confirm our expectations. Teacher collaboration is correlated with school support, but teacher participation shows virtually no correlation with the other two. Teacher collaboration and school support reflect school culture, whereas teacher participation apparently reflects teachers' individual professional activities. 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 organizations, as well as an item about the number of hours taking 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 subscale 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. We conclude that our results show validity evidence for considering collaboration and school support as aspects of a school culture factor. We should not include teacher participation in this factor, however.

The External Aspect

The external aspect has to do with how well constructs are measured similarly with multiple methods. This is a multitrait/multimethod analysis of the variables that the log and questionnaire instruments have in common and is useful for using log data to validate questionnaire data. These variables are shown in green and in yellow on Table 1. The analysis is limited by the fact that five of the eight measurements of the four constructs are measured as items and not as subscales. Customization and adequacy (Nos. 2, 3 6, and 7 on Table 1) do not lend themselves to measurement as subscales. Because item results are much less reliable than subscale results, our pattern of correlations is less likely to be interpretable than if all the items were measured as subscales. The analysis is also limited by the fact that, strictly speaking, it is not a multimethod study, because both the log and questionnaire are teacher self-reports. However, the instruments vary on when teachers complete them (the log as soon as they finish teaching lessons and the questionnaire once for the year), which we believe makes our analysis a form of multimethod study.

We conducted three analyses addressing the external aspect. The first was a monotrait/multimethod examination of the correlation between questionnaire implementation and log implementation. Given that the log measurement of implementation is the average of self-reports made soon after completing investigations and the questionnaire measurement of implementation is done once for the year, a high correlation will validate the questionnaire results. As shown in the green cells on Table 1, the correlation is .66. This high correlation is higher than any other correlation shown on Table 1. It validates the questionnaire measurement of implementation.

The second analysis for the external aspect was a comparison of the set of multimethod correlations (shown in the diagonal row of red-shaded cells in Table 1) with the two sets of monomethod correlations (shown in the two triangles of grey-shaded cells). This analysis was a review of three constructs that the log and the questionnaire had in common: the extent to which materials for teaching FAST were adequate, the extent to which lessons were customized, and the extent to which the teachers used supplemental materials to teach the lessons. To provide validity evidence, the correlation for a variable across the log and questionnaire should be greater than the correlations of this variable with other variables within either instrument. We found that the correlation of customization across methods was higher than the correlation of customization with other variables within methods. The same result was found for adequacy of FAST resources. However, supplementation showed somewhat varied results: The supplementation item on the questionnaire was more highly correlated with the customization on the questionnaire than it was with supplementation on the log.

The third analysis was a comparison between the two sets of monomethod correlations. For

each pair of variables, validity evidence is found when the rank of correlations is consistent in both sets of correlations. We found that the correlations between customization and supplementation were the highest in both sets of correlations, as expected. These two measurements address similar constructs. Furthermore, they both show negative correlations with adequacy of resources, which also is in the direction expected. Their ranks with adequacy of resources across instruments vary, probably reflecting the differences in measurement between the two methods. (Adequacy on the questionnaire is measured by a seven-item subscale and on the log is measured by a single item). Given that both methods show negative correlations, we believe that the multimethod analysis provides evidence for the validity of the questionnaire.

Conclusions About Validity

In conclusion, our instrument development methods provide good validity evidence for the content and substantive aspects. Our log results, summarized from data collected several times for most teachers, provide validity evidence for our questionnaire results, which are collected once: (a) the log implementation subscale supports the validity of our questionnaire implementation subscale, (b) teacher collaboration and school support are correlated as expected, (c) the correlations on single variables between methods are higher than the correlation of these variables with other variables, and (d) the patterns of correlations tend to be similar across methods. The results suggest that we should not consider teacher participation in science activities outside the classroom to be an aspect of school culture. Altogether, we believe that the findings are positive validity evidence for the questionnaire.

Further analyses must be done. We need to conduct validity studies of logs by correlating direct observers' results on the log with teachers' results on the log. We need to compare log and questionnaire results with observation results. Additional structural analyses that examine the relationships between predictor and criterion variables within the set of questionnaire results would be valuable (e.g., examining how well selected variables predict implementation), but care must be taken not to go on fishing expeditions without strong theory (Messick, 1989). Until we have strong theory about what predicts inquiry science implementation, we will very cautious about these kinds of analyses. We plan to continue our analyses and present the full set in a later report.

References

Ball, D. B., 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.

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

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

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.

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

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

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

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

Creemers, B. (1993). *Toward a theory on educational effectiveness*. (Report No. EA025038). Paper presented at the annual 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., Reynolds, D., & Swint, F. (1996). *Quantitative and class study data, 1992-1994.* (Report no. EA027569). The International School Effectiveness Research Project. The Netherlands: Groningen Institute for Educational Research. (ERIC Document Reproduction Service No. ED395380).

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*(1), 23–45.

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

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.

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.

Fullan, M., & Stiegelbauer, S. (1991). *The new meaning of educational change.* (2nd ed.). New York: Teachers College.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. *Review of Educational Research, 47*, 335–397.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review, 18*(1), 37–50.

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

Harlen, W. (2004). The development of assessment for learning: Learning from the case of science and mathematics. *Hodder Arnold Journals*, *21*(3), 390–408.

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.

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

The Inquiry Synthesis Project, Center for Science Education, Education Development Center. (April, 2004). *Technical report 2: Conceptualizing inquiry science instruction*. Retrieved March 14, 2004 from Education Development Center, Inc. Web site: http://www.cse.edc.org/work/research/technicalReport2.asp

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.

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 annual meeting of the American Educational Research Association, Montreal.

Mayer, D., Mullens, J., Moore, M., & Ralph, J. (2000). *Monitoring school quality: An indicators report.* (NCES 2001-030). Washington, DC: National Center for Education Statistics.

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.

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

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 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*(1), 19–35.

Muthen, H., Huang, L. C., Jo, B., Khoo, S. T., Goff, G. N., Novak, J. R., & Shih, J. C. (1995).

Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis*, *17*, 371–403.

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

National Center for Education Statistics. (2005) *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 Research Council (2000). *The national science education standards.* Washington, DC: National Academy Press.

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.

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

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)

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

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

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

Pottenger, F. M. III, & 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.

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

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

Rowan, B., Harrison, D. M., & Hayes, A. (2003). Michigan: University of Michigan, College of Education.

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

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.

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., 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.

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

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

Teddlie, C. & Reynolds, D. (2001). Countering the critics: Responses to recent criticisms of school effectiveness research. *School Effectiveness and School Improvement*, *12*(1), 41–82.

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.

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.

Table 1. Correlations for Analysis of the External and Structural Aspects of Validity[a]

Variable	1	2	3	4	5	6	7	8	9	10
1. Implementation level (26-item questionnaire scale, α = .93)										
2. Extent teacher customized FAST (questionnaire item)										
3. Extent teacher supplemented FAST (questionnaire item)		.41								
4. Adequacy of FAST resources (seven-item questionnaire scale, α = .70)		-.03	-.19							
5. Implementation level (five-item log scale, α = .72)	.66									
6. Extent teacher customized FAST (log item)		.45								
7. Extent teacher supplemented FAST (log item)			.26			.24				
8. Adequacy of FAST resources (log item)				.27		-.14	-.04			
9. Collaboration among teachers (nine-item questionnaire scale, α = .91)										
10. Teacher participation in science activities outside classroom (four-item questionnaire scale, α = .76)									.02	
11. School support for FAST (three-item questionnaire scale, α = .81)									.35	-.06

[a]81 teachers responded to the questionnaire subscales and items that are correlated only with each other, and 66 teachers responded to the subscales and items that are correlated between logs and questionnaires.