American Journal of Evaluation
Volume 29 Number 3
September 2008 235-250

10.1177/1098214008319175
http://aje.sagepub.com
hosted at
http://online.sagepub.com

The Complexity of Measuring the Quality of Program Implementation With Observations

The Case of Middle School Inquiry-Based Science

Paul R. Brandon
Alice K. H. Taum
Donald B. Young
Francis M. Pottenger III
Thomas W. Speitel
University of Hawai'i at Mānoa

In the growing literature on the evaluation of program implementation, less has been said about evaluating program quality than about evaluating other aspects of program implementation. Furthermore, most articles and reports in the program-implementation evaluation literature have presented only brief descriptions of how implementation instruments have been developed. In this article, the authors describe a method for evaluating the quality of implementation of middle school inquiry-based science using data from observations scaled with paired comparison judgments. The authors show the complexities of developing and applying the method, describe how they tried it out, present the results of validity and reliability analyses, and describe the method's strengths and weaknesses.

Keywords: *program implementation; evaluation of quality; inquiry-based science; observations*

Evaluation theorists, methodologists, and practitioners are increasingly calling for measuring program implementation when conducting evaluations. Some of the published descriptions of measuring implementation use observations, an implementation data-collection method that is complex and laborious but necessary for examining implementation quality. Many of the reports of observational methods to measure implementation, however, do not present thorough, detailed descriptions of the development, trial, and validation of the methods. The descriptions of conducting observations in implementation studies have been so brief that evaluators (particularly novices) have been unable to glean a full understanding of the complexities of, and the difficulties encountered in, developing and using observation methods.

In this article, we address this deficit in the literature. The purpose of the article is to describe in detail the development, trial, and validation of an observational method for

Paul R. Brandon, Professor, Curriculum Research & Development Group, University of Hawai'i at Mānoa, 1776 University Avenue, ULS 2-214A, Honolulu, HI 96822-2463; phone: (808) 956-4928; e-mail: brandon@ hawaii.edu.

Authors' Note: This article summarizes some of the results of a project funded by the National Science Foundation (Grant No. REC0228158). 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.

measuring program implementation quality in a manner that is intended to assist evaluators who seek full accounts of developing and using such methods. We accomplish this by presenting an observational method for studying the quality of middle school, inquiry-science program implementation and by examining its strengths and weaknesses. Many years of research have demonstrated the positive effects of inquiry-science teaching and learning (Cohen & Spillane, 1993; Lee, Hart, Cuevas, & Enders, 2004; Lott, 1983; National Research Council [NRC], 1996; Shymansky, Kyle, & Alport, 1983; Tamir, 1983; Von Secker & Lissitz, 1999; Wise & Okey, 1983; Wu & Hsieh, 2006). This article adds to the growing body of research (e.g., Lynch & O'Donnell, 2005; O'Donnell, 2008) on studying the implementation of science in K-12 classrooms.

Background on Assessing Program Implementation

The evaluation of program implementation has to do with examining the extent to which programs are implemented with *fidelity* (e.g., Bond, Evans, Salyers, Williams, & Kim, 2000; Mowbray, Holter, Teague, & Bybee, 2003; O'Donnell, 2008) to the intended program. Evaluators and researchers have examined a number of aspects of implementation in their studies. Probably the most widely studied aspects of implementation have been *adherence*, *exposure* (sometimes called *dose*), and *quality* (Dane & Schneider, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003; O'Donnell, 2008). Adherence has to do with the extent to which the steps and procedures of a program are delivered as intended, exposure has to do with the frequency of program units, and quality has to do with the how well a program implements the techniques or methods of the program (O'Donnell, 2008). Adherence and exposure can be measured as frequencies of implementation and are typically examined with self-report questionnaires or teacher logs, whereas quality requires a qualitative assessment of merit by an external observer. Examining quality in general is a "search for goodness and badness, for merit and shortcoming" (Stake, 2001, p. 3); examining the quality of classroom implementation of a program involves evaluating the extent to which teachers' instructional strategies show deep and sophisticated application of best classroom practices in a manner consistent with the intentions of the program developers.

Quality is more difficult to measure than adherence and exposure. Collecting data on adherence and exposure requires that evaluators measure how fully or frequently discrete steps, units, or components are implemented, which can be a simple, quantitative checklist task. Logs for reporting adherence and exposure typically need minimal development time, because (a) the definitions and descriptions of the steps and units are usually well known to the program personnel, (b) program personnel can provide the data without the assistance of evaluators or the need for raters, and (c) minimal instrument pilot-testing is needed. In contrast, collecting data on quality requires that evaluators measure how well the steps, units, or components are implemented—a complex judgment task that should not be measured with logs because of self-report bias and because people are not good judges of the quality of their own activities. Methods for collecting data on quality require considerable development time and financial resources, because evaluators choosing to use the best methods must carefully identify criteria for judging quality, develop forms for conducting ratings or other judgments, develop training procedures, videotape program sites or have observers visit sites, and make judgments about quality.

It is apparent from our wide sampling of the implementation literature, and in particular from our examination of implementation literature reviews, that of the aspects of implementation, quality has been addressed the least (Dane & Schneider, 1998). For example, Mihalic

(2002) identified a number of studies that examined the relationship between program outcomes and adherence, exposure, or quality and found the fewest number of studies examining the relationship between outcomes and quality. Dusenbury, Brannigan, Hansen, Walsh, and Falco (2005, p. 309) went so far as to state, "A methodology for examining the role of quality of implementation under real-world conditions has not yet been developed." In part, the relative dearth of studies of implementation quality might be because of the expense and extensiveness of the resources that are necessary for measuring quality (or, indeed, of any process variable similar to quality). Broad and deep studies of quality require observations, the most resource intensive of the common data-collection methods in educational and social science evaluation.

Background on Inquiry Science

The implementation-quality method that we discuss in this article was prepared as part of a National Science Foundation (NSF) project to develop and validate instruments addressing the implementation and outcomes of middle school inquiry science. K-12 inquiry science is based on the constructivist theory that students develop knowledge incrementally and through interaction with teachers and other learners. Students answer questions about natural phenomena or events by conducting scientific investigations in which they develop designs, collect and explain data, connect explanations of the results to existing scientific knowledge, and communicate the explanations (NRC, 1996). Inquiry-science students work in groups that mimic teams of professional scientists, deciding about scientific processes, procedures, analyses, and outcomes. The form of these steps depends in part on the extent that students conduct their investigations independent of teacher input. Inquiry-science teachers' role is to structure opportunities for their students to learn in mini-scientific communities, provide their students with materials and instruments, and use questioning strategies to guide the students. The National Science Education Standards (NRC, 1996, p. 32) describe how inquiry-based teachers "focus and support inquiries while interacting with students," "orchestrate discourse among students about scientific ideas," "challenge students to accept and share responsibility for their own learning," "recognize and respond to student diversity and encourage all students to participate fully in science learning," and "encourage and model the skills of scientific inquiry, as well as the curiosity, openness to new ideas and data, and skepticism that characterize science."

The version of inquiry science that we examined was the Curriculum Research & Development Group's (CRDG) Foundational Approaches in Science Teaching (FAST) program, a 3-year, interdisciplinary, middle school inquiry-based science program (called here an *inquiry-science program*) that is aligned with the National Science Education Standards (CRDG, 1996; Rogg & Kahle, 1997). FAST has been evaluated in several settings using a variety of designs and outcome measures, but its implementation has not been previously examined in summative evaluations.

FAST models the experience of practicing scientists. Students working in research teams develop hypotheses, do physical experiments, organize and analyze data, and develop a consensus about conclusions. FAST teachers serve as "research directors," questioning, stimulating, facilitating, and probing students as they conduct science investigations.

Experienced inquiry-science teachers use a variety of questioning strategies. 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., Coker, Lorentz, & Coker, 1980; Gall, 1970, 1984; Hamilton & Brady, 1991; Redfield

& Rousseau, 1981; Samson, Sirykowski, Weinstein, & Walberg, 2001; Soar, 1973; Stallings & Kaskowitz, 1974). The FAST developers believe that questioning is the heart of inquiry science and that teachers' use of questioning strategies is the most essential component of inquiry-science teaching. Questioning is the preferred method of interaction in inquiry-science classes because the teachers' role 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.

FAST teachers learn questioning strategies in professional development institutes provided for the program. Traditionally, teachers have been required to participate in a 2-week FAST professional development institute for any of the 3 years of the program (FAST I, II, or III) before they have been able to buy and use FAST materials for that course. In the institutes, teachers conduct all FAST investigations for the 1-year course of study, are immersed in science investigations that model the various teaching behaviors inherent in FAST, and are provided opportunities for reflective discussions of the learning, teaching, and assessing experiences. Trained, certified instructors model teaching strategies when the teachers do the science investigations that their students would undertake. Recently, a 1-week version of the training has been developed with an increased focus, among other topics, on the effective use of strategies to question students during the science investigations.

Developing the Inquiry Science Questioning Quality Method

To address the quality aspect of inquiry-science program implementation, we developed the Inquiry Science Questioning Quality (ISQQ) method. The ISQQ uses observation methods to judge teachers' quality of implementation (described in detail below) and analyzes the results of the observations with the paired comparison method (David, 1963; Torgerson, 1958).

The ISQQ is one of a suite of instruments that was developed in an NSF project to examine the implementation and outcomes of inquiry science. Others include the Inquiry Science Implementation Questionnaire (ISIS; Brandon, Young, Taum, & Pottenger, 2008), a log used to collect data for validating the ISIS results, the Inquiry Science Observation Coding Sheet (ISOCS; Brandon, Taum, Young, & Pottenger, in press), and the Inquiry Science Student Assessment Suite (ISSAS; Ayala, 2005). The implementation instruments were developed to collect data on aspects of implementation other than quality, including adherence and exposure. Because of its expense, which is apparent in the description below, the ISQQ is probably most appropriate for collecting implementation data after teachers have been sufficiently trained and have had sufficient time to become proficient with inquiry science. This is a use of the instrument to capture the best quality.

Expert judges use the ISQQ to evaluate the quality of the implementation of questioning strategies in inquiry science. The judges are trained in the criteria for judging the teachers, view videotapes of samples of teachers' behavior, take extensive notes about the teachers, compare each teacher with other teachers individually using the paired-comparison method, and record a preference vote for each pair. The method requires the judges to consider each teacher in light of the performance of each other teacher. The judges' conclusions about a teacher's performance are made in light of other teachers' performances, potentially making more rich, holistic, nuanced assessments of worth or merit than are possible with rating rubrics. Thus, the judgments are relative, and the results of the analysis do not provide information about performance on a scale of absolute standards, such as a rating rubric. The method produces an ordinal scale.

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

Developing 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 could accurately and reliably compare a sample of videotapes of FAST teachers.

When developing the criteria that the ISQQ addresses, we built on our development of other instruments in our inquiry-science NSF project. One of these instruments was the companion observation protocol, the ISOCS. When developing the ISOCS, the instrument developers and the program developers had narrowed the number of observed characteristics of implementation during approximately 40 develop–tryout–review–revision iterations. The ISOCS developers strove during these iterations to narrow the targeted teacher and student behaviors to a feasible list reflecting the aspects of inquiry science that the program developers deemed critical for program success. Eventually the ISOCS came to focus almost exclusively on teacher questioning, because inquiry-science teachers interact with students primarily by questioning them throughout all steps of their science investigations. Questioning is the preferred method of interaction because inquiry-science teachers should not instruct students directly but simply guide them as they learn science. Inquiry science is successful in large part by the extent to which teachers proficiently use question-asking strategies.

The other instrument from which we drew when developing the ISQQ was the ISIS, a teacher questionnaire. The ISIS addresses several features of conducting inquiry-science investigations, including asking questions. When developing it, we reviewed the FAST instructional guide (Pottenger & Young, 1992a), student book (Pottenger & Young, 1992b), and teacher's guide (Pottenger & Young, 1992c); identified potential variables for the ISIS; and conducted an iterative review and revision of the variables during the course of several months. From the list of variables, we chose those that all FAST student investigations had in common. Program experts reviewed the list several times, with revisions between reviews. Eventually, the list of variables was deemed to address the essential features of FAST inquiry science, including aspects of teachers' use of questions.

The results of the ISOCS and ISIS development helped us focus the preparation of the ISQQ. We were assisted in ISQQ development by an FAST monograph on inquiry science (Pottenger, 2005) that was being prepared independently during the same time period. Drawing on all this work, the senior FAST developer prepared a draft description of teacher questioning. The second senior FAST program developer and the project researchers reviewed and revised the description for clarity, conciseness, and thoroughness. After several iterations of development, review, and revision, the team's final version of the criteria, shown in the appendix, was finalized.

Preparing 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 3-day meeting. 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, was 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.

Trying Out the ISQQ and Conducting Validity Studies

Collecting Data for Judging Quality

To provide data for developing and validating the ISQQ, we videotaped FAST physical science in the classrooms of a sample of 16 public- and private-school teachers on the four major Hawaiian islands during the 2004-2005 school year. The videotaping targeted the five FAST physical-science student investigations that occurred at key junctures in the sequence of 14 student investigations on buoyancy and density (Shavelson et al., in press). We hired part-time employees on each island, trained them how to videotape lessons, and provided them with video cameras and other equipment, including boom and lavaliere microphones, digital cassette tapes, tripods, 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 videotaping personnel recorded comments about events and activities at the school that might have affected the class during the taping sessions and noted the sessions in logs. The guidelines, checklists, and logs are described in greater detail in Brandon et al. (2007). We asked the teachers to keep the videotaping personnel apprised of their progress through the early FAST 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 lessons 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 per teacher. We transferred the videos to DVD-ROMs (one DVD-ROM per class period). We then viewed samples of every 5-minute increment of every DVD-ROM and classified the quality of audio and video of the teacher. These quality checks showed that we had 91 DVD-ROMs 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 DVD-ROMs were deemed to be inadequate for coding the classes.

Selecting a Sample of Teachers and Observation Segments to Judge

The first step in our validity study of the ISQQ was to select a sample for the judges to examine teachers' use of questioning during inquiry-science classes. DVD-ROMs were selected from those of the nine teachers for whom we had taped the most student investigations. For all but one of the teachers, only one investigation was selected, and for each of the five investigations taped at the key junctures, two teachers were sampled. (See Brandon et al., 2007, for a table showing the sample.)

Segments of DVD-ROMs were selected for viewing and judging in paired comparisons. A FAST scientist/educator with expertise in videotape editing reviewed all the sampled DVD-ROMs and selected approximately 15 minutes of the three phases of FAST student investigations

(the Introduction phase, the Investigation—or 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 cleanup 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.

Recruiting Judges

The next step was to recruit judges. We chose to recruit five; our rationale was that this number was used in a previous evaluation study using paired comparison methods (Heath & Brandon, 1982) and that it was more than a sufficient number because many observational studies of implementation use fewer. We wanted to ensure that at least one pair of our judges was consistent among themselves and deemed that of five, we would identify at least a pair. The judges were FAST experts from five states who had taught FAST and served as FAST trainers. All agreed to participate in a 3-day quality-judging workshop immediately after a 3-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.

Training the ISQQ Judges

The five judges were trained and the ISQQ was administered during a 3-day period. On Day 1, the judges were introduced to the study and were trained in how to use the quality criterion statement. The day 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 flip chart was used to record comments when appropriate. The project's principal investigator served as the meeting facilitator; the project manager/project researcher and the project co–principal investigator (one of the two lead FAST developers) participated in the discussion among the judges.

The training began with a description of the purpose of the study, the overall NSF project, and the role 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. This step was included to help establish the content validity of the criteria.

Next the judges were trained in how to apply, in two steps, the quality criteria to DVD-ROMs of teachers. 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 his opinion, one at a time and without interruption, about the quality of questioning that the teacher displayed. Next the group discussed the members' 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 another 15-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 DVD-ROM. 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 DVD-ROM segments individually and to compare each teacher with other individual teachers. However, because there were a limited number of good-quality segments (not including the DVD-ROMs that were to be judged later), we were unable to view more DVD-ROMs.

When the viewing was complete, the meeting concluded at about midday with instructions for the tasks for the second phase of the process. The judges were told that (a) their task until noon of Day 3 was to view three 15-minute samples for each of nine teachers, (b) they were to work at places of their choosing except where laptops might be damaged, (c) they were not to discuss any of their work with each other, (d) they were to take notes about the extent to which each teacher addressed the quality criteria, (e) their notes should address all aspects of the criteria, (f) they should write summary statements for each teacher, (g) they should review the DVD-ROMs as many times as necessary to make global judgments of quality, and (h) they would reconvene after lunch on Day 3 to make judgments about quality, using a method that would be described at the time. The judges were shown how to access the DVD-ROMs 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). On Day 2 and the morning of Day 3, the judges independently observed the DVD-ROM segments for each teacher.

Making the Paired Comparison Judgments

In the early afternoon of Day 3, the judges made paired comparison judgments. In paired comparisons, each member of a set of objects is paired with each other, and trained judges select the member of each pair that addresses a specified criterion the most. This is a *preference vote*.

Rationale for using the paired comparison method. The paired comparison method has a long, venerable history, dating back to the psychophysical work of Fechner in the 1860s (David, 1963) and reflects Thurstone's (1927) law of comparative judgment. It has been used in consumer choice tests (e.g., taste tests), personnel ratings, establishing the anchor points on measurement scales, examining the seriousness of crimes, patients' mental health, visual illusions, life goals, food preferences, political ideology, personality research, and sociometric measurement (Rounds, Miller, & Dawis, 1978). In addition, it has been used at least once in an educational evaluation for judging the relative performance of 15 small school districts on six attributes of the districts' special education programs (Heath & Brandon, 1982). Heath and Brandon (1982) found that the results for three of five criteria were highly consistent among the five participating judges and distinguished well among most of the districts. The findings on the other two criteria suggested that within-site variability affected the judgments more than between-site variability.

The paired comparison method produces scales of objects (in our case, teachers) that are based on the comparison of every object with every other object by several judges. This results in a more nuanced conclusion about the scaled object than does a simple comparison of the object to the characteristics described on a rubric and an assignment of a score based on that comparison. In paired comparisons, all judgments are made in light of the judgment of all other objects, one object at a time. The method grounds the judgments in the context of a program.

A primary value of the paired comparison method is its use of the human observer as the data collection instrument (e.g., Guba & Lincoln, 1988). As data collection instruments, humans are responsive and adaptable to the data collection context. They are capable of absorbing and processing a continuing stream of information about the objects of observations—for example, teachers conducting multiple tasks in multiple settings on multiple occasions. Judgments are based on the dynamic teacher–student interaction; they are not bound by the constraints of tightly stated rubrics. In the case of the ISQQ, judges keep in mind not only all the characteristics of good questioning but also the context within which teachers ask questions. They make holistic judgments, in which they consider all aspects of the observed object simultaneously when comparing teachers (Guba & Lincoln, 1988).

Description of the ISQQ paired comparisons. The ISQQ judges brought the notes that they had 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 were to compare each teacher with each other teacher and judge (a) which member of each pair showed greater quality; 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 about 15 to 30 minutes.

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, as follows:

1. Viewing the two training samples was sufficient to feel comfortable about assessing quality.
2. Viewing the videotaped segments took from 1 to 2 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 note-taking.
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 strong 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 teachers waited long enough for answers to questions and whether the teachers missed questioning opportunities. One judge said that he frequently had to continue to return to the statement of quality criteria because he had additional criteria of his own in mind when viewing the video 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 students' behavior. Another tended to look at the students to see if they were engaged. One approach was to listen but not to watch.
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 and Reliability Analyses

Reports of observations should include evidence of validity and reliability (Evertson & Green, 1986; Herbert & Attridge, 1975; Hintze, 2005). Minimally, evidence of content validity should be included. Other sources of validity evidence might be from convergent, discriminant, or criterion validity studies. Reliability evidence for paired comparison data is found through analyses of interobserver agreement or reliability. Ideally, reports of observations of the quality of implementation will discuss the likelihood of error caused by central tendency error (i.e., rating objects toward the middle of scales), leniency error (rating objects favorably because of bias), the recency effect (having judgments be based primarily on recent events or in light of immediately prior influences), observer drift along a scale, misclassifications because of simultaneously observed events or behaviors, logical errors (i.e., making "judgment errors based on theoretical, experiential, or commitment-based assumptions" [Evertson & Green, 1986, p. 183]), misclassifications because of simultaneity of events, and observer drift (Evertson & Green, 1986).

We conducted validity and reliability analyses of the results of the paired comparison scaling. These results, described in this section, compare favorably with the levels of reliability and validity reported in the literature on measuring program implementation using observational methods.

Preparing and Scaling the Data

The validity analysis began with the analysis of the comparison results using the Thurstone Case 5 paired-comparison scaling method (Dunn-Rankin, Knezek, Wallace, & Zhang, 2004; Edwards, 1957), a common approach to analyzing this form of data. The Thurstone Case 5 method produces an ordinal scale with unequal distances among the scaled objects. The lowest scale value is set to zero, and the remaining values are linearly transformed to adjust to a scale value of zero for the worst-performing scaled object (in our case, teachers). The scale values are shown in Table 1.

Reliability of the Preference Data

Data cannot be valid unless they are reliable. We conducted two reliability analyses of the results of the paired comparisons, including Thurstone's absolute average discrepancy coefficient (Edwards, 1957; Gulliksen & Tukey, 1958; Heath & Brandon, 1982) and Guttman's coefficient of reproducibility (Edwards, 1957; Heath & Brandon, 1982).

1. Thurstone's absolute average discrepancy coefficient = .022. The coefficient is based on a comparison of the empirical proportions of preference votes Matrix 2 with theoretically expected proportions; the lower the value, the better the result. The coefficient that we found is comparable to values reported by Edwards (1957) and suggests fairly high reliability.
2. Guttman's coefficient of reproducibility = .80. This value indicates the percentage accuracy with which responses to the paired comparisons can be reproduced from ranks (Edwards, 1957). Our result indicates good reproducibility.

Table 1
Thurstone Case 5 Scale Scores (Analysis Matrix 2)

Teacher ID No.	Thurstone Scale Score
20	0.00
5	0.05
15	0.69
16	0.82
2	0.95
3	0.96
13	1.52
21	1.71
7	2.43

To examine further the agreement among judges, we calculated Spearman's ρ; correlations among the ranks that the judges assigned to the nine observed teachers, and we analyzed circular triads among the paired comparisons (Dunn-Rankin et al., 2004). The between-judge Spearman correlations 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. 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. The total number of circular triads (found using Dunn-Rankin et al.'s TRICIR software program) in the judges' preference data = 16. Of these, 11 were made about two of the judged teachers, and 8 were made by the judge whose Spearman's ρ; correlations with the remaining judges was the lowest of all five judges. Clearly, it tended to be difficult for some of 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 support for content validity is found in the manner in which the method was developed and conducted, as described earlier, including (a) the development of the criteria by inquiry-science experts, (b) the careful development of materials and guidelines for conducting the training and the quality judgments, and (c) the manner in which the training and judgments were conducted.

The judges' feedback at the conclusion of the workshop provides somewhat mixed evidence about validity, however. Evidence supporting validity includes the comments about the adequacy of the number of training samples, about the ease of taking notes while making judgments, and 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.

Convergent Validity

Convergent validity findings in the form of a correlation of the ISQQ results with the results for an alternative method for assessing questioning in FAST classrooms can provide additional validity evidence. Preferably, the alternative method should assess questioning analytically in contrast to the holistic approach of the ISQQ.

We compared the results on the ISQQ with the results on the ISOCS (Brandon, Taum, et al., in press). The primary purpose of the ISOCS is to measure the extent to which teachers use questioning strategies in inquiry-science classrooms. The instrument is a measure of the adherence aspect of implementation: It focuses primarily on the frequency with which teachers initiate questions and the frequency of teacher–student interactions that occur following the teachers' questions. The ISOCS was developed during about 2 years of approximately 40 review-and-revision cycles. All of the drafts of the instrument were tried out with trained coders working with the same DVD-ROMs that were used for the ISQQ. Eventually, coded behaviors were narrowed to those focusing on the question-and-answer cycle of Socratic inquiry (Pottenger, 2005). Two observers used the ISOCS to code the DVD-ROMs of the teachers whom we examined with the ISQQ and reconciled the code differences between them. For the purpose of validating the ISQQ data, we correlated the Thurstone Case 5 scale scores with two types of ISOCS results: (a) the percentage that codes for student comments constituted of all the codes that had been assigned to the teacher ($r = .52$) and (b) the percentage that codes for the teachers' use of follow-up statements and of probing questions constituted of all the codes that had been assigned to the teacher ($r = .45$). 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.

Conclusions and Strengths and Weaknesses of the Method

This article describes and discusses an observational method for collecting and analyzing data on the quality of program implementation in K-12 inquiry science. Several aspects of program implementation are commonly examined in evaluation studies. Some focus on breadth, and others focus on depth. The measurement of implementation quality addresses the latter: It focuses not on whether program steps and activities are implemented but on how well they are implemented.

Observations are necessary for measuring quality because they avoid self-report biases by providing external viewers' perspectives. The more that evaluators are fully aware of the steps of conducting observation studies, the better informed their method choices will be, and the more likely they will be to avoid methodological pitfalls. However, the methods of developing observation protocols, conducting observations, and analyzing observation data about program implementation are not often fully described. Thus, evaluators' understanding of observational methods is likely to be limited without digging deep into the literature (some of which is decades old) about conducting observations.

In this article, we endeavor to describe an implementation-quality observation method in sufficient depth for evaluators seeking information about the complexities of conducting observations to get a full picture of the process of development and application. For example, evaluators developing observations might find that, like us, they must winnow down their list of program features to examine the essential characteristics that most directly address program quality and are most likely to affect student learning and understanding. This winnowing down occurs, in part, because evaluation resource constraints

require diligence in the selection of quality characteristics. Observations are expensive, and careful examination of the program's core features is necessary if they are to be feasible with time and financial limitations. In the case of the ISQQ, we decided in several reviews of program documents and many iterations with the program personnel to focus on how well the teachers used questioning strategies when interacting with students during science investigations.

The paired comparison method potentially has many strengths that are commonly overlooked when researchers and evaluators prepare measurement scales. The method takes advantage of the comparative manner in which judges make decisions about quality. Instead of judgments about a teacher's quality being made in light of an abstract standard, they are made in light of the breadth of quality of all the teachers within a program and in light of the context within which the program is conducted and implemented. The nature of this form of judgment builds a more nuanced sense of quality than judging quality against a rating rubric. Furthermore, paired comparisons are not subject to central tendency error (i.e., rating objects toward the middle of a scale), leniency error (rating objects favorably because of bias), the recency effect (having judgments be based on recent events or in light of immediately prior influences), observer drift along a scale, misclassifications because of simultaneously observed events or behaviors, or observer personal bias. Finally, the method requires the judge to make holistic judgments, drawing on the strength of using humans as a measurement instrument.

One weakness of the paired comparison method is that the potential exists for logical error bias (i.e., the tendency to make errors because of inappropriate assumptions about inquiry science) while judging questioning quality. We have reason to believe that this error was manifested somewhat in our study, because (a) the correlations for two of the judges with the remaining three were lower than the correlations of the three judges among each other and (b) the circular triad results confirmed that the two judges were outliers. The potential for outliers is one of the reasons that we used five judges. The results confirm that our caution in selecting several judges was appropriate. Indeed, the presence and potential effects of outliers should always be a concern of evaluators.

Training potentially mitigates the effects of logical error bias. Despite the judges' feedback that they believed the two samples of videos were sufficient to learn how to apply the criteria, if we had provided more samples for the judges to view during training, some of the disagreement among judges' paired comparisons might have been mitigated. Furthermore, additional training might have reduced the tendency of judges to apply criteria for judging quality in addition to those presented on the quality criterion statement. If we had provided for longer training, it is likely that the judges' use of other criteria might have become apparent during the training period. In the future, we intend to add additional training segments when using the method. The degree to which we will achieve this goal will depend in large part on resources of time and money—both of which tend to limit most evaluators' ability to conduct observation studies, particularly in small evaluations. Clearly, the method we describe here cannot be used for small or poorly funded evaluation studies.

Finally, the paired comparison method in our study ordered teachers on quality but did not yield an interpretation of how good or bad that quality was. This, of course, is an disadvantage of not using a rating rubric. A remedy for this would be to develop and administer a method for judges to collaboratively define levels of quality at various points on the final Thurstone scale at the conclusion of the study.

Appendix
Inquiry Science Questioning Quality Criteria

The exchange of teacher questions and student responses is the sign of good, middle school inquiry-based 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 the following:
   a. asking clear, unambiguous questions at the appropriate opportunities for the purposes of initiating discussions and encouraging student curiosity,
   b. using Socratic question-and-answer chains,
   c. asking the children to reflect on possible answers to their own questions (e.g., "What do you think?"),
   d. asking questions that gain insight into students' behavior (e.g., "What might happen if did you X?"),
   e. asking questions about how investigations might be conducted (e.g., "How might that be found?"),
   f. asking questions that seek comparisons or contrasts (e.g., "How do these results compare with our previous results?" and " How are they different?"),
   g. asking questions about the sufficiency of evidence (e.g., "What is the evidence for that, and what is the quality of the evidence?"), and
   h. asking questions about connecting the findings to everyday life.

References

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

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.

Brandon, P. R., Taum, A. K. H., Young, D. B., & Pottenger, F. M., III. (in press). The Inquiry Science Observation Coding Sheet: Development and findings about validity. *Evaluation and Program Planning*.

Brandon, P. R., Taum, A. K. H., Young, D. B., Speitel, T. W., Pottenger, F. M., III, Nguyen, T. T., et al. (2007). *Final report of a Phase-I study of the effects of professional development and long-term support on program implementation and scaling up*. Honolulu: University of Hawai'i at Mānoa, Curriculum Research & Development Group.

Brandon, P. R., Young, D. B., Taum, A. K. H., & Pottenger, F. M., III. (2008). *The Inquiry Science Implementation Scale: Instrument development and the results of validation studies*. Manuscript submitted for publication.

Cohen, D. K., & Spillane, J. P. (1993). Policy and practice: The relationship between governance and instruction. In S. H. Fuhrman (Ed.), *Designing coherent education policy: Improving the system* (pp. 35-95). San Francisco: Jossey-Bass.

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

Curriculum Research & Development Group. (1996). *Alignment of Foundational Approaches in Science Teaching (FAST) with the National Science Education Standards Grades 5-8*. Honolulu, HI: 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.

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

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

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.

Dusenbury, L., Brannigan, R., Hansen, W. B., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understand the diffusion of preventive interventions. *Health Education Research*, *20*, 308-313.

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

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

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

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

Guba, E. G., & Lincoln, Y. S. (1988). *Effective evaluation*. San Francisco: Jossey-Bass.

Gulliksen, H., & Tukey, J. W. (1958). Reliability for the law of comparative judgment. *Psychometrika*, *23*(2), 95-110.

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.

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.

Herbert, J., & Attridge, C. (1975). A guide for developers and users of observation systems and manuals. *American Educational Research Journal*, *12*, 1-20.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review*, *34*, 507-519.

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.

Lott, G. W. (1983). The effect of inquiry teaching and advanced organizers upon student outcomes in science education. *Journal of Research in Science Teaching*, *29*, 437-451.

Lynch, S., & O'Donnell, C. (2005, April). *The evolving definition, measurement, and conceptualization of fidelity of implementation in scale-up of highly rated science curriculum units in diverse middle schools*. Paper presented at the meeting of the American Educational Research Association, Montreal, Canada.

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.

Mihalic, S. (2002). *The importance of implementation fidelity*. Boulder: University of Colorado at Boulder, Center for the Study and Prevention of Violence. Retrieved March 17, 2005, from www.incredibleyears.com/research/fidelity-importance.pdf

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.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, *78*, 33-84.

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 & Development Group.

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

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

Rounds, J. B., Miller, T. W., & Dawis, R. V. (1978). Comparability of multiple rank order and paired comparison methods. *Applied Psychological Measurement*, *2*, 413-420.

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.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., et al. (in press). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education*.

Shymansky, J. A., Kyle, W. C., & Alport, J. M. (1983). The effects of new science curricula on student performance. *Journal of Research in Science Teaching*, *20*, 387-404.

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 No. ED 106 297)

Stake, R. E. (2001). Representing quality in education. In A. Benson, D. M. Hinn, & C. Lloyd (Eds.), *Visions of quality: How evaluators define, understand and represent program quality* (pp. 3-12). New York: JAI.

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

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

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286.

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

Von Secker, C., & Lissitz, R. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching*, *36*, 1110-1126.

Wise, K. C., & Okey, J. R. (1983). A meta-analysis of the effects of various science teaching strategies on achievement. *Journal of Research in Science Teaching*, *20*, 419-435.

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