Development, Validation, and Trial of a Method for Judging the Quality of Using Questioning Strategies in a Middle-School Inquiry Science Program

Paul R. Brandon, Alice K. H. Taum, Donald B. Young,
Francis P. Pottenger III, Thomas Speitel, and Mary Gray
University of Hawai'i at Mānoa

Paper presented at the annual meeting of the
American Educational Research Association
Chicago, April 2007

Introduction and Background

The purpose of this paper is to describe the development, validation, and trial of a method for judging the quality of teachers' use of strategies for questioning students in middle-school inquiry-based science classes (called here *inquiry science).* Quality has been discussed in the program implementation literature as an aspect of implementation that should be examined in educational and social science research and program evaluation studies.

An NSF Study of Foundational Approaches in Science Teaching

The work described here was conducted by a team of researchers and curriculum developers at Curriculum Research & Development Group (CRDG), University of Hawai'i at Mānoa, as part of a project funded by a National Science Foundation (NSF) Interagency Educational Research Initiative grant (No. REC0228158). The purpose of the NSF project was to develop and validate a suite of instruments for measuring the implementation and student outcomes of Foundational Approaches in Science Teaching (FAST), a widely disseminated middle-school inquiry science program. FAST is aligned with the National Science Education Standards (CRDG, 1996; Rogg & Kahle, 1997) and has been shown in several studies to positively affect student outcomes (CRDG, 2000 [three studies]; Dekkers, 1978; Pauls, Young, & Lapitkova, 1999; Tamir & Yamamoto, 1977; Young; 1982; Young, 1993). It consists of three inquiry courses entitled, "FAST 1: The Local Environment," "FAST 2: Matter and Energy in the Biosphere," and "FAST 3: Change Over Time." Of these, the NSF instrument-development project focused on the density and buoyancy lessons in FAST 1.

Teachers' role in FAST and their questioning of students. FAST models the practice of science: Students working in small groups develop hypotheses, test them in experiments, report the results, and arrive at conclusions about their findings

consensually. FAST teachers are "research directors" who guide the students, primarily with questioning strategies. Questioning is the preferred method of interaction in inquiry-science classes because the teachers' role is not to directly instruct students but to guide them as they develop, implement, and interpret small scientific investigations. Inquiry science teachers, of course, interact with students without questioning them, but the primary means of helping students learn in a constructivist, hands-on fashion is by asking questions. Findings about the effects of teachers' use of questions vary among studies, but research in general has shown that teachers' proficient use of the appropriate questioning strategies improves student learning (e.g., Gall, 1970; Gall, 1984; Redfield & Rousseau, 1981; Samson, Sirykowski, Weinstein, & Walberg, 2001).

Studying Program Implementation

It is widely agreed (e.g., Dane & Schneider, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003; Ruiz-Primo, 2005) that evaluators assessing program implementation should address five aspects of implementation: *adherence*, *exposure*, *quality of delivery*, *participant responsiveness*, and *program differentiation*. Adherence is the extent to which program implementation follows the prescribed sequence, procedures, lessons, steps, and so forth; it addresses the detailed processes that occur in programs. 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. To evaluate adherence and exposure is to answer the question, "How fully was the program implemented?"; to evaluate quality is to answer the question, "How well was the program implemented?" Of these three aspects of implementation, we addressed the first two in instruments that we have previously reported on extensively (e.g., Brandon & Taum, 2005).[1]

In this paper, we describe a method for measuring the quality of implementation that we developed, validated, and tried out during the NSF project: the Inquiry Science Questioning Quality (ISQQ) method. The ISQQ is a paired comparison method (David, 1963; Torgerson, 1958). In paired comparisons, each object in a set

[1]We did not address participant responsiveness, which has to do with students' (or other program beneficiaries') engagement in a program, because it was beyond our resources to do so, and we did not address program differentiation because it has more to do with a feature of good design that any experiment should address than it has to do with program implementation.

of objects is paired with each other object a pair at a time, and trained judges select the member of each pair that addresses a specified criterion more than the other member. This is a *preference vote*. For example, the first author used the paired comparison method to compare a group of schools (the *objects*) on each of several criteria that defined effective special education; analyses of the results showed that the paired comparisons were conducted by two observers reliably (Heath & Brandon, 1982). The method results in objects on a scale that yields both their criterion rankings and the distances between the objects (i.e., the scale values are not equidistant).

Development of the ISQQ

In the present study, expert judges evaluated the quality of the implementation of questioning strategies by a sample of FAST teachers. The judges recorded a preference vote for each pair of teachers. They made holistic judgments, which are more feasible for addressing many characteristics than an analytic method such as an observation checklist. Judges of quality using the ISQQ kept in mind not only all the characteristics of good questioning but also the context within which each pair of teachers asked questions.

The development of the ISQQ included four steps: (a) development of the description of the criteria on which teachers were to be compared, (b) selection of a sample of teachers and of segments of videotapes of teachers in the classroom, (c) recruitment of judges, and (d) development of the procedures for conducting the paired-comparisons.

Development of the Description of the Criteria

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

the implementation of inquiry science, with a focus on the FAST program (Brandon & Taum, 2005). In that earlier phase of the NSF project, the researchers (the first two authors) met with the FAST developers and teachers (the remaining authors) on several occasions to whittle down a list of the features of inquiry science that were gleaned from a review of FAST student and teacher documents. After each meeting, items on the list were edited, deleted, or added.

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

Selection of a Sample of Teachers and Observation Segments to Judge

The second step was to select a sample of videotaped FAST student investigations (which had been transferred to DVD-ROMs) for the judges to examine teachers' use of questioning during inquiry science classes. The sample was selected from 135 classroom observations of 16 FAST teachers who we had videotaped on the four major Hawaiian islands. The videotaping had targeted five FAST-1 Physical Science (PS) student investigations (PS4, PS7, PS10, PS12, and PS13), which occurred at key junctures in the sequence of 14 investigations that students conducted on buoyancy and density. The key junctures had been identified in a previous study of the FAST program (Stanford Education Assessment Laboratory & Curriculum Research & Development Group, 2005). The ideal goal was to videotape all the teachers in each of the targeted FAST lessons. However, scheduling difficulties resulted in an incomplete collection of videotaped investigations, and the quality of the audio was less than desirable on some parts of some of the tapes. This limited the number of teachers we could include in the

sample. Furthermore, the number of teachers that we could sample was less than the pool of 16 because it would not have been feasible to conduct comparisons between all possible teacher pairs (N= 120).

The final criteria that we arrived at for choosing the investigations for the ISQQ were that (a) they were taught by a variety of teachers (only one teacher was represented twice in the set), (b) they represented all five of the targeted PS investigations (two instances of the teaching of each investigation were chosen), and (c) the videotapes were of sufficient quality to hear the teacher well throughout the lesson. We eliminated the lesson for the teacher who was taped twice, resulting in nine teachers who were included in the study. The population of videotaped investigations and teachers and the selected sample are shown in Table 1.

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

Judge Recruitment

The third step in the development of the project was to recruit six judges—a number deemed sufficient for reliable results and feasible within the fiscal resources of the project. In the end, five judges were available to participate. The five were male FAST experts from five states who had taught FAST and served as FAST trainers.[2] All agreed to participate in the three-day study immediately following a three-day FAST training workshop on another topic. They were

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

Table 1
Population of 100%-Usable Videotaped FAST Physical Science Investigations and the Sample (Shaded) Examined in the Quality Study [a]

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

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

compensated for airfare, local travel, and lodging and given a taxable stipend of $1,000 each.

Preparation of the Facilities, Equipment, Materials, and Procedures

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

Tryout of the ISQQ

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

Introduction and Validation of the Quality Criterion Statement

Day 1 began with an introduction to the study and the training of the judges. The conference room was equipped with a laptop computer, computer projector, and

a screen for group viewing, as well as individual laptops for the judges' use. The participants received folders describing the study, and the study administrators used the training and administration manual. A flip chart was used to record comments when appropriate. The first author served as the primary trainer; the second author, who was the project manager and a project researcher, and the third author, who was one of the FAST developers, participated in the discussion among the judges.

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

Judge Training

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

The second step of the training was to have the judges view another15-minute video segment individually on laptops with headphones. They took notes and

viewed the tapes without discussion. Then the judges again presented their opinions about the level of quality exhibited on the video. The differences among the judges' opinions about levels of quality varied little.

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

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

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

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

Making the Paired Comparison Judgments

In the early afternoon of Day 3, the judges reconvened in a University conference room for the final session. They brought the notes that they made when viewing the videotape segments and were provided with judgment recording forms. The paired-comparison method was then described in detail. It was explained that the method can be used to compare a set of "objects" on any attribute and that it produces an interval-level scale of the objects. It was contrasted with ratings, and an example of using the method was presented. The judges were told that, referring to their notes,

they would compare each teacher with each other teacher and judge (a) which member of each pair showed greater quality than the other member of the pair and (b) the similarity of their quality was on a scale of 1 to 7. (The similarity results are not reported in this paper.) Questions were fielded. Finally, the judges made the paired comparison judgments (using forms on which the teachers were randomly sorted in a different order on each form), which took from about 15 minutes to one-half hour.

Judges' Feedback About the Process

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

- Viewing the two training samples was sufficient to feel comfortable about assessing quality.
- Viewing the videotaped segments took from one to two hours per teacher.
- Entering the notes into computer files while viewing the video segments did not complicate note-taking; the judges alternated between viewing and videotaping.
- One judge stated that he found it difficult to summarize quality across the segments for the three FAST investigation phases; another found that having three segments ensured that he had a good sense of whether the teacher used good questioning strategies.
- The judges tended to apply some additional criteria such as the extent to which the teacher waited long enough for answers to questions and whether the teachers missed questioning opportunities. One judge said that he had to continue to return to the statement of quality criteria because he had additional criteria of his own in mind when viewing the videotape segments. Another offered additional statements to include in the quality statement. A third judge reported that he found it difficult to focus only on teacher questioning. For example, at first he tended to look at the children's behavior. One remedy was to listen but not to watch. Another tended to look at the students to see if they were engaged.
- 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.
- 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.

- 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.
- A judge stated that having more than nine teachers to view and assess would have been onerous.

Preparing and Scaling the Data

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

The second step was to total the preference votes and rank the totals. The preference data in each of the judges teacher-by-teacher matrixes, which were prepared in the first step, were transformed to ones and zeros, with *1* entered if the teacher in the column heading was preferred over the teacher in the row heading and *0* entered if the teacher in the row heading was preferred over the teacher in the column heading. The diagonals were assigned the value of *.5*. The cells were then totaled across the five judges, resulting in one matrix. The columns of this matrix were totaled, and the results were ranked. The ranks formed a set of scale data that we call Analysis Dataset No. 1. These are shown in Table 2.

The second set of scale data that was analyzed was formed using the Thurstone Case 5 paired comparison scaling method (Dunn-Rankin, Knezek, Wallace, & Zhang, 2004; Edwards, 1957), which produces a scale with unequal distances among the scaled objects. Using the Thurstone method, the cell totals in Analysis Dataset No. 1 were converted to proportions of the total possible number of votes (five); then the columns were summed and were sorted on the sums. The cells were then converted to normal deviates (.5 = 0), and the columns were summed. The distance between column sums was calculated; these distances resulted in scale values. For ease of interpretation, we transformed these values to a scale with a minimum value of 0.0. The scale is shown in Table 3, which we call Analysis Dataset No. 2.

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

Table 2
Results of Paired Comparisons (Analysis Dataset No. 1)

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

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

Validation

Reliability of the Preference Data

Data cannot be valid unless they are reliable. We conducted five reliability analyses of the results of the paired comparisons, each coming from a different measurement tradition. The coefficients we produced included Kendall's coefficient of concordance (*W*), Thurstone's absolute average discrepancy coefficient (Edwards, 1957; Gulliksen & Tukey, 1958), Guttman's coefficient of reproducibility (Edwards, 1957), and the intraclass correlation coefficient (ICC), Model 2 (Shrout & Fleiss, 1979). For *W* and the ICC, we used the total row of Analysis Dataset No. 1; for the Thurstone analysis we used Analysis Dataset No. 2; and for the Guttman analysis, we used Analysis Dataset No. 3. We also calculated the average percent agreement. The results (shown in Table 4) are:

1) Kendall's *W*, corrected for ties, = .55. This is a measure of the degree of agreement among judges. Howell (1992) suggested translating *W* into

Spearman's rho, because the latter is more interpretable. Our calculations shows that rho = .44, a value indicating modest reliability.

2) The Thurstone's absolute average discrepancy method produces a coefficient (.022) based on comparing empirical proportions with theoretically expected proportions; the lower the value, the better the result. The coefficient that we found (.022) is comparable to values reported by Edwards (1957) and suggests fairly high reliability.
3) Guttman's coefficient of reproducibility = .80. This value indicates the per cent accuracy with which responses to the paired comparisons can be reproduced from ranks (Edwards, 1957). Our result indicates good reproducibility.
4) The ICC, Model 2 (.48) is a measure of association among raters that takes into consideration the proportion of variance that raters have in common. According to Barrett (2001), Fleiss (1981) and Cicchetti and Sparrrow's (1981) interpretation of a coefficient of this magnitude is that it indicates a fair level of reliability.
5) The average percent agreement = 63. This was calculated as the average agreement on teacher preference between each judge paired with each other

Table 3
Thurstone Case 5 scale scores (Analysis Dataset No. 2)

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

judge. We interpret this as a fair level of agreement; a minimum of 80% would have been preferable.

These results clearly show favorable, albeit not uniformly high, levels of reliability. To examine further why the results were somewhat less consistent than desirable, we examined the differences among the judges' teacher rankings in two ways. First, we correlated (Spearman's rho) the five judges with each other on the ranks that they assigned to the nine observed teachers. The results ranged from .14 to .63, with the correlations for two of the judges with the remaining three judges clearly standing out as lower than the correlations of the three judges among each other. Second, we conducted a correspondence analysis (Benzècri, 1992; Clausen, 1998; Greenacre, 1984). This is a method for analyzing the relationship between categorical variables. The method is useful because it allows us to simultaneously plot the judges and the teachers who they judged. The results, presented here as Figure 1, show that the two judges (Nos. J1 and J2) whose Spearman's rho correlations were the lowest with the other judges were outliers in the correspondence analysis plot.

Together, these results show discrepancies of two of the judges' results with the others. The most obvious reasons for the discrepancies are insufficient attention of the two judges to the questioning quality criteria or insufficient differences among

Table 4
Results of Analyses of the Reliability of the Thurstone Case 5 Scale Values

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

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

The paired comparison method is designed to ensure that distinctions are made between closely performing objects and that ties, such as might occur with ratings, are avoided. However, the distinctions apparently were difficult for differentiating

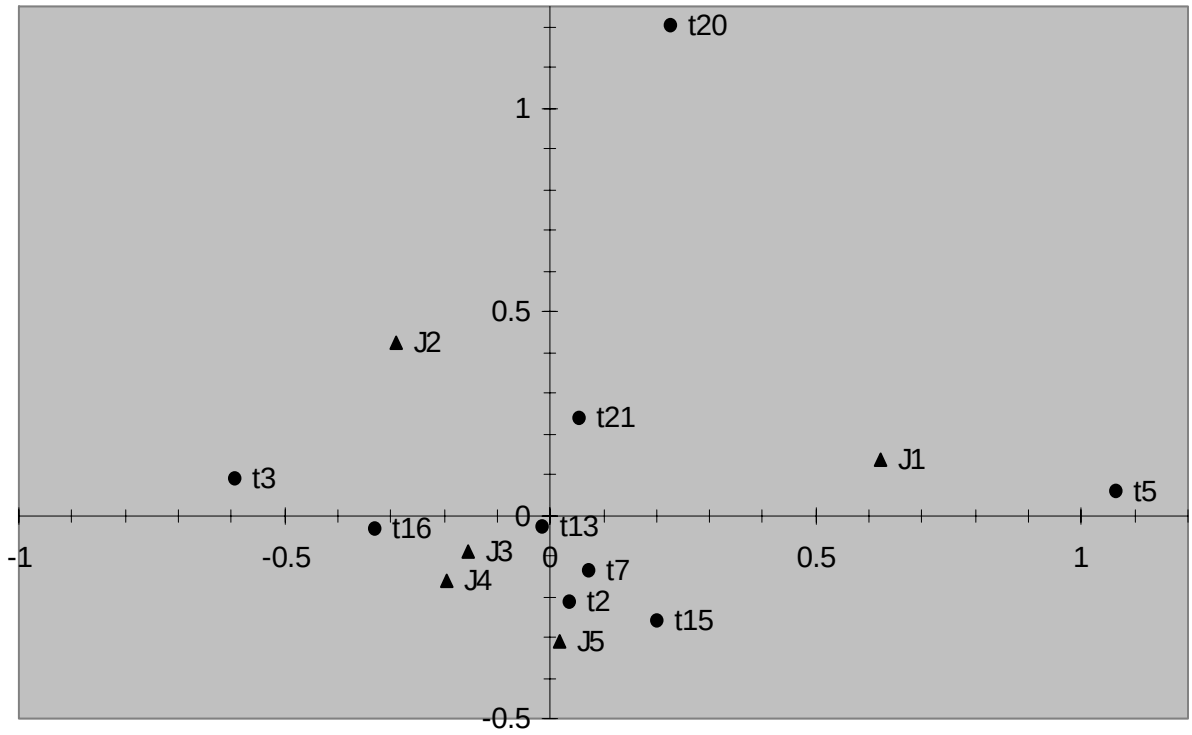

Figure 1. Results of the correspondence analysis of judges and their teacher rankings

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

Other Findings About the Validity of the Preference Data

Validity has to do with the adequacy of inferences that researchers make from the data collected with a method. We applied Messick's (1989, 1995) conceptualization of the unified theory of validity, in which construct validity has six aspects, including the content, substantive, structural, generalizability, external, and consequential aspects. Of these, we have evidence for the content and external aspects, and potentially for the substantive aspect, which we will explore at a later date, as noted below.

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

The judges' feedback at the conclusion of the workshop, however, provides somewhat mixed evidence about validity. Evidence supporting validity includes the comment about the adequacy of the number of training samples, the comment about the ease of taking notes while making judgments, and the comment about the appropriateness of providing three videotaped segments of the work of each teacher. Evidence not supporting validity is found in the comment that it was difficult to make holistic judgments about quality. Other evidence that particularly does not support content validity is found in the comments by multiple judges about their tendency to add quality criteria of their own to those specified in the statement that

the judges were instructed to use. These comments suggest that the judges' conceptualization of the task tended to be insufficiently well bounded. This might help explain the mixed findings about reliability.

External aspect of validity. The external aspect of validity has to do with the extent to which the data collected with an instrument are congruent with the data collected on the same construct with another method. In much of the validity literature, this is called concurrent validity. Concurrent-validity evidence in the form of a correlation of the ISQQ results with the results for an alternative method for assessing questioning in FAST classrooms will help assure us that the data are valid. Preferably, the alternative method should assess questioning analytically, in contrast to the holistic approach of the ISQQ.

For evidence about the external aspect of validity, we compared the results on the ISQQ with the results on an observational measure that we had previously developed and used: the Inquiry Science Observation Coding Sheet (ISOCS) (Taum & Brandon, 2005, 2006). One of the purposes of the ISOCS is to measure the extent to which teachers use questioning strategies in inquiry science classrooms. The ISOCS is a measure of the adherence aspect of implementation (e.g., Ruiz-Primo, 2005): It focuses primarily on the frequency with which teachers initiate questions and the frequency of teacher-student interactions that occur following the teachers' questions."Higher frequencies of teacher questions have been found to be related to higher levels of student achievement (as measured by either standardized achievement tests or course content mastery tests)" (Hamilton & Brady, 1991, p. 253; as support for this statement, Hamilton and Brady cite Brophy & Evertson, 1976; Coker, Lorentz, & Coker, 1980; Soar, 1973; Stallings & Kaskowitz, 1974; Weil & Murphy, 1982).

The ISOCS was developed during about two years of approximately 40 review-and-revision cycles. All of the drafts of the instrument were tried out with trained coders working with DVDs made from videotapes of about 100 lessons in 19 FAST 1 teachers' classrooms on the four major Hawaiian islands (Taum & Brandon, 2005). Eventually, the codeable 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 DVDs of the teachers who we examined with the ISQQ and reconciled the code differences between them. As evidence of the concurrent validity of the ISOCS results, we correlated ISOCS data for a sample of six teachers with

mean achievement test results for the teachers' students and found a .96 Pearson correlation.

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 student comments constituted of all codes for the teacher and (b) the percentage that the teachers' use of follow-up statements and of probing questions constituted of all codes for the teacher. The Spearman's rho correlation of the ISQQ Thurstone Case 5 scale scores with the percentage that student comments constituted of all teacher codes = .52, and the Spearman's rho correlation of the Case 5 scores with the percentage that the teachers' used follow-up statements and probing questions = .45. We believe that these correlations provide solid evidence of a substantial relationship between the two sets of results, thus supporting the validity of the data collected with the ISQQ.

Substantive aspect of validity. The substantive aspect of validity has to do with the appropriateness of the types of information that our judges recorded when considering questioning quality. Substantive validity is found to the extent that this information is appropriate to the task at hand. Evidence for this might be found in the judges notes recorded while observing the DVDs of the teachers. We plan to analyze these notes for validity purposes at a later date.

Conclusions

The paired comparison procedure is a venerable social science research method that requires people to serve as judges of each of several judged objects in light of the characteristics of each other object. The method is touted as appropriate for making distinctions among closely performing evaluands by avoiding global ratings and helping ensure differences among the results. The correlations with the ISOCS data, which measures the frequency of inquiry-science questioning, are sufficiently high to strongly provide validity evidence for the ISQQ. However, only two of the four reliability coefficients strongly support reliability, making us question somewhat whether the correlation with the ISOCS might have been higher had the judges agreed more strongly with each other.

A likely reason for the disagreement among judges was that the training period was too brief. Instead of two trial ratings, we believe in hindsight that we should have allowed time for viewing additional samples of teachers until we showed on two or three occasions that the judges viewed the teachers similarly. However, the

time necessary to do paired comparisons on multiple teachers on multiple occasions and then to discuss the differences among judges was beyond our resources (one day, with five judges).

Nevertheless, we believe that the results show that judges can assess teacher questioning quality sufficiently well. This finding is encouraging because it suggests that in the studies we hope to conduct in the future, we will be able to judge the quality of teachers' use of questioning strategies arguably the central feature of teaching inquiry science, and that together with assessments of the adherence and exposure aspects of program implementation, we will provide a full picture of the implementation of inquiry science.

References

Benzècri, J.-P. (1992). *Correspondence analysis handbook.* New York: Dekker.

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

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

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.

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.

Clausen, S.-E. (1998). *Applied correspondence analysis: An introduction.* Thousand Oaks, CA: Sage.

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

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

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

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.

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

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.

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

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

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

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* New York: Academic Press.

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

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.

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Higher Education.

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.

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

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.). Honolulu: University of Hawai'i at Mānoa, Curriculum Research & Development Group.

Pottenger, F. M. III, & 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. III, & Young, D. B. (1992c)..*The local environment: FAST 1, Foundational Approaches in Science Teaching, teacher's guide* (2nd ed.). Honolulu: 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, OH: Miami University of Ohio.

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

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.

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

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

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

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

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

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

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

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

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

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

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

Appendix A

Teacher Quality Criteria

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

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

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