An Examination of the Quality of Implementation of an Integrated Arts Education Program

Paul R. Brandon and Brian Lawton
Curriculum Research & Development Group
University of Hawai'i at Mānoa

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

The findings of a quasi-experimental evaluation of a three-year project to infuse the arts into elementary school reading and mathematics instruction showed statistically significant positive effects on the project's intended quantitative student outcomes, including achievement, attitudes toward school, and interest in the arts (Lawton & Brandon, 2006). Without data on implementation, however, it is unknown whether we can clearly attribute differences in these outcomes to the project. In this paper, we address how we collected some of our data on project implementation. The paper is one instance of our ongoing efforts to study the quality of implementation (e.g., Brandon et al., 2007).

Measuring implementation requires examining *how fully* a program is implemented and *how well* it is implemented. Measuring *how well* is more complicated than measuring *how fully* and to our knowledge is not widely discussed in the implementation literature. In this paper, we attempt to address this deficit by describing the development and use of a method for examining the quality of implementation of the arts education project. We (a) briefly provide background describing the research on the effects of arts education on elementary school students' achievement, (b) describe the project we evaluated, (c) report the development of our methods for measuring quality, (d) describe how we implemented the methods, and (e) discuss the methods' strengths and weaknesses and how the methods might have been improved. Our intent here is not to report evaluation results; instead it is to report the development and use of methods and therefore contribute to the field's understanding of how to collect some of the data that are necessary if evaluators are to know about the implementation of the educational programs that they are evaluating. In many ways, it is a "study of a study" conducted in the spirit of Cronbach et al. (1980, p. 214), who stated,

> Evaluators gain much experience in the course of designing and redesigning a study. Unfortunately, little of that experience is recorded for the benefit of the evaluation community. . . . Methods of evaluation would improve faster if evaluators more often wrote retrospective accounts of design choices.

Background

Previous Research on the Effects of Teaching the Arts on Elementary School Students' Achievement

Considerable research has been conducted on the effects on academic

achievement of (a) teaching children the arts and (b) integrating the arts into instruction (e.g., see Cornett, 2006; Darby & Catterall 1994; Deasy, 2002; Fiske, 1999). In a review of meta-analyses examining the effects of arts education on non-arts academic outcomes, Hetland and Winner (2004) provided summaries for 10 "instrumental claims" shown throughout the study of art education. For three of their claims, they found supporting evidence demonstrating a causal connection between (a) classroom drama and verbal skills, (b) music listening and spatial reasoning, and (c) music instruction and spatial reasoning. These claims were based on the results of experimental research. For five of their claims, they concluded that there was little support for causal relationships between (a) studying the arts and verbal and mathematics achievement, (b) studying the arts and creativity, (c) visual arts and reading, (d) dance and reading, and (e) music and reading. The lack of strong support for these claims was because the results were correlational or because researchers did not clearly define the arts forms that they studied. For the final two instrumental claims, Hetland and Winner found equivocal evidence to support the claim about causal relationships between dance and spatial reasoning and between music and mathematics. Overall, they concluded that greater methodological rigor is needed in future studies. Furthermore, they concluded that researchers typically do not report the quality of arts instruction, that techniques for integrating the arts often are poorly described, and that more research is needed on non-cognitive outcomes such as school attitudes and absenteeism. In particular, the lack of shared agreement among arts educators about the definition and description of arts integration strategies is a major problem in the research on arts integration (Mishook & Kornhaber, 2006). This is unfortunate, because a strong body of evidence supporting the effectiveness of these strategies is particularly important in this NCLB era, in which the use of the arts to improve basic skills is being cut back or eliminated in favor of less imaginative approaches to teaching basic skills (Eisner, 2000).

The ARTS FIRST Windward Research Project and Its Evaluation

In this paper, we discuss the evaluation of a project that trained teachers how to use the arts when teaching students reading and mathematics, with a focus on how we developed and implemented methods to study the quality of teachers' use of the arts and on our assessment of how well the methods were implemented. The evaluation examined a U.S. Department of Education Model Development and Dissemination Grant project implemented in three project schools, with three matched control schools, on the island of O'ahu in the state of Hawai'i. The project trained teachers how to integrate arts strategies into basic skills instruction, with the ultimate goal of improving students' achievement in basic academic skills, attitudes toward school, and interest in the arts. In professional development (PD) workshops, the participating teachers were trained how to use the strategies of four art forms: drama, dance, music, and the visual arts. Artist mentors worked with the teachers while they implemented the strategies in the classroom.

The core of the project was a series of daylong group PD institutes and, at each

of the participating project schools, in-class residency/mentoring sessions conducted by professional artists. In the institutes, the project team introduced participating teachers to the various strategies for integrating the arts into basic skills instruction and provided opportunities for discussion and reciprocal feedback. These were followed by in-class visits of the artist mentors to each participating teacher's classroom, in which the teachers first observed mentors modeling the strategies in art activities and then were viewed by the mentors as they "soloed" the various art forms. Each project teacher received a minimum of 20 hours of in-class PD.

The methods of the project evolved over the years it was implemented. By the second project year (School Year 2004–05), the project had developed and was implementing 13 arts activities, each of which was specific to one of the four art forms. Based on the experience of the team of developers, trainers, and mentors and on the formative evaluation findings that the authors of this paper gathered and reported during the second year, the team categorized the activities into three fundamental strategies to use across the art forms. These are shown in Table 1. Having three overarching strategies provided for more generalizability and ease of use of the activities by the participating teachers during the final project year.

Outcome evaluation. In the third year of the study, we examined the differences between the treatment and control group students on valid and reliable measures of achievement, attitudes, and interest in the arts in a propensity-score matching design. Propensity score matching helps remove the effects of selection bias in quasi-experiments (e.g., Luellen, Shadish, & Clark, 2005). Using a fixed-effects model, the analyses showed statistically significant differences between the project and control groups on students' interest in dance and in music (although not in visual arts or drama), attitudes toward school, and some of the student achievement measures (Lawton & Brandon, 2006). Other analyses showed that project teachers' attitudes toward using the arts in instruction improved significantly from the beginning to the end of the third project year.

Implementation evaluation. The outcome evaluation results suggested that the project had an effect on the students. However, because the study was a quasi-experiment, it could not be stated unequivocally that the PD accounted for the difference between groups. To the extent that it could be shown that the differences in project implementation were correlated with means on student outcome measures, stronger conclusions about project effects could be made. Therefore, we developed and tried out methods for collecting data on the level of implementation of project activities in the classroom.

Implementation has been defined in part as addressing the extent to which a program (a) adheres to its prescribed steps, (b) exposes students to the breadth of the program, and (c) is administered with quality (Dane & Schneider, 1998; Ruiz-Primo, 2005).[1] Adherence and exposure can be measured validly with self-

[1]This definition is by no means intended to describe all that has been discussed in the burgeoning literature on evaluating implementation. It simply is intended to point out the aspects of

Table 1
The Three Arts Strategies Used in the Third Year of the Arts Education Project

Strategy	Activities addressing the strategy, by art form		
	Drama	Dance/music	Visual arts
Observing	Focusing, empathizing, and using multi-sensory awareness ("muscle memory" and "emotional memory").	Focusing, listening, and using "muscle memory" and kinesthetic awareness.	Focusing on details, imagining, and visualizing.
Patterning	Sequencing, story building, and structuring using beginning, middle, and end.	Sequencing, arranging, organizing, and structuring using beginning, middle, and end.	Seeing relationships, sequencing, repeating, and arranging and organizing.
Representing	Interpreting and representing ideas through gestures and words.	Shaping, interpreting, and expressing ideas through movement and sound.	Replicating, interpreting, symbolizing, and expressing ideas though a variety of media.

report questionnaires and logs. We developed a weekly log for teachers to report exposure—that is, the extent to which they used the arts activities to help teach reading or mathematics.[2] We had limited success in consistently collecting log data, even though teachers had a choice of online or paper versions and we reminded them weekly to submit their logs.

Quality is even 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—a simple

implementation that we addressed in our study. We intend to present our study within a much broader description of the program implementation literature when the material in this paper is included in a manuscript prepared for publication.

[2]We chose not to measure adherence because we believed that the instrument would have taken more time to complete than the teachers were willing to give. To a certain extent, the case could be made that adherence was measured in our quality study. This deserves further consideration.

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 on survey questionnaires 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. Methods for collecting data on quality require considerable development time, because evaluators choosing to use the best methods must carefully identify criteria for judging quality, develop forms for conducting ratings, develop rater training procedures, videotape program sites or have raters observe sites, and conduct the quality ratings.[3]

In this paper, we describe how we measured the quality of videotaped teachers as they implemented arts activities in their classrooms after the completion of their final mentoring session. Our study of implementation turned out to be more of a trial than a full-blown study of implementation because the activities that the teachers were taught continuously evolved over the three years of the project and because the teachers did not practice and become accomplished with the activities on their own as much as anticipated. The study included four major steps: (a) videotaping teachers, (b) developing teaching-quality rating criteria, (c) selecting and training the judges, and (d) conducting the ratings. We describe each of these steps in the remainder of this paper and then discuss the "quality of the quality study," including the reliability and validity of the rating data, the strengths and weaknesses of our procedures, and how the procedures might be improved in future studies judging quality of implementation.

Methods

Videotaping the Teachers

The first step in conducting the study was to videotape teachers implementing the arts activities that they had been taught during the project. We chose to use videos for rating the teachers instead of rating them live in the classroom because videos allow for judges' multiple reviews of the teachers and thereby increase the likelihood of having nuanced ratings. The teachers were told that the videotape would be used to assess of the quality of their teaching using the arts. We reminded the teachers that the videos would not be used for the purpose of evaluating them

[3]An alternative to raters' judgments of quality is, of course, for project personnel to report their perceptions of the quality of implementation. We have used this method in the past and found relationships between the resulting self-report implementation data and student outcomes (e.g., Heck, Brandon, & Wang, 2001). The potential for a social desirability bias, which is substantial in self-reports of quality, was less in these instances than in the present instance, because, in the Heck et al. study, teachers reported the implementation of a project that was schoolwide and shared by all, whereas in the present study, teacher self-reports would have required respondents to report the quality of their individual work. This was one of the reasons of choosing in the present study to have external raters judge quality.

as teachers but rather to examine the extent to which the project had properly trained them in how to use the arts activities.

During the videotaping, the participating teachers worked with their artist mentor on an activity within an art form of their choice. Of the 12 teachers who participated in the third year of the project, five chose a drama activity, three chose a dance activity, and four chose a visual arts activity. Over the course of the semester, the mentors guided the teachers through three stages of mastery: First, each teacher observed his or her mentor conducting the activity in the teacher's classroom. Second, the teacher and mentor co-taught the activity. Finally, the mentor observed the teacher conducting the activity (the "solo" session). Each of these sessions were video-taped to allow the teachers and students to adjust to being taped. After teachers finished their solos, their final taping (the "super-solo") was scheduled. The super-solo was taped without the presence of an artist mentor. These videotapes were used for the final assessment of the teachers' quality.

Developing Quality Rating Criteria

The second step in the process was to develop criteria for rating the quality of the use of arts activities in teaching reading and mathematics. The criteria were aspects of integrating the arts that the teachers had learned and practiced in the project PD. The goal of this step was to have a list of criteria, with definitions, that judges would use for judging quality in a later step of the process.

The step began by working with a list of the aspects of quality that had been discussed in the PD. All the artist mentors had participated with the project manager and an external consultant in developing the PD over the three years of the project. This team made many changes in the methods and emphases of the PD over the course of the three years (Lawton & Brandon , 2006); by the third year, the team had arrived at a list of characteristics of good integration and use of the arts. The evaluators reviewed the list with the project manager and confirmed that she continued to consider them essential to good arts integration. The project manager added another list of criteria that she had developed over the years. The evaluation team combined the two lists into one, resulting in a document showing 29 criteria in three categories: content, planning and organization, and delivery.

Next, the combined list, which obviously was too extensive to use for rating quality in a reasonable period of time, was reviewed in a meeting of the project participants. The evaluators convened a six-hour meeting of the project manager, three of the artist mentors, a university arts-education teacher trainer, a laboratory high school fiber artist/arts educator (who was affiliated with the evaluators' organization and had served as an observer during the three years of the project), a university lecturer with previous experience as a K–12 classroom teacher, and the authors of this paper. Of the artist mentors, one was a drama expert and two were visual arts experts. The project manager was a dance expert, and the teacher trainer was a drama expert. The goals of the meeting were to prepare definitions of each criterion, discuss instances of when the criteria are seen in the classroom, and identify the most essential criteria. Each criterion in the final list had to apply to all

the art forms. They had to be clear enough to judge quality reliably, and the list had to be short enough to make the judgments manageable.

The participants engaged in an extensive discussion about each criterion. The first author of this paper served as the meeting facilitator. The second author, who was the evaluation project manager, the hands-on formative project evaluator, and a classroom observer and videographer for the length of the project, participated extensively, contributing his knowledge of the criteria and how they were manifested in the classroom. As the meeting proceeded, some criteria and examples were revised, some were deleted, and some were combined. The meeting resulted in a list of 10 criteria, with examples of each

After the meeting, the evaluators reviewed the list and identified redundancies and unclear portions of the criteria. With the concurrence of the project manager, they eventually narrowed the list of criteria to eight. The criteria are shown as Appendix A.

Selecting and Training the Judges

The third step in the process was selecting and training judges. The arts education project manager, two of the project's five artist mentors (a dance expert and a visual arts expert), the laboratory school fiber artist/arts educator, and the university arts-education teacher trainer participated as judges. (The other three artist mentors were unavailable.) The university lecturer with previous experience as a K–12 classroom teacher and the second author, both of whom had helped develop the criteria, sat in on the training for the purpose of trying out the role of judge, engaging in the discussion, and providing feedback about how the rating process might be improved.

A judge training workshop was prepared and conducted. The plans called for (a) reviewing the quality criteria, (b) conducting one round of group practice and discussion of the application of the criteria, and (c) conducting as many as two more individual rounds of practice, each to be followed by discussion. The evaluators selected three of the participating project teachers' solo videotapes. The videotapes were transferred to DVD-ROMs for training the judges. The evaluators then prepared a rating form that instructed the judges to rate the teachers on the videotapes on a 1–4 scale (with half-points allowed) for each teaching quality criterion. It was decided not to label anchor points with extended rubric descriptors; instead, *1* = no acceptable quality, *2* = acceptable quality, *3* = good quality, and *4* = excellent quality. The anchor descriptors were weighted toward good quality rather than inadequate quality to allow more variation in the ratings of the observed teachers, each of whom had practiced using the activities during the final year of the project and were thought to have mastered the activities fairly well. The first page of the rating form, giving the instructions and the section for rating the first of the criteria, is shown as Appendix B.

The workshop was convened for a period of about six hours. The first author served as the facilitator. After introducing the workshop purpose and agenda, the evaluators reviewed the criteria and explained that they had been slightly revised

Table 2
Rating of the first training DVD-ROM[a]

Criterion	Judge						
	1	2	3	4	5	6	7
1	3	3	3.5	3	4	4	4
2	2.5	3	2.5	3	3	3	3.5
3	2.5	3	2.5	2.5	3	3	4
4	3	3.5	2.5	2	2	3	3.5
5	3	3	2	2	2.5	3	4
6	2.5	4	3	2.5	2.5	3	4
7	2.5	2.5	2.5	3	3	4	3.5
Overall	2.5	3	2.5	2.5	2.5	3	3.5

[a]Judges 6 and 7, the evaluation project manager, who served as the formative evaluator, and the university lecturer with experience as a classroom teacher, were not trained artist mentors.

since their initial development. The group then began viewing a sample DVD-ROM on a computer projector and immediately engaged in an extensive discussion about the meaning of the criteria, when the criteria were applicable, and the level of quality being exhibited on the video. The discussion about the video continued for almost two hours.

At the conclusion of viewing the sample classroom video, each judge privately rated the teacher on each of the seven criteria and also did an overall assessment. The facilitator then recorded the results on a flip chart, and the group reviewed and discussed the results, shown here as Table 2. As seen in the table, all the judges except Nos. 6 and 7, who were the evaluation project manager and the university lecturer with previous classroom teaching experience, were within one-half point of each other. The participants agreed that the criteria on which their ratings diverged the most tended to be the more difficult to apply. They also decided to add the overall assessment as the eighth criterion

The next step was to have been an individual viewing of a second videotape without discussion or interruption. However, the evaluators and the group agreed that more discussion might be in order as the viewing proceeded. The evaluators showed a second DVD-ROM, and the judges viewed it and took notes. Compared to the first viewing, little discussion ensued until after the viewing was completed. When asked for their independent ratings of the second video, the results shown in Table 3 were collected. (The university lecturer and the second author did not

Tab le 3
Rating of the second training DVD-ROM

Criterion	Judge				
	1	2	3	4	5
1	1.5	1.5	1.5	1	1.5
2	1.5	1	1.5	1	1
3	2	1.5	1	2	1
4	1.5	1.5	1.5	1.5	1.5
5	1	1	1	1	1
6	—	—	—	—	—
7	1.5	1.5	1	1	1
8	1	1	1	1	1

participate in this round of ratings.) As seen in Table 3, the ratings were more similar this round than the previous. Although the increased similarity in the second round probably was caused in part by the low quality of the implementation of the arts activity by the teacher shown in the viewed DVD (resulting in a floor effect), the results were deemed sufficiently close to end the training.[4]

At the end of the meeting, the five judges were given a DVD-ROM for each of the 12 third-year project teachers whose arts-integration quality was to be rated. They were asked to complete their ratings at the places of their choosing within four weeks. They were instructed not to discuss any of their work with each other. Finally, they were asked to take extensive notes about the extent to which the teachers addressed the quality criteria and provide written statements justifying their final ratings on each criterion. (We discuss the ratings but not the written comments in the next section of this paper.)

After the ratings were completed, we conducted a post-rating focus group, in which we asked the judges for their feedback about the process, which we draw upon later in the paper.

[4]The final step of the process was for the artist mentors attending the training to weight the criteria. As experts in the project, they were an appropriate group to decide the weights of each criterion. Each of the four artist mentors assigned a weight to each criterion. However, the weights were not used in the analysis because the unweighted and weighted results did not vary considerably.

Table 4
Mean Differences in Ratings (Across All Rated Teachers) Within Pairs of Judges for Each of the Eight Quality Criteria

Criterion	J1-J2	J1-J3	J1-J4	J2-J3	J2-J4	J3-J4	All
1	.92	.71	.42	.63	.67	.79	.69
2	.50	.58	.63	.67	.38	.71	.58
3	.75	.79	.58	.79	.58	1.04	.76
4	.33	.67	.63	.50	.71	1.13	.66
5	.88	1.08	.50	.88	.79	1.08	.87
6	.71	.96	.54	.67	.58	1.00	.74
7	.63	.71	.67	.67	.79	1.21	.78
8	.67	.58	.33	.50	.58	.75	.57
Mean	.67	.76	.54	.66	.64	.96	.70

To prepare the data for analysis, we entered the ratings of each judge of each teacher on each criterion into a computer file and verified the results. These raw data are shown in Appendix C, Table C1.

The Quality of our Study of Quality

Interrater Reliability

The primary purposes of this paper are to report and reflect on our methods and their implications for the quality of the quality judgments. Essentially, this is about the validity of the ratings and of the inferences that we can make from them. Our first step was to examine interrater reliability. Reliability is a key aspect of validity; before we know whether we can proceed to use the quality ratings to report the level of implementation, we need to know about the consensus and consistency of the ratings (Stemler, 2004).

Results are reported here for only four of the judges. Early in the course of our analyses, we discovered that the ratings of the fiber artist/arts educator, who we thought might be an appropriate person to be a judge because she had served as an observer during the three years of the project, correlated very poorly with the other judges' ratings. Therefore, we eliminated her results from further analyses.

Consensus estimates. Consensus is about the extent to which judges agree on the ratings. It reflects the extent to which judges' interpretation of the level of quality is the same. Our measure of consensus was the average difference within each of the pairs of judges ($n(n\text{-}1)/2$ = six pairs). For each criterion, we calculated the absolute value of the difference between the two judges' ratings (in a pair) for each teacher and averaged these differences, resulting in an average difference

Table 5
Correlations Among Judges on Each of the Eight Quality Criteria

Criterion	J1 & J2	J1 & J3	J1 & J4	J2 & J3	J2 & J4	J3 & J4	All
1	.70	.50	.87	.58	.58	.24	.58
2	.45	.44	.54	-.05	.71	.14	.67
3	.54	.28	.58	.31	.68	-.04	.54
4	.76	.66	.50	.75	.58	.30	.63
5	.37	.43	.42	.07	.43	.06	.49
6	.59	.25	.67	.53	.81	.30	.53
7	.44	.52	.55	.38	.75	.20	.57
8	.46	.49	.73	.46	.65	.21	.69
Mean	.54	.45	.61	.38	.65	.18	.59

value for each pair on each criterion. We also calculated the mean difference across criteria for each judge pair. The results, given in Table 4, show that the mean differences ranged among the criteria from .04 to 1.13. The smallest mean difference across criteria was between Judges 1 and 4 (.54). This value was about 20% less than the next lowest average (.64 for Judges 2 and 4). Thus, Judges 1 and 4 agreed with each other more than the other judges. The average difference for this pair of judges was about ½ point, which was the largest difference between judges that we hoped to find in the study. In contrast, the average across all judge pairs was .70.

Consistency estimates. Consistency estimates address the extent to which the judges in our study consistently applied the criteria when judging the quality of teachers' implementation of arts activities in their reading or mathematics instruction. To examine consistency, we (a) calculated correlations among the four judges' criteria ratings for each of the 12 teachers and (b) calculated coefficient alphas, resulting in 12 correlation matrixes, with an alpha coefficient for each. The purposes of this step were to examine the correlations among judges and to determine the extent to which the criteria ratings could be analyzed as one scale.

The Pearson correlations among the judges are shown in Table 5. They partially confirm the findings of the percent agreements and the average differences: The correlations of Judges 1 and 4 (mean = .61) are not the highest, but they are nearly the highest (at .65, the mean correlation between Judges 2 and 4 is slightly greater).

Of the alpha coefficients (one calculated for the results for each of the 12 teachers), five were greater than .90, three were from .80 to .89, two were from .70

Table 6
Four Judges' Overall Quality Scores (Mean Criterion Ratings) of 12 Teachers' Quality

Teacher	J1	J2	J3	J4	Mean
1	3.00	3.63	3.25	3.00	3.22
2	3.13	2.25	3.69	2.00	2.77
3	1.31	2.56	2.94	1.63	2.11
4	2.94	3.81	2.94	3.38	3.27
5	2.06	2.69	3.06	2.44	2.56
6	2.81	2.81	2.44	2.94	2.75
7	1.38	2.56	1.81	1.81	1.89
8	2.13	2.06	1.88	1.88	1.98
9	1.5	1.50	1.94	1.75	1.67
10	2.75	3.19	3.44	2.38	2.94
11	1.94	2.44	3.44	1.75	2.39
12	2.69	2.94	2.94	2.00	2.64
Mean	2.30	2.70	2.81	2.25	2.52

to .79, and two were below .70. We concluded that we were justified in considering the criteria as one scale.

Having determined that we could analyze the results on the eight criteria as one scale, we calculated the mean rating across the eight criteria for each judge, resulting in four means for each teacher (one for each judge). Then we averaged the mean ratings across judges, resulting in an *Overall Quality Score* for each of the 12 teachers. They are seen in Table 6. As shown in this table, the quality scores ranged across judges from somewhat below an acceptable level of quality (1.67) to somewhat above a good level of quality (3.27), with a mean across teachers of 2.52 (st. dev. = .52), indicating an overall project level of quality midway between acceptable and good. The average ratings of Judges 1 and 4 (2.30 and 2.25, respectively), who showed the greatest consensus and consistency, were considerably less than the average ratings for the other two judges (2.70 and 2.81).

Kendall's coefficient of concordance (W) is another measure of consistency. For the four judges' Overall Quality Scores for the 12 teachers, $W = .60$. Howell (1992) suggested translating W into Spearman's rho, because the latter is more interpretable. We found that Spearman's rho = .49, a moderate correlation. For an

Table 7
Kendall's Coefficient of Concordance (*W*) and Interclass Correlation Coefficients on the Overall Quality Scores, for Pairs of Judges

Statistic	J1 & J2	J1 & J3	J1 & J4	J2 & J3	J2 & J4	J3 & J4	All
Kendall's *W*	.74	.71	.89	.62	.88	.60	.60
W translated to Spearman's rho	.48	.42	.78	.24	.76	.20	.49
Intraclass correlation coefficient	.48	.33	.67	.47	.58	.03	.45

Overall Quality Score consisting of the average for Judges 1 and 4 only, *W* = .89, and Spearman's rho = .78—a considerable improvement over the results for all four judges combined. The Judge 2/4 combination showed the second best results.

The final consistency analysis that we conducted was to calculate the intraclass correlation coefficient (ICC), Model 2 (Shrout & Fleiss, 1979). The ICC is a measure of association among judges that takes into consideration the proportion of variance that judges have in common. According to Barrett (2001), Fleiss (1981) and Cicchetti and Sparrrow (1981) interpret the ICC that we found across all judges, .45, as indicating a "fair" level of reliability. For the pair of Judges 1 and 4, the ICC = .67, showing a fairly high proportion of variance that the two judges have in common and thus indicating a good level of agreement. Again, the Judge 2/4 combination showed the second best results.

Content-Related Validity

The content aspect of validity addresses "content relevance, representativeness, and technical quality" (Messick, 1995, p. 745). In addition to the reliability results, which show technical quality, evidence for the content aspect of validity is found in our description, earlier in the paper, of (a) our methods for developing the quality criteria, (b) the procedures for developing and implementing judge training, and (c) the manner in which the ratings were conducted. We believe that this evidence shows content validity, although the judges' feedback at the conclusion of the workshop qualified the strength of the evidence somewhat. The judges tended to find that Criteria 4 and 5 overlapped, suggesting either that the two criteria should have been combined or that the training should have been more explicit. They also tended to agree that the training should have been longer. For example, it was suggested that the judges rate one or two videos on their own immediately after the initial training and then, on a day soon after, reconvene to discuss the results.

Criterion-Related Validity

As a measure of validity, we correlated the Overall Quality Scores with results from data collected with our student and teacher outcome measures for the final year of the project. These included student attitudes towards school, student interest in the arts, and teacher attitudes toward teaching with the arts. We did not correlate the results with student achievement because we interpreted the between-group differences in achievement scores to be due in large part to reading programs that had been used at the project schools. We did not find any statistically significant results with either the results for the entire set of four quality judges or with the pair of Judges 1 and 4. The highest correlation that we found was .32, which was between the average rating for Judges 1 and 4 and teachers attitudes after partialling out teachers' attitudes at the beginning of the year. We believe that this provides some evidence of the validity of the Overall Quality Scores.

Overall Summary and Conclusions

Results of the validity and reliability analyses. The primary results of the validity and reliability analyses that we have reported here are as follows:

1) We found early in the analyses that the ratings of the fiber artist, who had observed and recorded all the PD sessions, did not correlate highly with the results for the other judges. This might have been due to her lack of experience integrating the arts into elementary school reading and mathematics. Therefore, we removed her results from further analyses, which was disappointing because we had endeavored to have as large a pool of judges as possible.
2) The average difference among the judges' ratings was .70, which was greater than desirable. Among one pair of judges (1 and 4), however, the results were on target. Furthermore, the correlation between this pair of judges was the second greatest of all pair of judges and only slightly lower than the correlation for another pair (judges 2 and 4).
3) Kendall's *W* and the ICC showed disappointing results for the group of four judges analyzed together. However, for the Judge 1/Judge4 pair and the Judge 2/Judge 4, pair, Kendall's *W* and the ICC were quite satisfactory. The three judges in these two pairs were the project manager and two of the participating artist mentors; one of the other two judges was the teacher trainer who had experience in integrating the arts but who did not participate in the project.
4) The correlation of the quality ratings with the results on the student outcome measures suggested virtually no relationship between implementation and outcomes in our study. There was no correlation between teacher attitudes toward teaching with the arts and the results for all four judges, but there was a small correlation for Judges 1 and 4.

Some tentative conclusions about the study. The reliability and validity findings lead us to several tentative conclusions:

1) The disappointing consensus results for the group of four judges suggests that the criteria were not sufficiently well defined or that there might have been insufficient consensus about levels of quality (i.e., acceptable, good, and so forth).
2) Although it is desirable to have several judges, it is possible to identify

subgroups of judges whose results are acceptable when the results for the entire group are not. The possibility that not all judges' results will prove psychometrically adequate underscores the need to enlist a pool of several judges.

3) Only experts in the project should participate as judges. We were not entirely surprised that two pair of judges proved superior to the others, because they consisted of full participants in the project, whereas one of the other judges was not a project participant. The fourth judge was a participant, however, underscoring the need to have a number of judges participate so that an acceptable subgroup can be identified.
4) Training in how to rate the quality of the arts education project probably should be of an open-ended length, ending after consistency and consensus are established in multiple instances. Issues about the feasibility of having longer training would have to be addressed if training were lengthened.
5) When conducting practice sessions during the training, judges probably should be asked to privately record and report their ratings. We chose to have judges report their ratings to the group aloud; some might have changed their ratings to accommodate their peers' opinions. If the judges had reported their ratings privately, it is likely that problems with lack of consensus could have been identified and addressed during the training.
6) Training might be improved if videotaped segments showing levels of implementation on each of the criteria were used to calibrate the teachers.
7) Judges probably need to be instructed explicitly that it is their obligation to come as close to consensus as possible during the training. Using only personnel who participate in a project probably would facilitate this process, because they are more likely than outsiders to have established group cohesion.
8) Evaluators should not necessarily expect relationships between the quality of implementation and student outcomes. In our case, there are several possible rationales for this conclusion. First, our student attitude and interest measures tended to show a ceiling effect. Perhaps a lack of variation in student outcomes made it difficult for other variables to show correlations with the outcome data. Second, the data we collected from our teacher logs suggested that the teachers used the arts strategies infrequently. The effect of quality of implementation is likely to be minimal if the frequency of implementation is minimal. Third, the relationship between quality and implementation might not be strong enough to show a relationship with outcomes under the best of circumstances. To what extent can we expect variations in levels of quality among high-quality teachers to show a statistical relationship with variation among other data?
9) The judges found that the process of the study was a useful formative evaluation task. The development and use of the criteria occurred at a time in the evolution of the theory and methods of the project when the integration model and its manifestation in the PD and in the classroom had substantially jelled. The systematic process of conducting the ratings provided the project manager and

the artist mentors with deeper insights into the quality of the implementation of the arts activities than they had had previously, despite having worked closely with the teachers in the institutes and in the classroom. The project personnel had trained the teachers with criteria in mind, but they had not systematically identified, defined, and described a short and explicit list of quality criteria before the PD and trained the teachers in the criteria. The project manager informed us that they would explicitly address the criteria in future training.

Future studies. In addition to considering these conclusions, future studies of how to measure quality might also take these additional steps:

1) The differences among criteria might be more closely examined to determine whether some criteria need revision. For example, the standard deviations (which we calculated but did not report here) of criterion ratings might be examined for consistency among judges.
2) Discriminant validity results might be obtained and examined. For example, the artist mentors' overall assessments of the quality of the teachers might be collected and compared with the quality ratings. (Of course, if the mentors' assessments and the quality ratings show a strong correlation, the conclusion might be to use the mentors' assessments to judge quality because they will be much cheaper to obtain!)
3) Teachers' general teaching quality needs to be assessed at the beginning of studies so that it can be partialled out of their quality ratings.
4) The appropriateness and adequacy of assigning teachers a simple three-level categorical implementation score (low, medium, and high) should be explored. Perhaps implementation scores are not sufficiently accurate and sensitive to serve as interval-level predictors of achievement, or perhaps teachers need only reach a minimal level of proficiency to be effective.
5) Generalizability coefficients should be calculated as another measure of reliability.
6) Multi-level analyses, which can tease out the relationship between teacher-level results and student-level results, might be conducted of the relationship between quality and student outcomes.
7)

References

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

Brandon, P. R., Taum, A. K. H. Young, D. B., Pottenger F. P., Speitel, T. & Gray, M. (2007, April.) *Development, validation, and trial of a method for judging the quality of using questioning strategies in a middle-school inquiry science program.* Paper presented at the meeting of the American Educational Research Association, Chicago, 2007.

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

Cornett, C. E. (2006). *Creating meaning through literature and the arts: An integration resource for classroom teachers* (3rd. ed.). Upper Saddle river, NJ: Pearson Education.

Cronbach, L. J., & Ambron, S. R., Donrbusch, S. M., Hess, R. D, Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation.* San Francisco: Jossey-Bass.

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.

Darby, J. T., & Catterall, J. S. (1994).The fourth R: The arts and learning. *Teachers College Record*, *96*, 299–328.

Deasy, R.J. (Ed.). (2002). *Critical links: Learning in the arts and student academic and social development*. Washington, DC: Arts Education Partnership.

Eisner, E. W. (2000). Arts education policy? *Arts Education Policy Review*, *101*(3), 4–6.

Fiske, E. B. (1999). *Champions of change: The impact of the arts on learning.* Washington, DC: Arts Educational Partnership.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd ed.). New York: Wiley.

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

Hetland, L., & Winner, E. (2004). Cognitive transfer from arts education to non-arts Outcomes: Research evidence and policy implications. In E. Eisner & M. Day (Eds.), *Handbook on Research and Policy in Art Education*. National Art Education Association.

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

Lawton, B., & Brandon, P. R. (2006, November). *Methods and results of a three-year evaluation to infuse arts strategies in elementary reading and mathematics instruction.* Paper accepted for presentation at annual meeting of the American Evaluation Association, Portland, OR.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*, 530–558.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741–749.

Mishook, J., & Kornhaber, M.L. (2006). Arts integration in an era of accountability. *Arts Education Policy Review*, 107(4), 3–11.

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

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

Stemler, S. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved March 19, 2007 from http://PAREonline.net/getvn.asp?v=9&n=4

Appendix A
Quality Criteria Statements and Examples

ARTS FIRST Windward Research Project
Quality Criteria Statements and Examples

No.	Criterion	*The teacher. . .*
1	Provision for an appropriate **physical environment**	. . .prepares the classroom layout and supplies, materials, and equipment in a manner that facilitates the lesson.
2	Maintenance of the **students' focus** on the activity	. . .divides his or her attention among students appropriately. . . .gets students on task quickly and keeps them on task. . . .structures effective grouping and collaboration. . . .manages the energy level of the students.
3	Establishment and maintenance of a clear **progression** of the lesson	. . .asks questions and provides articulate instructions. . . .presents the appropriate sequencing. . . .provides smooth transitions between activities. . . .manages the pacing of the activities.
4	Facilitation of students' **reflection**	. . . provides opportunities for the students to (a) *describe* ("tell what you see"), *interpret* ("tell what you imagine is happening"), and *evaluate* ("tell what you like most and why"). . . .provides timely feedback. . . .asks questions to assess the students' understanding. . . .suggests possible improvements.
5	Fostering of students' **creative expression**	. . .facilitates brainstorming. . . .asks open-ended questions. . . .accepts students' choices. . . .encourages risk-taking and experimentation. . . .projects an open, relaxed attitude.
6	Allocation of the necessary **time** to the three artistic processes	. . .allocates the appropriate time to *creating* (students generate art; this includes problem solving, decision-making and exploration); *performing* (students share their art); and *responding* (students reflect on products and process).
7	**Use of the arts strategies to teach subject matter**	. . .facilitates understanding in various stages: (a) introducing the strategy; (b) reviewing understanding; (c) adding a technique. . . .makes a clear connection between the arts strategies and the arts benchmark.
8	**Global assessment** (rater's summarization of overall teacher quality)	

Appendix B
Teacher Quality Criteria Rating Form

ARTS FIRST Windward Research Project
Teacher Quality Criteria Rating Form
(Page 1 only)

Instructions: The purpose of this form is to rate each of the 12 Windward Research Project teachers *no later than* Wednesday, February 21, 2007. Rate the teachers on each criterion with 1 = *not acceptable quality*, 2 = *acceptable quality*, 3 = *good quality*, and 4 = *excellent quality*. Enter your rating into the appropriate column. Do your best to use only these numbers; however, you may give a rating of 1.5 if you think that the quality shown on the criterion was above *not acceptable* but below *acceptable*, a rating of 2.5 if you think that the quality shown on the criterion is above *acceptable* but below *good*, or a rating of 3.5 if you think the quality was above *good* but below *excellent*.

It is essential that you record your evidence of particular behaviors and events, including time stamps, in the Notes column that influenced your rating. In addition, please provide a final statement that justifies/summarizes your final rating. If you need more room than is provided please attach a separate sheet of paper indicating criteria, teacher name, and your name.

Your name: _____ Teacher name: _____

Criterion	Rating (1-4)	Notes
1. Provision for an appropriate **physical environment**		

Appendix C
Ratings of Teacher Quality

Table C1

Rating of 12 Teachers' Quality on Eight Criteria by Four Judges

Teacher	Criterion 1				Criterion 2				Criterion 3				Criterion 4				Criterion 5				Criterion 6				Criterion 7				Criterion 8			
	J1	J2	J3	J4	J1	J2	J3	J4	J1	J2	J3	J4	J1	J2	J3	J4	J1	J2	J3	J4	J1	J2	J3	J4	J1	J2	J3	J4	J1	J2	J3	J4
1	3	4	4	3.5	3.5	4	2.5	4	3	3.5	2.5	3	3	3.5	4	3	2.5	3	3	2.5	3	4	4	2.5	3	3.5	3	2.5	3	3.5	3	3
2	3	4	4	3	3	2	4	3	3	2	4	2	3	2	3.5	1	3	2	3.5	1.5	3	2	3.5	2	3.5	2	3.5	1.5	3.5	2	3.5	2
3	2	4	2.5	2	1.5	3	3	3	1	2	2.5	2	2	2.5	3.5	1	1	2	2.5	1	1	2	4	1	1	2.5	2.5	1.5	1	2.5	3	1.5
4	3	4	2.5	4	3	3.5	2	4	3	4	2.5	4	3	3.5	4	2.5	2.5	4	3	2.5	3	4	3.5	4	3	3.5	3	3	3	4	3	3
5	2	3.5	3.5	2	2	3	2.5	2.5	2	2.5	3	2	2.5	2.5	2.5	2.5	2	2	3.5	3	2	3	3	3	2	2.5	3.5	2	2	2.5	3	2.5
6	3	3.5	2.5	4	3	3	3	4	3	3.5	2.5	4	2	2	2	1.5	2.5	2.5	1.5	2	3	3	3	2	3	2.5	2.5	3	3	2.5	2.5	3
7	1.5	3	2	2.5	1.5	2.5	1.5	2.5	1.5	2.5	2	1.5	1.5	2	2	1	1	3	1	2.5	1.5	2.5	2	1.5	1	2.5	2	1.5	1.5	2.5	2	1.5
8	2	3	2	2.5	2	2	2	2.5	3	2	1.5	2	2	1.5	1.5	1.5	1	2	2.5	1.5	3	2	1.5	2	2	2	2	1	2	2	2	2
9	1.5	1.5	2	2	2	2	2.5	2.5	1.5	1	1.5	1.5	1.5	1.5	2	2	1.5	2	2.5	2.5	1	1.5	1	1	1.5	1	2	1	1.5	1.5	2	1.5
10	2	2.5	2.5	2	3	3	3.5	3	2.5	3.5	3	2	3	3	3	2	3	3.5	4	3	2.5	3	4	2	3	3.5	4	2.5	3	3.5	3.5	2.5
11	2	3	3	2.5	3	3	3.5	3	2	3	3.5	1.5	1.5	1.5	3	1.5	1.5	2.5	4	1.5	1.5	2	3	1.5	2	2	4	1	2	2.5	3.5	1.5
12	4	4	3	4	3	2.5	2.5	2	3	3	3.5	2	2	2.5	3	1	1.5	3	3	1.5	3	3	2.5	2.5	2.5	2.5	3	1	2.5	3	3	2