INSTITUTE FOR LEARNING INNOVATION

Thinking Through Art
Isabella Stewart Gardner Museum
School Partnership Program
Year 3 Research Results



April 2007

Prepared by: *

Marianna Adams
Susan Foutz
Jessica Luke
Jill Stein

***Authors are listed alphabetically**

Thinking Through Art Isabella Stewart Gardner Museum School Partnership Program
Summary of Final Research Results
April 2007

Museums and schools have a long history of working together to facilitate students' learning in and through the arts. While art museums have traditionally served school audiences through a range of single-visit tours, increasingly they offer more extensive school programs in an effort to provide students with in-depth, comprehensive learning experiences. Studies suggest that as many as half of American museums offer some form of a multiple-visit school program in which students might visit the museum from two to ten times a year (Wetterlund & Sayre, 2003). Museums also offer extended experiences such as pre- and post-visit activities in the classroom (Adams, Luke & Manuel, 2003; IMLS, 2002)

Recent research suggests that many multiple-visit programs focus on creative and critical thinking skills, skills that are considered increasingly important in the general education of young people (Adams, Luke & Manuel, 2003). Yet, until now, the museum education field has neither articulated exactly what is meant by critical thinking skills,[1] nor how the museum provides a unique environment for learning such skills.

In 2003, the Isabella Stewart Gardner Museum (ISGM), in partnership with the Institute for Learning Innovation (ILI), received a 3-year grant from the Department of Education to research students' learning in and from an art museum multiple-visit program. The ISGM's *School Partnership Program (SPP)* provided the context for this study and focused on three overarching goals described below in more detail. Launched in 1996, the *SPP* is a multiple-visit program serving K-8 students from neighboring inner-city public schools. Over the three years of the study, the pedagogy for the *SPP* shifted from a Socratic-method to more open-ended questions, using the Visual Thinking Strategies (VTS) questioning model which focuses on learning to look at and make meaning from works of art, as well as gaining familiarity with the museum environment in order to feel comfortable using the Gardner as a community resource.

Goal 1: Develop reliable and valid measures for assessing Critical Thinking Skills (CTS) in and from an art museum that are applicable to a wide range of school/museum educators across the country.

The Year 3 study was designed as a one-year, quasi-experimental investigation, using a posttest-only control-group design to assess the effects of participation in an art museum multiple-visit program on elementary students' CTS. Multiple methods were used to assess critical thinking, including individual student interviews with a poster reproduction, group "untours" in the museum where children were allowed to view a gallery on their own, and performance on state-wide standardized tests. Students from grades 3, 4, and 5 in two schools

[1] It is important to note that comprehensive research has been conducted in critical thinking and school-based arts education. The work of Abigail Housen most notably influenced this study. See specific citations in Reference section.

participating in the *SPP,* Farragut and Tobin, comprised the treatment group. Students in grades 3, 4, and 5 from three non-participating schools, Hamilton, Kennedy, and Philbrick, comprised the control group. Control schools were matched to treatment schools based on test scores, race/ethnicity distribution, socio-economic status of the school population as measured by percent of students eligible for free and reduced lunch, and the absence of regular visual arts instruction by a certified art teacher.

A total of 135 students (64 treatment group; 71 control group) participated in individual poster interviews that assessed their ability to employ critical thinking skills while looking at a poster reproduction of a work of art. A total of 116 students (56 treatment group; 60 control group) participated in the "untours" that assessed students' ability to apply their knowledge and skills not just to one individual work of art, but to the larger museum context and the multiple objects within this context. Finally, standardized test scores were collected from 410 students (199 treatment group; 211 control group) and analyzed to determine the degree to which *SPP* participation influenced students' critical thinking abilities in the broader context of general reading and comprehension on the Massachusetts Comprehensive Assessment System (MCAS) test and the Stanford-9 (SAT-9).

Years 1 and 2 of the study focused on the development of a CTS rubric based on exploratory data from students and teachers in the *SPP*, as well as critical thinking frameworks from Costa and Kallick (2000) and Housen (1983). Seven individual critical thinking skills comprised the final Year 3 version of the CTS rubric, Observation, Interpretation, Evaluation, Association, Problem-Finding, Comparison, and Flexible Thinking. Further, three of the CTS categories, Observation, Interpretation, and Evaluation, were divided into sub-categories, in an effort to breakdown the larger CTS category into more discrete and nuanced parts. A separate, over-arching critical thinking skill, Evidence, was applied to any of the seven individual critical thinking skills when used to support an assertion or opinion.

This rubric proved itself to be a robust, valid, and reliable assessment tool for critical thinking. However, it is quite cumbersome and requires substantial training to administer properly. To make the instrument more user-friendly for museum practitioners, we developed a short-form without the sub-categories and pilot-tested it with thirteen museum practitioners in six museums across the country. The results of this pilot test were encouraging and provide insight into the applicability of the condensed rubric. Practitioners saw a need to measure critical thinking within their programs, and perceived the short-form rubric as a relevant and useful way to do that despite some areas that need revision to be more relevant and easier to use. Given the limited context provided to these practitioners, and the fact that they were not familiar with this data collection method, it is impressive that they could make sense of the rubric and offer such valuable insights. It was exciting to see that the exercise successfully stimulated practitioners to think differently about their practice. Clearly the rubric can be a valuable diagnostic and training tool for practitioners as they strengthen the ways that their programs can enhance young people's critical thinking skills.

Goal 2: Better understand how an art museum multiple-visit program can facilitate critical thinking amongst a generalizable sample of inner-city elementary students.

The study design described above sought to address the following three research questions:
Research Question 1: How does participation in the *SPP* influence students' individual critical thinking skills? When asked to "think out loud" about the poster reproduction, treatment

students were consistently able to say more than control students about the artwork, demonstrating higher frequencies of critical thinking skills and evidence for their assertions. Data showed that treatment students talked significantly longer about the art work, averaging 28 lines per interview compared to 14 lines for control students.[2] More importantly, treatment students generated significantly more instances of the critical thinking skills of Observation, Interpretation, Association, Comparison, Flexible Thinking, and Evidence.

Research Question 2: How does participation influence students' critical thinking skills within a social context in the museum? When given time to explore and discuss works of art in the ISGM, treatment students again had more to say than did control students. Treatment school students averaged 45 lines per small group "untour," while control students averaged 27 lines. Similar to the findings in the individual poster interviews, treatment students in the "untour" surpassed control students on the CTS categories of Observation, Interpretation, Association, Comparison, Flexible Thinking, and Evidence.

Research Question 3: How does participation influence students' performance on standardized tests? When standardized test scores from the 2004-2005 MCAS and the 2005-2006 SAT 9 were analyzed, there were no differences between treatment and control students. This finding is not particularly surprising as the richness and complexity of responses to art and art museums is quite different than the type of multiple choice or short answer responses required on standardized tests.

Goal 3: Broadly disseminate study results to the museum education profession, and the formal education sector, through channels such as lectures and symposia, publications, website features, and a documentary video.

Over the years of the study, the ISGM and ILI have consistently shared the methodology, process of rubric development, and results of each year's study with the larger museum field. The final report will serve as a springboard for continued dissemination to inform the field's understanding of critical thinking in a museum multiple visit program. It is encouraging to have powerful data to support the assertion that such programs can positively influence the quality and quantity of children's critical thinking skills. Further, the most gratifying implication is that museum multiple-visit programs can be extremely effective in working with populations of urban at-risk children, many of whom have few opportunities to enrich their critical thinking through the visual arts or learn how to learn through art and art museums.

References

Adams, M., Luke, J. & Manuel, D. (2003). *Multiple-Visit Museum/School Programs in the Arts: What do students learn?* Technical research report. Annapolis, MD: Institute for Learning Innovation.

Costa, A. & Kallick, B., Eds. (2000). *Assessing and reporting on habits of mind*. Alexandria, VA: Association for Supervision and Curriculum Development.

Housen, A. (1983) *The Eye of the Beholder: Measuring Aesthetic Development*. Unpublished Ed.D. Dissertation. Harvard University.

Wetterlund, K., & Sayre, S. (2003). *2003 Art Museum Education Programs Survey*. Retrieved from http://www.museum-ed.org.

[2] Student interviews and "untour" conversations were digitally recorded and transcribed for analysis.

TABLE OF CONTENTS

INTRODUCTION **2**

RESEARCH RATIONALE 2
GOALS OF THE OVERALL THINKING THROUGH ART RESEARCH STUDY 2
RESEARCH RESULTS FROM YEARS 1 & 2 OF THE STUDY 3
FOCUS OF THE YEAR 3 RESEARCH STUDY 5

METHODS **7**

PARTICIPATING SCHOOLS 7
DATA COLLECTION PROCEDURES & PARTICIPATING SCHOOLS 8
Individual Student Interviews *8*
"Untours" in the Museum *9*
Standardized Tests *10*
CRITICAL THINKING SKILLS RUBRIC 10
DATA ANALYSIS 12

RESULTS AND DISCUSSION **13**

TO WHAT EXTENT DID PARTICIPATION IN THE SPP INFLUENCE STUDENTS' INDIVIDUAL CRITICAL THINKING SKILLS? 13
Students' Use of Overall Critical Thinking Skills during the Poster Interview *13*
Students' Use of Critical Thinking Sub-Categories during the Poster Interview *18*
Students' Use of Evidence to Support their Assertions during the Poster Interview *25*
TO WHAT EXTENT DID PARTICIPATION IN THE SPP INFLUENCE STUDENTS' CRITICAL THINKING SKILLS WITHIN A SOCIAL CONTEXT IN THE MUSEUM? 26
Students' Use of Overall Critical Thinking Skills during the "Untour" *26*
Students' Use of Critical Thinking Sub-Categories during the "Untour" *30*
Students' Use of Evidence to Support their Assertions during the "Untour" *34*
TO WHAT EXTENT DID PARTICIPATION IN THE SPP INFLUENCE STUDENTS' CRITICAL THINKING ON STANDARDIZED TESTS? 34
3rd Grade Reading Comprehension *34*
4th Grade Reading Comprehension *35*
5th Grade Reading Comprehension *35*
HOW DOES THE CTS RUBRIC TRANSLATE TO THE LARGER PROFESSIONAL FIELD? 36

CONCLUSIONS **37**

KEY FINDINGS 37
REFLECTIONS 40
REFERENCES 41
Appendix A: Year 1 Critical Thinking Skills Rubric *42*
Appendix B: Year 2 Critical Thinking Skills Rubric *46*
Appendix C: Individual Student Interview Protocol *49*
Appendix D: Small Group "Untour" Protocol *52*
Appendix E: Year 3 Critical Thinking Skills Rubric & Scoring Manual *55*
Appendix F: Critical Thinking Short Form *73*
Appendix G: Research to Practice: Testing a Tool for Assessing *75*
Critical Thinking in Art Museum Programs *75*

Thinking Through Art
Isabella Stewart Gardner Museum
School Partnership Program
Year 3 Research Results

April 2007

INTRODUCTION

Research Rationale

Museums and schools have a long history of working together to facilitate students' learning in and through the arts. While art museums have traditionally served school audiences through a range of single-visit tours, increasingly they are offering more extensive school programs in an effort to provide students with in-depth, comprehensive learning experiences. Studies suggest that as many as half of American museums offer some form of a multiple-visit school program in which students might visit the museum from two to ten times a year (Wetterlund & Sayre, 2003). Many other museums offer extended experiences such as pre- and post-visit activities in the classroom (Adams, Luke & Manuel, 2003; IMLS, 2002)

Looking across art museum multiple-visit programs, many aim to facilitate the development of young people's critical thinking skills (Adams, Luke & Manuel, 2003). However, most are unclear in their definition of critical thinking and their notions of how critical thinking skills are facilitated. Studies are needed that investigate what and how students learn from multiple-visit programs, and more specifically the extent to which these programs make a difference in young people's abilities to think critically about works of art.

Goals of the Overall Thinking Through Art *Research Study*

In 2003, the Isabella Stewart Gardner Museum (ISGM) and Institute for Learning Innovation (ILI) received a 3-year grant from the Department of Education to research students' learning in and from an art museum multiple-visit program. Research efforts within this grant were focused on assessing the effects of participation in an art museum multiple-visit program on the development of elementary students' critical thinking skills. The study had three overarching goals:

- To develop reliable and valid measures for assessing Critical Thinking Skills (CTS) in and from an art museum that are applicable to a wide range of school/museum educators across the country;

- To better understand how an art museum multiple-visit program can facilitate critical thinking amongst a generalizable sample of inner-city elementary students;
- To broadly disseminate study results to the museum education profession, and the formal education sector, through channels such as lectures and symposia, publications, website features, and a documentary video.

The ISGM's *School Partnership Program (SPP)* provided the context for this study. Launched in 1996, the *SPP* is a multiple-visit program serving K-8 students from neighboring inner-city public schools. At the beginning of this study in 2003, the program consisted of museum educators working together with participating classroom teachers to design lessons that connected with school curricula and were relevant to students' experiences inside and outside the classroom. Topics ranged from mythology and portraiture to the science involved in tapestry weaving and art conservation to turn-of-the century Boston history. Each class came to the museum for three to four visits during the school year, and every museum visit was preceded by a classroom lesson taught by museum staff.

By the end of the three year study the program evolved to a set of specific lessons designed to encourage development of students' observation and interpretation skills. Featured artworks encompass a wide range of styles, cultures, time periods, and media, and include more advanced discussions focusing on Museum archives, furniture, and art conservation. The number and type of visits remains the same. These ongoing experiences allow students to become familiar with the Gardner building and collection, develop their own artistic sensibilities, and feel at home at an important community institution.

Research Results from Years 1 & 2 of the Study

In Year 1 (2003-2004), exploratory research was conducted with teachers and students in the *SPP*. In an effort to develop a rubric for critical thinking skills relevant to learning in art museums, the Year 1 study sought to address three questions:

a) What is critical thinking?
b) How does the program address critical thinking?
c) What is the best way to assess critical thinking in the program?

What is critical thinking? Since the definition of critical thinking has many components, this study sought information on the ways teachers thought about critical thinking, directly and obliquely through discussions of meaning making and inquiry methods, as well as more open-ended discussions of teachers' core values and beliefs about teaching and learning. Findings from the teacher component of this study were analyzed for patterns and trends across the data. The in-depth teacher interviews were helpful in articulating areas for the *SPP* staff to clarify and reinforce. For the most part, teachers were in alignment with the goals and outcomes of the SPP and these findings provided the ISGM staff with information on how to further increase alignment between the program outcomes and teacher perceptions and needs.

How does the program facilitate critical thinking? The student component of the Year 1 study sought to document the ways in which the *SPP* facilitated the development of students' critical thinking skills. Student data were collected in two different contexts – in the museum while looking at and talking about authentic objects, and at school while looking at and talking about a reproduction of a work of art. Key findings from the student data, focused specifically on two themes: 1) *breadth* of students' critical thinking skills; and 2) *depth* of students' critical thinking skills. In addition, data from the teacher interviews provided further insight into teachers' perceptions of what and how students learn as a result of participation in the *SPP*.

To begin the process of creating a set of useful critical thinking skills that related directly to the multiple-visit program, ISGM staff articulated how they thought students enhanced their thinking skills as a result of participation in the *SPP.* These initial thoughts were then mapped onto critical thinking frameworks from Costa and Kallick (2000) and Housen (1983) to develop an initial set of nine categories of critical thinking skills that the ISGM staff felt the program most closely addressed. Then students' responses were coded across the following nine categories of critical thinking skills: 1) Observation and Articulation; 2) Interdependent Thinking; 3) Flexible Thinking; 4) Evaluation and Reflection; 5) Curiosity and Problem-Solving; 6) Activation of Prior Knowledge and Experience; 7) Categorization and Analytical Thinking; 8) Exploration and Speculation; and 9) Awareness and Respect for their environment. (See Appendix A: Year 1 Critical Thinking Skills Rubric)

During Year 1 of the grant, ISGM educators noticed that the *SPP* lessons, which often featured directed questions and activities, were not helping the students to learn to look closely, to remain open to multiple Interpretations, or to puzzle over meanings. For example, the 5th grade mythology lesson involved asking students to look for clues to a particular god or goddess's identity. Once the students "figured out" the main characters in a large ceiling painting, they felt they had accomplished the task. When museum educators then asked students to further discuss an image, students disengaged from the discussion.

What is the best way to assess critical thinking? In Year 2 (2004-2005), the first phase of a quasi-experimental study was conducted, using a posttest-only control-group design. At a broad level, it was intended as an initial effort to employ the CTS rubric in order to assess the impact of the *SPP* on students' critical thinking skills. Three complementary perspectives were taken in order to broadly and deeply investigate the effect of the multiple-visit program on students' critical thinking skills: 1) application of critical thinking skills to other contexts, measured using specific items from standardized tests; 2) school-situated assessment of critical thinking, measured through recorded conversations of students in front of three reproductions; 3) museum-situated assessment of critical thinking, measured through recorded conversations of students in the museum. The rubric was revised based on Year 2 findings and the nine categories were reduced to seven, with the addition of a rating scale for evidence that applied to student use of any of the seven basic critical thinking skills to support a statement or assertion (See Appendix B: Year 2 Critical Thinking Skills Rubric).

Results of the analysis of control and treatment group performance on two standardized tests, the Massachusetts Comprehensive Assessment System (MCAS) and the Stanford Achievement Test 9 (SAT 9), were not particularly illuminating. Patterns in any statistically significant differences

between Treatment/Control, grade level, or by school were not consistent, making it impossible to confidently draw any helpful conclusions. What differences did emerge had to do with the effect of one of the Treatment schools, which had more opportunities to attract advanced students than did the other three schools (1 Treatment and 2 Control groups). The differences might also have been affected by the fact that the *SPP* program was still more customized to each participating school than was to be the case in Year 3.

Findings from the student assessment were much more informative and encouraging. There were no statistical differences between the Treatment and Control groups in the School-Situated assessment. Both groups were similar in the types of Critical Thinking Skill categories they used, the amount of times they used a critical thinking skill, and the depth of their expansion and evidence provided relative to the critical thinking skill categories.

However, there were several areas of significant difference between Treatment and Control groups in the Museum-Situated assessment. The Treatment group used a wider variety of Critical Thinking Skill categories than did the control group and they used those categories more often. Even more encouraging was the finding that the Treatment group students used more descriptive language in the articulation of their thoughts about the works of art as well as a higher quality of evidence to support their assertions and opinions overall. It was curious, however, that there were no differences between groups when students talked about the poster reproductions at their school. Logically, one would expect similar findings in both the museum and school environments. Consequently, adjustments were made to the Year 3 study design to refine the sensitivity and strengthen the reliability of the rubric and coding system. At the same time, the ISGM program staff diagnostically assessed the program for what students are asked to know, do, and feel during the pre-visit experiences at the school. What students were accustomed to do during the pre-visit session was perhaps different enough from the school-situated assessment as to not yield a difference, while the open-ended discussions during the museum visits were more aligned with the museum-situated assessment.

In Year 2 of the grant, the pedagogy for the *SPP* shifted from Socratic-method to more open-ended questions, often using the Visual Thinking Strategies (VTS) questioning model. With this pedagogy, students were encouraged to apply previous experiences and knowledge to puzzle over meaning in the artworks; listen, respect, and share ideas with their peers; express their ideas through discussion, writing, and art making, and build confidence through support and encouragement from museum staff and classroom teachers. Videotaped lessons from both the classroom and museum context yielded evidence that this teaching strategy was more effective in developing students' critical thinking skills.

Focus of the Year 3 Research Study

In Year 3, research efforts were focused on assessing the extent to which an art museum multiple-visit program can facilitate critical thinking amongst participating 3rd, 4th, and 5th graders. Specifically, the study was designed to answer 3 specific research questions:

1. How does participation in the *SPP* influence students' individual critical thinking skills?
2. How does participation influence students' critical thinking skills within a social context in the museum?
3. How does participation influence students' performance on standardized tests?

To develop the curriculum during Year 3, ISGM educators considered teacher interviews and informal comments, as well as carefully examined emerging trends in the student data gathered during the first year of the project. With these trends in mind, the *SPP* was restructured to use VTS to better meet the needs of participating students and teachers, and to take full advantage of the multiple-visit nature of this unique partnership and the opportunities to interact and learn from actual artworks and objects in the museum environment. *SPP* goals were redefined to: 1) Learn to look at and make meaning from works of art and 2) Gain familiarity with the museum environment and feel comfortable using the Gardner as a community resource.

ISGM educators collaborated with *Thinking Through Art* project advisor and Visual Understanding in Education (VUE) co-founder Philip Yenawine to carefully identify artworks in the Gardner Museum collection that were developmentally appropriate for each grade in the Partnership. VUE's other co-founder, cognitive psychologist Abigail Housen, has repeatedly demonstrated that students are capable of building their personal understanding of art and ability to make meaning through regular, sequential discussions of artworks with their peers.[3]

The Gardner's *SPP* curriculum starts with simple images that lend themselves to storytelling, featuring multiple figures in a familiar or recognizable situation or activity. Then, as the students become more adept at decoding images, the imagery increases in complexity, with puzzling attributes that will challenge them to build upon previous experiences and scaffold off each others' responses. Pedagogy used by the museum educators encourages students to make observations and inferences based on what they see, and also to support their assertions with evidence. To support the older students' more analytical ways of decoding works of art, additional questions are added by the museum educators.

In addition, writing and sketching activities were added to both classroom and museum-based lessons so that students have the opportunity to individually reflect and respond to works of art. Prompts for writing mirrored the VTS questions used during the group discussions to reinforce ways of looking and to encourage the students to provide evidence for their thoughts. The corresponding studio projects were also redesigned to be more open-ended, paralleling the group discussions rather than reinforcing concepts introduced by museum educators. We were confident that by modifying the curriculum we would engage students in discussions and activities that not only complemented each other but grew in complexity over their three years in the program.

[3] Housen, Abigail, *Aesthetic Thought, Critical Thinking, and Transfer*. Arts and Learning Journal, Volume 18, No. 1, pp. 99-132, 1992.

METHODS

The Year 3 study was designed as a one-year, quasi-experimental investigation, using a posttest-only control-group design to assess the effects of participation in an art museum multiple-visit program on the development of elementary students' CTS.[4] Multiple methods were used to assess critical thinking, including individual student interviews, group "untours" in the museum, and performance on state-based standardized tests.

Participating Schools

Students in 3rd, 4th, and 5th grades from 5 elementary schools participated in the study, including 2 treatment schools (Farragut Elementary and Tobin Elementary) and 3 control schools (Hamilton Elementary and Kennedy Elementary initially; Philbrick Elementary was added later). While every effort was made to match the control schools with the treatment schools along several factors explained below, the realities of selecting, recruiting, and retaining control school matches proved challenging.

Initially, Treatment schools were selected in Year 1 of the study, based on their longevity within the *SPP* program: these were Farragut and Tobin Schools which remained in the study all three years. Because the two Treatment schools differed enough from each other, each Treatment school was matched to a Control school in Year 2 of the study, based on comparability of test scores, race/ethnicity distribution, and socio-economic status of the school population as measured by percent of free and reduced lunch eligibility (See Table 1 for specific data). Importantly, control schools were also matched based on similarity of arts instruction. Specifically, neither the Treatment nor Control schools had any form of regular visual arts instruction by a certified art specialist. The original Control schools selected were Hamilton (comparable to Farragut) and Kennedy (comparable to Tobin). At the start of the Year 2 study the fifth-grade teacher at Hamilton no longer wished to participate so one fifth-grade class of students from Philbrick Elementary School was added to the Control group. There was one class of combined fourth and fifth graders at Hamilton that continued to participate in the study.

Table 1: Description of the Participating Schools

Comparison Factors	Treatment Schools		Control Schools		
	Farragut	Tobin	Hamilton	Kennedy	Philbrick
Percent 3rd grade Passing MCAS	86%	70%	70%	82%	94%
Percent 4th grade Passing MCAS	86%	55%	49%	70%	95%
Percent of 3rd & 4th Grade Students eligible for Free- or Reduced-Lunch	75%	98%	88%	94%	78%

[4] Given that the program is situated within the social setting of the classroom, a true experimental study is not possible since subjects cannot be assigned to groups randomly. The use of a posttest-only control-group design permits researchers to make causal inferences about the effect of the program on students' learning, and rules out the potential test bias created through the use of a pretest.

Percent Hispanic Students	26%	69%	20%	70%	29%
Percent African-American Students	49%	30%	44%	28%	54%

Source: Boston Public Schools website 2003-2005 – www.boston.k12.ma.us

Data Collection Procedures & Participating Schools

Individual Student Interviews

Interviews with individual students assessed their ability to employ critical thinking skills while looking at a reproduction of a work of art. This "think aloud" technique was piloted in January 2006 to ensure students were able to successfully engage in this task. Interviews were conducted from March through June 2006. Students were shown the same work of art in both the pilot interviews and the study interviews. The work was Caravaggio's *Supper at Emmaus*; this poster was one of the three used in the Year 2 study; it was selected because of the style of the work (a 17th century European old-master painting similar to works students encounter at the Gardner)), and the subject (a group of people seated around a table but with the nature of their interactions open to Interpretation). Students were taken out of class one at a time by a researcher and brought to a quiet space in the school where they were shown the reproduction. They were asked to talk on their own into a digital audio recorder about the work of art, with the researcher providing minimal prompts, such as "anything else" or "tell me more" (See Appendix C: Individual Student Interview Protocol).[5] Student comments and the researcher's prompts were later transcribed in full.

At the outset of the Year 3 study, parents from all five schools were mailed consent letters which explained the nature of the research study and requested the participation of their child. From students who returned their participation form, researchers selected a random sample for individual interviews. The sample was restricted to students who did not participate in the Year 2 study in order to reduce the possibility of confounded results based on previous study experiences. The treatment sample consisted of students who had previously participated in the *SPP* to varying degrees.

A total of 135 students participated in individual reproduction interviews (See Table 2 below for specific percentage distribution of students across groups). Of these, 48% (n=64) were treatment students and 52% (n=71) were control students. The distribution of interviews by grade at treatment and control schools was comparable; 34% of the students interviewed were in third grade, 36% in fourth grade, and 30% in fifth grade.

[5] The interview protocol is an adaptation of Abigail Housen's Aesthetic Development Interview (ADI) in that it provides an open environment for students to talk about the work of art in any way they wish. There are two important differences between the interview protocol for the ISGM study and the ADI. First, researchers administering the ADI can say "Where is your eye moving now?" when students stop talking or say they have nothing else to talk about. The ISGM protocol does not permit researchers to prompt the students to look in this way. Secondly, the ADI is administered with the intention to determine a subject's aesthetic development as well as assess critical thinking. This study does not seek to assess people as to their position on Housen's developmental scale. Rather, the ISGM interview protocol seeks data on how students employ critical thinking skills when looking at art in specific contexts.

Table 2: Distribution of students who participated in individual interviews by school

School	Percent of interviews (n=135)
Treatment	
Farragut	30%
Tobin	18%
Control	
Hamilton	9%
Kennedy	39%
Philbrick	4%

"Untours" in the Museum

Small group "untours" assessed students' ability to apply their knowledge and skills not just to one individual work of art, but to the larger museum context and the multiple objects within this context. These "untours" were completed during March through June 2006 with students who had previously completed an individual interview. Students were brought to the ISGM in class-based groups of five to six students, anywhere from one day to 20 days after their individual interview. Each student was outfitted with a wireless microphone and recording device, and taken to the Raphael gallery, where they were free to go wherever they wanted within the room and look at and talk about whatever they wanted (See Appendix D: Small Group "Untour" Protocol).[6] Although multiple students could be heard on any one recording, transcribers focused on the voice of the child wearing the microphone. The comments of that student and any adult voices on the recording were transcribed.

A total of 116 students participated in the "untours." Of these, 48% (n=56) were treatment students and 52% (n=60) were control students. The distribution of interviews by grade was comparable; 35% of the students interviewed were in third grade, 36% in fourth grade, and 29% in fifth grade. Since the *SPP* program was conducted solely in English, students were instructed to speak English during the interviews; however, five students spoke a very small amount of Spanish during the course of the "untour." In these cases, the Spanish was not translated or coded; but all statements the child made in English were coded.

[6] The protocol for the "untour" was first developed in a 1999 study by the Institute for Learning Innovation with elementary school students at the Dallas Museum of Art in an effort to study the effect of a pre-visit interactive CD experience on the museum visit. Removing the direction or guidance of an adult allows children to explore a gallery on their own, providing a glimpse of what matters to them.

Table 3: Distribution of students who participated in "untours" by school

School	Percent of "untour" (n=116)
Treatment	
Farragut	28%
Tobin	21%
Control	
Hamilton	9%
Kennedy	39%
Philbrick	4%

Standardized Tests

Finally, standardized test scores were collected and analyzed to determine the degree to which *SPP* participation influenced students' critical thinking abilities in the broader context of general reading and comprehension. Researchers worked with Boston Public School staff to secure class-based test scores for all relevant students in the study. Student test score data collected for the purpose of this study included the Massachusetts Comprehensive Assessment System (MCAS) test and the Stanford-9 (SAT-9). This report includes results from the 2005 SAT-9; a follow-up report will summarize results from the 2006 MCAS, since those results were not available at the time of writing this report.

A total number of 199 treatment and 211 control students participated the SAT-9 testing. See Table 4 for an overview of the data sample by school.

Table 4: Number of students from each school who participated in the SAT-9 (Fall 2006).

SAT9 Tests	**Treatment Schools**		**Control Schools**	
	Farragut	**Tobin**	**Hamilton**	**Kennedy**
3rd grade	34	47	19	64
4th grade	42	18	14	65
5th grade	40	18	0	49

Source: SAT-9 data provided by the Boston Public School System.

Critical Thinking Skills Rubric

The major dependent variable in this study was critical thinking, measured using an empirically-grounded rubric informed by the work of Abigail Housen (2002 & 1983) and Costa & Kallick (2000), and tested and refined in Years 1 and 2 of the study. The development of the critical thinking skills rubric over the years of this study was a challenging and time-intensive process. The first year began with nine critical thinking categories that were drawn from the literature and from the *SPP* educators (See Appendix A). After analyzing the Year 1 data with the critical thinking skills rubric it became clear that it needed some refinement. In some cases there were

categories that seemed to overlap each other, making coding more difficult and therefore less reliable. Many of the category descriptions in the Year 1 rubric tried to measure multiple behaviors that did not necessarily happen together, thus further threatening the reliability of the rubric.

For the Year 2 study quasi-experimental design, the rubric was revised to include seven basic critical thinking skills based on the analysis of the Year 1 data (See Appendix B). In addition, a separate over-arching critical thinking skill of providing evidence for assertions was included. This skill could be applied to any instance of the seven critical thinking skills when used to support a statement. The use of support for assertions (Evidence) was embedded in the Year 1 rubric exploration and speculation category. Some of the literature consulted for this research separated out evidence as an independent critical thinking skill. However, the Year 1 data was such that separating out evidence was problematic.

For this final Year 3 study, the CTS rubric was further refined (See Appendix E: Year 3 Critical Thinking Skills Rubric & Scoring Manual). Some of the categories were renamed, such as changing the Year 2 category of "Analysis" to "Comparison" and changing Year 1 category of "Revision" to "Flexible Thinking" to better reflect what students actually did in the data. More importantly, three of the categories, Observation, Interpretation, and Evaluation, were assigned sub-categories, in an effort to further breakdown the larger CTS category into more discrete parts. This decision was made based on Year 2 data findings that there were different ways in which students observed, interpreted, and evaluated. Specifically, students observed what something was, what action was happening, physical characteristics, and media. Label reading behaviors were also recorded as a sub-category for Observation at the request of the *SPP* staff. Sub-categories for Interpretation were interpreting how something was used, conditions or emotional states, relationships and identity, and intentions. There were two kinds of Evaluation students engaged in, an evaluation based solely on personal preference or opinion and an evaluation based on the perceived merits of a work. Breaking down the most frequently used CTS provided researchers with the flexibility to look at student responses on both a collective (looking at all the sub-categories together in one category) and a fine-grained level (splitting the single categories apart).The other critical thinking skills were not divided into sub-categories because the Year 2 data analysis did not suggest this was necessary.

Results from the two prior years of data analysis showed that many of the responses from children were fairly straightforward, simple, and easily isolated phrases or sentences. In these cases, the codable unit is a single critical thinking skill, and can consist of one or two words, a phrase, or a simple sentence. We found that Observation, Interpretation, Association, and Evaluation was always coded as isolated, while Problem-Finding, and Comparing were in some cases isolated, and Flexible Thinking was never an isolated skill, as it requires a combination of isolated skills to occur. Some students employed two or more isolated thinking skills to create a different kind of thinking skill. In these cases, the codable unit was a combination of two or more critical thinking skills, and usually consisted of several phrases or a compound sentence ("synergistic"). Only three of the critical thinking skills were ever found to be synergistic: Problem-Finding, Comparing, and Flexible Thinking.

The final development of the CTS rubric was largely dependent on the prior two years of data analysis. However, some skills were left in the Year 3 rubric despite the fact that students engaged in them much less frequently than they did in skills such as Observation and Interpretation. This decision was made because *SPP* staff wanted to continue to assess these skills as they wanted to enrich the curriculum to provide more opportunities for students to engage in those skills. Therefore, the rubric could serve as a reflection of what is actually happening with students and as a diagnostic tool for practitioners to assess the effectiveness of programmatic changes. This was, after all, one of the primary intentions for this research – to develop a model for assessing critical thinking skills in museum multiple visit programs that practitioners could use to self-assess their programs. A practitioner checklist was developed after the analysis of the Year 3 data and is included in Appendix F: Critical Thinking Skills Short Form for Practitioners.

One of the many variables that could possibly influence the development of critical thinking skills the researchers considered tracking was that of contact hours. While this was not a developmental study in that we did not seek to track how children grew aesthetically, we did recognize that time-on-task might be an important variable. Table 5 illustrates the range of contact hours by average classroom, museum, and total time spent. As is clearly evident by this table, the variation in contact hours was substantial which meant that as important as contact hours probably is to the development of critical thinking skills, the size of the sample together with the variability in range of contact hours prevented any deeper analysis by this variable.

Table 5: Distribution of average contact hours by school and grade level

Treatment School	Avg CLASSROOM contact hours	Avg MUSEUM contact hours	Avg PROGRAM contact hours
3rd grade TOBIN	10	6	16
3rd grade FARRAGUT	11	7	18
4th grade TOBIN	15	9	24
4th grade FARRAGUT	15	9	23
5th grade TOBIN	18	12	29
5th grade FARRAGUT	30	20	50

Data Analysis

Transcripts were coded by three trained coders in the Spring/Summer of 2006. Since more than one person served as a coder and, being human, individuals are not always consistent in the ways they look at the world, it was important to check for inter-rater reliability (IRR)[7] to determine if the coders were consistent with each other in the choices they made when assigning codes to the student transcripts. In the individual interviews the IRR for the seven CTS categories was 81% and 33% for coding of evidence. Reliability for the seven CTS categories in the untour was 78% with 100% agreement for evidence. An IRR percentage of 60% or more is considered acceptable and assures that the reliability or consistency of the coders was sufficient and results can be

[7] IRR was calculated using Cohen's Kappa. For the individual interview setting, Kappa=.81 for the CTS coding and Kappa=.33 for the evidence coding. For the "untour" setting, Kappa=.78 for the CTS coding and Kappa=1 for coding of evidence.

confidently interpreted. The lower IRR percentage for evidence coding in the interviews (33%) indicates that the results for the evidence category should be cautiously interpreted.

RESULTS AND DISCUSSION

To what extent did participation in the SPP influence students' individual critical thinking skills?

When asked to "think out loud" about the Caravaggio reproduction, treatment students were consistently able to say more than control students about the artwork, demonstrating higher frequencies of critical thinking skills. After researchers factored out extraneous, repetitious, and unrelated language, data showed that treatment students talked significantly longer about the art work, averaging 28 lines per interview compared to 14 lines for control students (Mann-Whitney U=828.500, p=0.000). More importantly, treatment students generated significantly more instances of critical thinking skills and evidence. What follows is a detailed summary of students' individual performance in interviews, in terms of 1) use of overall Critical Thinking Skills; 2) use of specific Critical Thinking Skills; and 3) provision of evidence in support of students' assertions.

Students' Use of Overall Critical Thinking Skills during the Poster Interview

Regardless of whether they were in the treatment or control group, students were more likely to use Observation and Interpretation skills when talking about the reproduction (see Table 6). Overall, the median number of times a student used observational skills during the reproduction interview was 10.00; the median number of Interpretational skills used was 10.00. The medians for the other CTS were 0.00 per interview, indicating that more than half of the students did not use these CTS at all. This trend towards greater usage of Observation and Interpretation was also observed during the pilot phase of the project (Year 2). Most likely this trend is related to the nature of observing and interpreting, which tend to be acquired at an earlier developmental stage; in addition, both Observation and Interpretation may often serve as building blocks for other skills, such as Comparison or Flexible Thinking.

Overall, students were least likely to use Comparison and Evaluation in the poster interviews, again regardless of whether they were treatment or control students. This may be due to the fact that they were focused on a single artwork, and thus were less apt to draw comparisons or state multiple preferences/opinions than if they were looking at several objects. Also, students may have been less interested in expressing likes or dislikes when talking out loud on their own than they might in a social context with their peers.

Table 6: Students' use of overall critical thinking skills during the poster interview (N=135).

Critical Thinking Skills	Median*	Range
Observation	10.00	0 – 172
Interpretation	10.00	0 – 288
Evaluation	0.00	0 – 11
Association	0.00	0 – 13
Problem-Finding	0.00	0 – 40
Comparison	0.00	0 – 12
Flexible Thinking	0.00	0 – 23

*The median is the midpoint of the distribution, meaning that half of the students are below this point and half of the students are above this point. Where the median is 0, it means that half or more of students did not use that critical thinking skill.

When the treatment and control group interviews were analyzed for differences in the amount of CTS used, significant differences emerged. Of the seven broad CTS, there were significant differences between the treatment and control groups for five skills: Observation (Mann-Whitney U=1092.000, p=0.000), Interpretation (Mann-Whitney U=942.500, p=0.000), Association (Mann-Whitney U=1829.000, p=0.028), Comparison (Mann-Whitney U=1762.500, p=0.002), and Flexible Thinking (Mann-Whitney U=1173.000, p=0.000) (see Table 7). For all of these CTS, treatment students used significantly higher amounts of the skill than did control students.

Table 7: Treatment and control students' use of overall critical thinking skills during the poster interview

Critical Thinking Skills	Treatment Group Median (n=64)	Control Group Median (n=71)	Significance
Observation	18.00	6.00	* *(p=.000)*
Interpretation	18.50	6.00	* *(p=.000)*
Evaluation	0.00	0.00	Not significant
Association	1.00	0.00	* *(p=.028)*
Problem-Finding	0.00	0.00	Not significant
Comparison	0.00	0.00	* *(p=.002)*
Flexible Thinking	1.00	0.00	* *(p=.000)*

For Observation and Interpretation, the differences between treatment and control students as portrayed in Table 7 are easily seen; in both cases, treatment students on average used significantly more of those two skills than did control students. However, for Association, Comparison, and Flexible Thinking, the differences between treatment and control students are seemingly more negligible, especially in the case of Comparison where there does not seem to be any difference between the two groups. For these three CTS, it is necessary to go beyond a comparison of medians in order to truly understand the significant differences between treatment and control students.

Association. Treatment students demonstrated significantly higher instances of Association during poster interviews than did control students. However, this is hard to see by simply comparing the average number of times each group used this critical thinking skill (i.e., Treatment median=1; control median=0). Instead, it is necessary to more closely examine the overall distribution of each group's use of Association. Figures 1 and 2 below show that while control students' cluster around 0, with only a few instances of Association, treatment students are more dispersed, with more of them actually using Association and in higher numbers.

Figure 1: Distribution of treatment students' usage of Association during poster interviews.
Note: To interpret this figure as well as figures 2-20, the X (horizontal) axis: (e.g. in figure 1 is Association) = number of times that CTS was used; Y (vertical) axis: Frequency = number of students who used that CTS that many times.

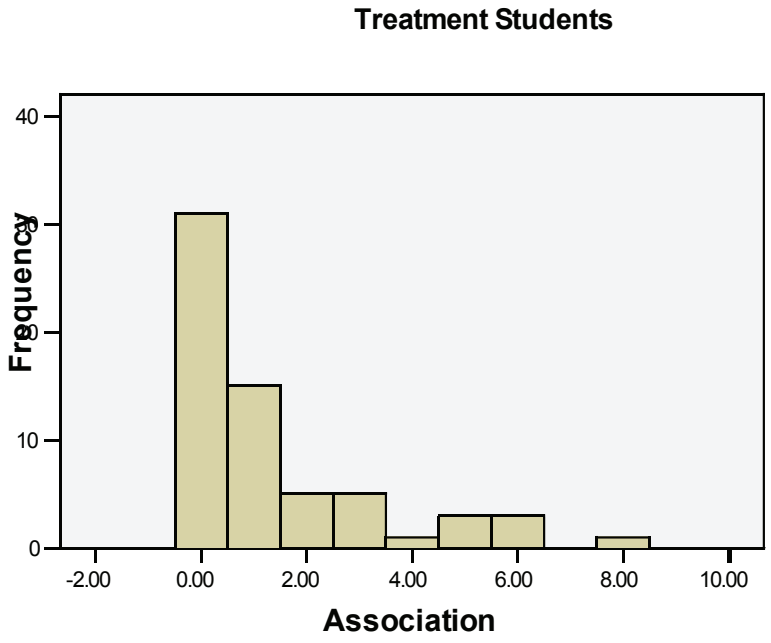
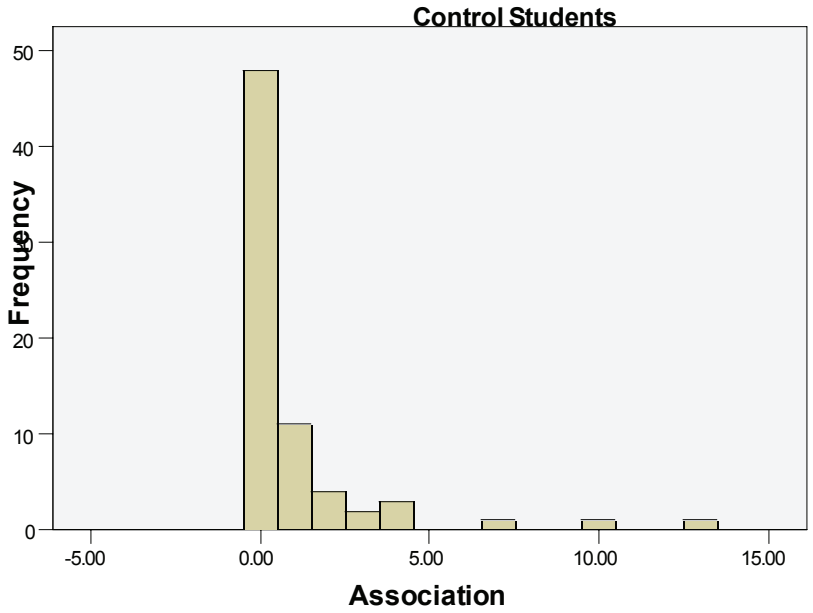


Figure 2: Distribution of control students' usage of Association during poster interviews.



Comparison. Interestingly, both treatment and control students used Comparison an average of 0 times (see Table 7). The question arises then as to how there can be a significant difference between these two groups if both used this CTS the same number of times. To answer this question, it is necessary to examine a more detailed distribution of each group's use of Comparison. Figures 3 and 4 below show that in both groups students cluster around 0 (which is why the average or median for each group is 0). However, there are clearly more treatment students with a frequency of higher than 0, making for a significant difference between treatment and control students in their use of Comparison.

Figure 3: Distribution of treatment students' use of Comparison during poster interviews.

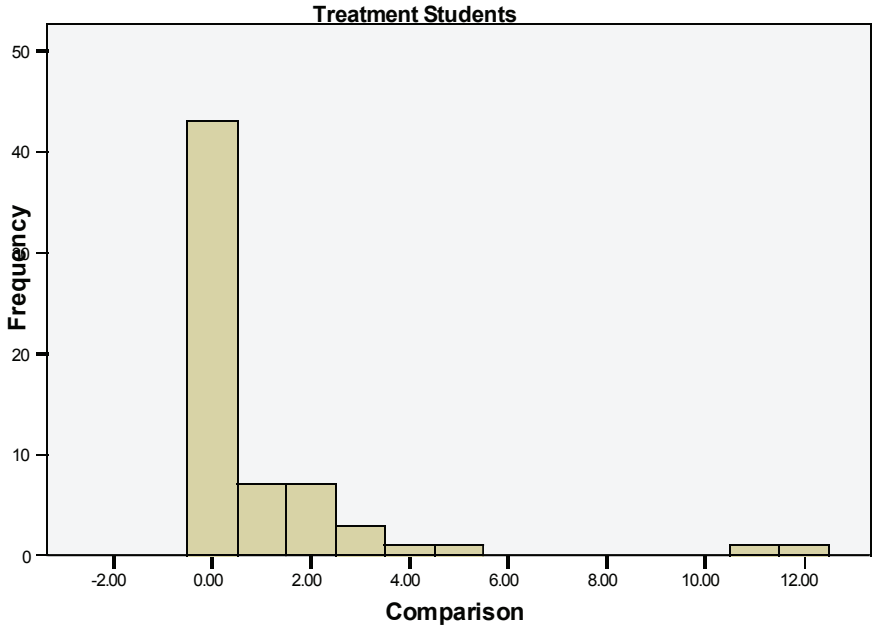
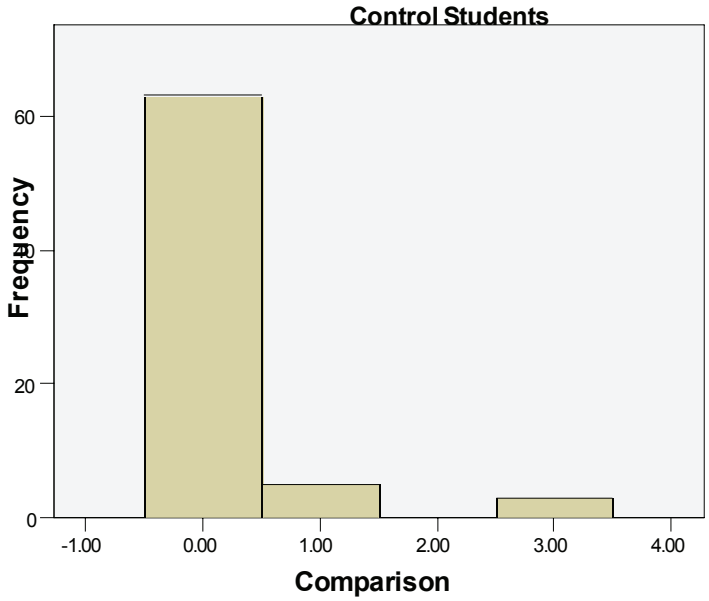


Figure 4: Distribution of control students' use of Comparison during poster interviews.



Flexible Thinking. Treatment students used Flexible Thinking significantly more times than did control students during poster interviews. Figures 5 and 6 below show the while students in both groups cluster around 0, many more treatment students used Flexible Thinking than did control students.

Figure 5: Distribution of treatment students' use of Flexible Thinking during the poster interview.

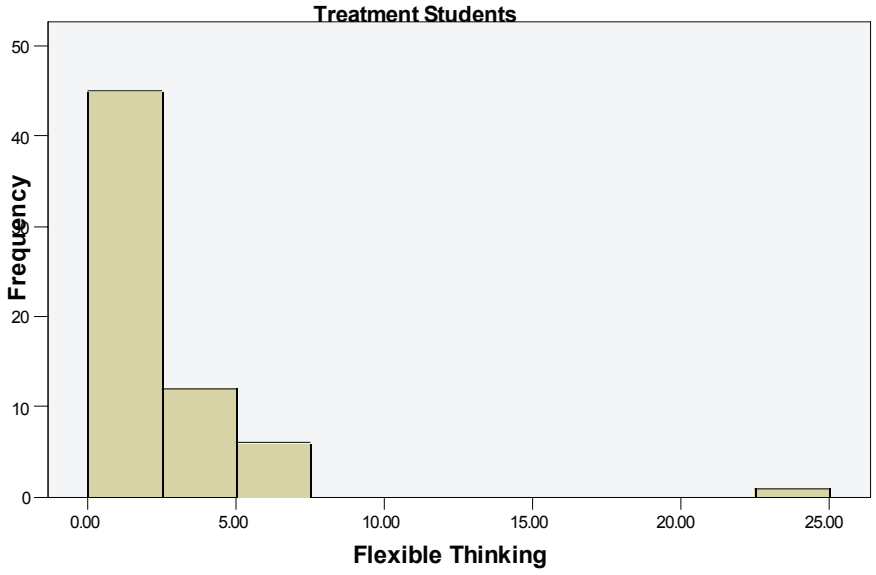
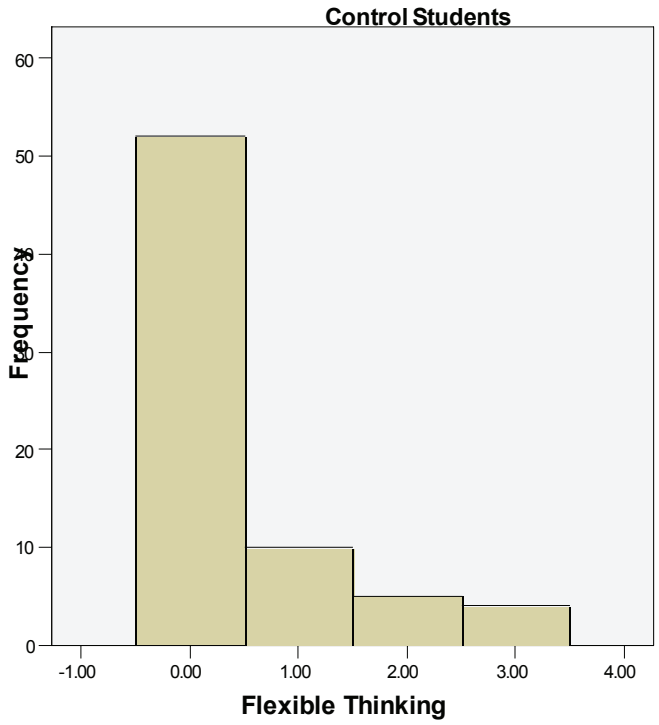


Figure 6: Distribution of control students' use of Flexible Thinking during the poster interview.



Students' Use of Critical Thinking Sub-Categories during the Poster Interview

Within the seven independent CTS, specific sub-categories were identified to more explicitly define how students thought about the works of art. As shown in Table 8, some sub-categories were used by students more often as they looked at the reproduction (regardless of whether they were treatment or control students). For example, within Observation, giving specific details, features, or the location of an object in the work (skill 1.3) was used five or more times by half of the students, more than any other sub-category used during the reproduction interview. Specifically, these instances included when the student noted the color, shape, size, or quantity of objects or people; identified particular features or missing features; or situated an object in a physical context. The following examples represent how students used this sub-category:

The tablecloth…is like dark, like deep red (treatment student, 5th grade)

It looks like there's a rag tied around his waist (treatment student, 5th)

They don't have fancy clothes (control student, 3rd)

I see a door behind someone's back (treatment student, 3rd)

And there's a big, big glass of water (control student, 3rd)

They're in a dark room (control student, 4th)

Table 8: Students' usage of critical thinking skill sub-categories during the poster interview (N=135).

Critical Thinking Skills	Median*	Range
Observation (1)		
What something is or is not (1.1)	2.00	0 – 19
Actions (1.2)	2.00	0 – 28
Details, features, locations (1.3)	5.00	0 – 129
What or how something is made (1.4)	0.00	0 – 5
Reading labels (1.5)	0.00	0 – 2
Interpretation (2)		
Use of objects (2.1)	0.00	0 – 3
Implicit Conditions (2.2)	1.00	0 – 30
Identity (2.3)	1.00	0 – 35
Intentions or narrating (2.4)	6.00	0 – 223
Evaluation (3)		
Personal opinion (3.1)	0.00	0 – 10
Merits of the work (3.2)	0.00	0 – 3
Association** (4)	0.00	0 – 13
Problem-Finding (5)		
Isolated	0.00	0 – 9
Synergistic	0.00	0 – 37
Comparison (6)		
Isolated	0.00	0 – 7
Synergistic	0.00	0 – 7
Flexible Thinking** (7)	0.00	0 – 23

*The median is the midpoint of the distribution, meaning that half of the students are below this point and half of the students are above this point. Where the median is 0, it means that half or more of students did not use that critical thinking skill.
** CTS has no sub-categories.

On the other end of the spectrum, students seldom observed how the work was made or what it was made of (1.4), most likely because they were looking at a reproduction, or simply focused their attention on the content of the reproduction.

Like Observation, the skill of Interpretation was divided into sub-categories in order to better define how students were using this skill. As Housen's theory of aesthetic development would predict, researchers found that both treatment and control students were equally likely to narrate a story about the painting (skill 2.4), with half of the students using this sub-category six or more times during the poster interview. These instances included interpreting the intentions of characters, or "setting the scene" with what's going on, what is about to happen, or where people are, as the following examples help illustrate:

I think it's Thanksgiving (treatment, 4th grade)

And it looks like they were about to eat before they started talking (control, 4th)

I think that they're gonna have dinner (treatment, 4th)

And [it] looks like she is telling them something bad (treatment, 3rd)

Or it looks like they've come to the person for a problem that they want to solve (control, 5th)

Students rarely interpreted how an object was used (skill 2.1), arguably because the context of looking at a single poster offers little opportunity to do so.

When treatment and control students were compared, differences were found between the two groups in terms of their use of the sub-categories of critical thinking. As illustrated in Table 9, treatment students used significantly more instances of sub-categories within Observation, Interpretation, Evaluation, and Comparison. Within Observation and Interpretation, a comparison of medians between treatment and control students clearly demonstrate their differing uses of CTS sub-categories. For instance, three of the five Observation sub-categories were used significantly more often by treatment students (1.1: Mann-Whitney U=1493.000, p=0.000; 1.2: Mann-Whitney U=1618.000, p=0.003; 1.3: (Mann-Whitney U=974.500, p=0.000), while three of the four Interpretation sub-categories were used significantly more often by the treatment group (2.2: Mann-Whitney U=1256.500, p=0.000; 2.3: Mann-Whitney U=1364.000, p=0.000; 2.4: Mann-Whitney U=102.500, p=0.000).

Table 9: Treatment and control students' usage of specific critical thinking skills during the poster interview

Critical Thinking Skills	Treatment Group Median (n=64)	Control Group Median (n=71)	Significance
Observation (1)			
What something is or is not (1.1)	3.50	1.00	* *(p=.000)*
Actions (1.2)	3.00	2.00	* *(p=.003)*
Details, features, locations (1.3)	12.00	2.00	* *(p=.000)*
What or how something is made (1.4)	0.00	0.00	Not significant
Reading labels (1.5)	0.00	0.00	Not significant
Interpretation (2)			
Use of objects (2.1)	0.00	0.00	Not significant
Implicit Conditions (2.2)	3.00	1.00	* *(p=.000)*
Identity (2.3)	3.00	1.00	* *(p=.000)*
Intentions or narrating (2.4)	12.00	4.00	* *(p=.000)*
Evaluation (3)			
Personal opinion (3.1)	0.00	0.00	Not significant
Merits of the work (3.2)	0.00	0.00	* *(p=.046)*
Association* (4)	-	-	-
Problem-Finding (5)			
Isolated	0.00	0.00	Not significant
Synergistic	0.00	0.00	Not significant
Comparison (6)			
Isolated	0.00	0.00	* *(p=.001)*
Synergistic	0.00	0.00	* *(p=.032)*
Flexible Thinking*(7)	-	-	-

* CTS has no sub-categories.skill.

Within Evaluation and Comparison, the differences between treatment and control students are more difficult to see (i.e., Table 9 shows each group with a median of 0), requiring a closer look at the distribution of each group in order to truly understand how treatment and control students are different.

Evaluation. Although giving a personal opinion about the work (3.1) was not significantly different between the two groups, commenting on the merit of the work or the artistry (3.2) was significantly different. In fact, control students were significantly more likely to comment on the merit of the work as compared to treatment students (Mann-Whitney U=2086.500, p=0.046). Although both groups commented on merit an average of 0 times, Figures 7 and 8 below show that there are more control students who actually comment on merit, which is why there is in fact a significant difference in favor of the control students within this sub-category of critical thinking.

Figure 7: Distribution of control students' use of Evaluation, sub-category 3.2 (commenting on the merits of artistry of the work) during individual poster interviews.

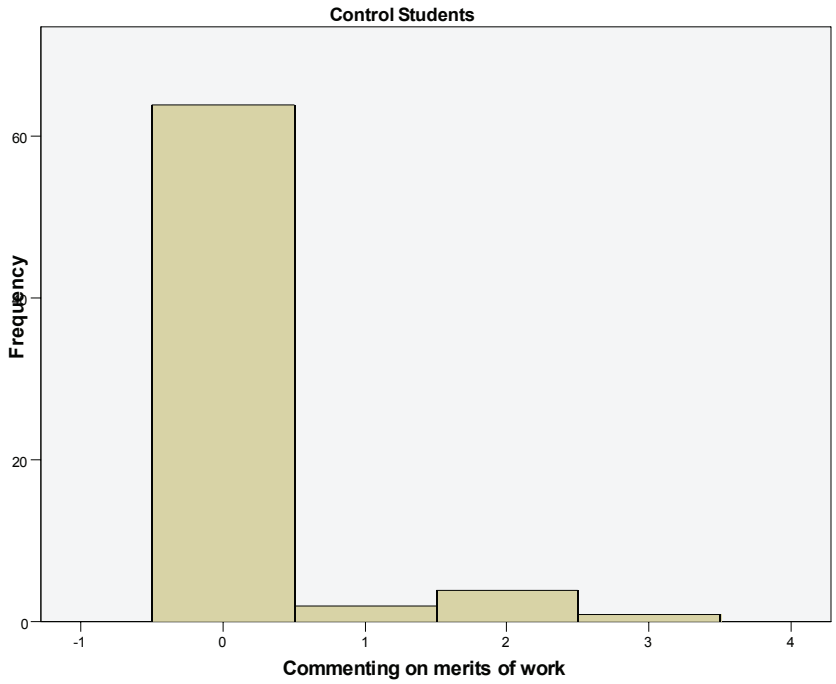
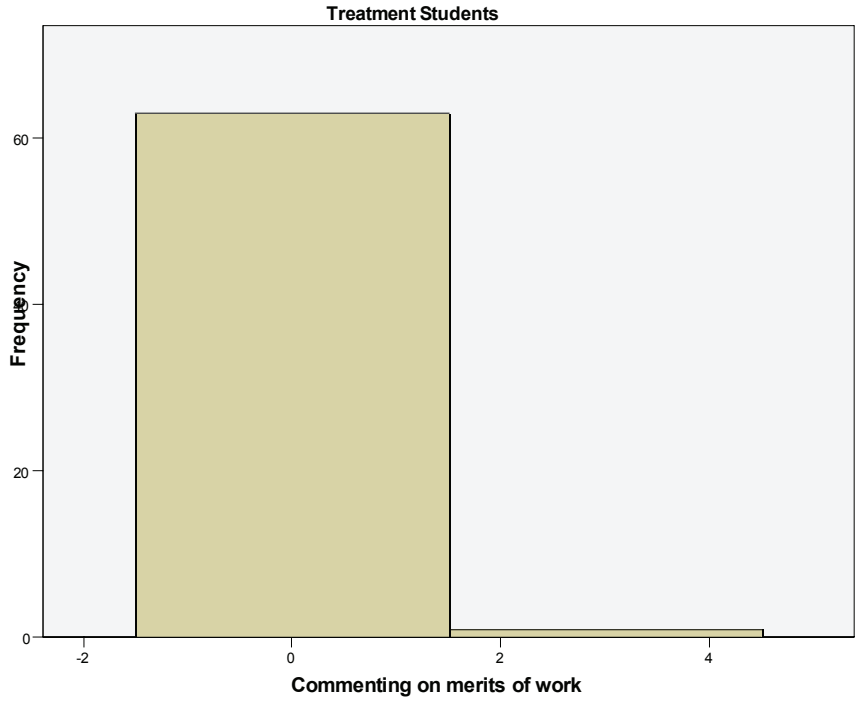


Figure 8: Distribution of treatment students' use of Evaluation, sub-category 3.2 (commenting on the merits or artistry of the work) during individual poster interviews.



Comparison. Both isolated and synergistic instances of Comparison were used significantly more often by the treatment group (6 isolated: Mann-Whitney U=1795.000, p=0.001; 6 synergistic: Mann-Whitney U=1991.000, p=0.032). Again, these differences are difficult to see at first, since a presentation of medians simply shows that both treatment and control students used Comparison an average of 0 times. Figures 9 and 10 below show that for isolated instances of Comparison, treatment students were more likely to actually demonstrate this skill.

Figure 9: Distribution of treatment students' use of isolated Comparison during poster interviews.

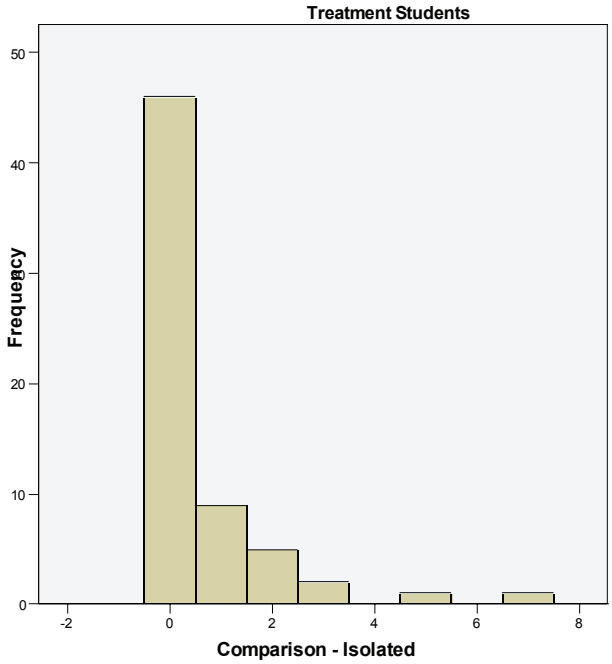
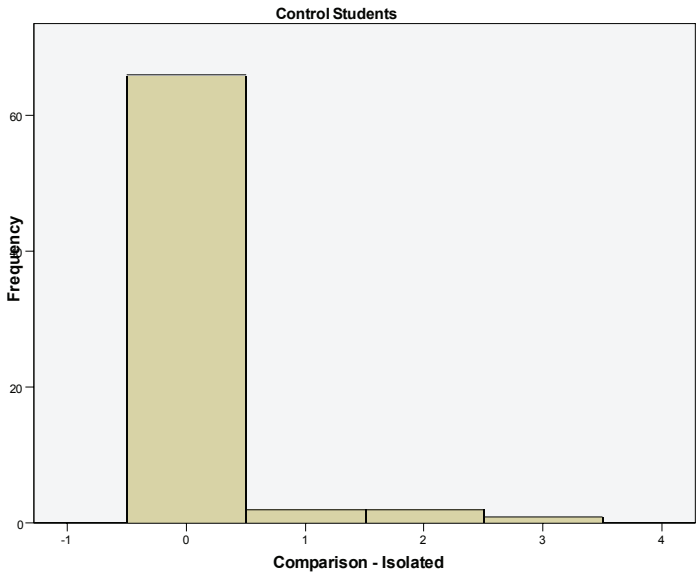


Figure 10: Distribution of control students' use of isolated Comparison during poster interviews.



Similarly, Figures 11 and 12 below show that treatment students were also more likely to use synergistic instances of Comparison.

Figure 11: Distribution of treatment students' use of synergistic Comparison during poster interviews.

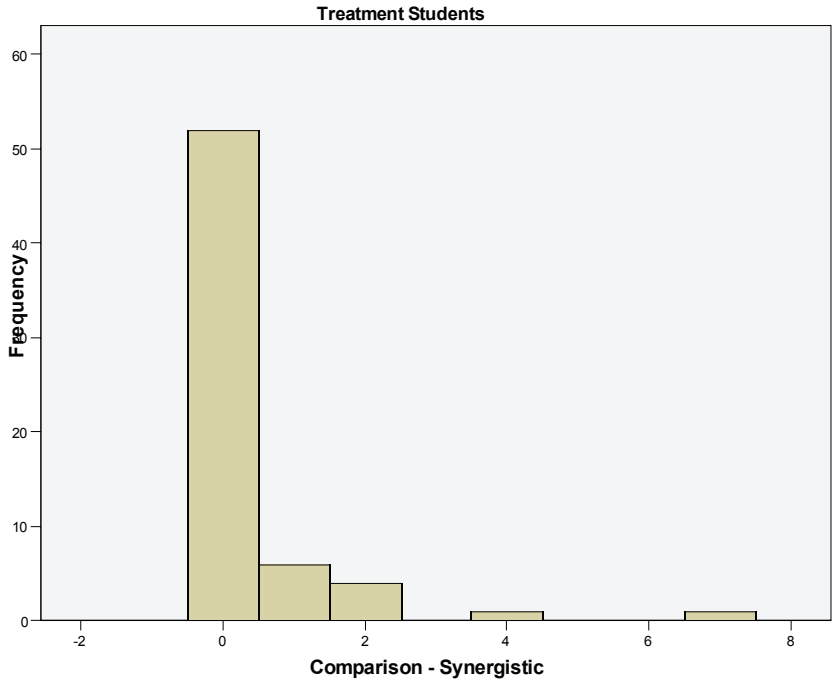
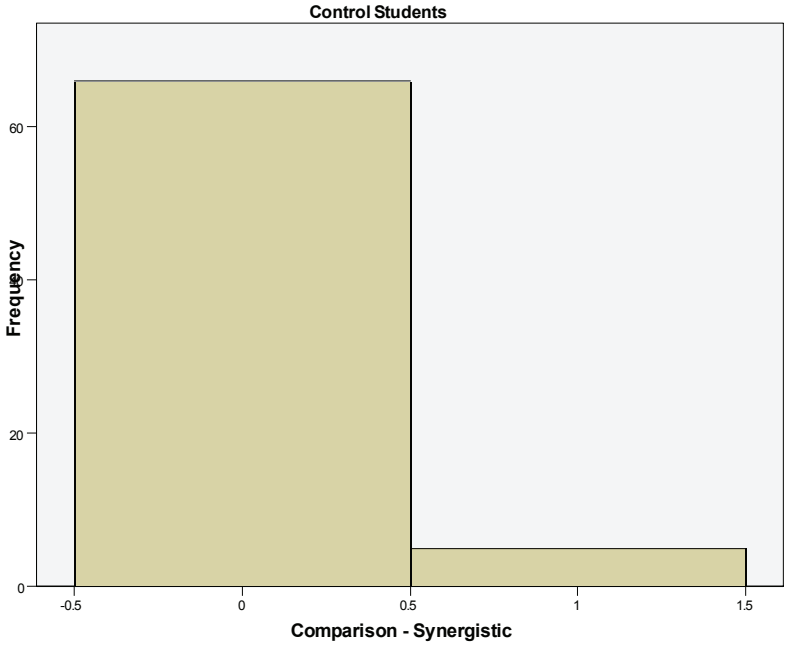


Figure 12: Distribution of control students' use of synergistic Comparison.



Students' Use of Evidence to Support their Assertions during the Poster Interview

Students were asked to provide evidence for their statements as they spoke about the work of art in their individual interview. Statements were coded as either weak or strong (see Appendix E). Looking across all students, regardless of whether they were treatment or control, most used weak evidence (mean=1.44). Evidence was defined as weak if the student offered vague, unspecific, or unclear assertions to support their statements, or if their arguments were "circular." The following examples are representative of weak uses of evidence:

> *I think they're arguing because he's doing like that (control student, 3rd grade)*
>
> *And I think the cloth, the cloth right there, is from China because of the different designs on it (treatment, 4th)*
>
> *He probably goes to find shells, collects shells, 'cause there's a shell on his shirt (treatment, 4th)*
>
> *It looks like they're from the old days because how they look and the clothes they're wearing (control, 5th)*

Evidence was considered strong if the student offered specific, clear, and reasonable support for their statements, most often drawing from the object or situation rather than personal opinion. Following are some representative examples of strong uses of evidence:

> *The guy in the green looks like he's laughing- he has a grin on his face (control student, 3rd grade)*
>
> *But they don't look exactly poor 'cause they have a lot of food on the table (treatment, 5th)*
>
> *And it looks like they're like praying, because she has her eyes closed and they like, they're holding hands (control, 5th)*
>
> *And I think there's a guy, like behind him. You see the little shadow? (treatment, 4th)*

A comparison of treatment and control students' use of evidence showed that treatment students were significantly more likely to support their statements about the artwork with evidence, whether weak or strong, than were control students (weak evidence: Mann-Whitney U=1134.000, p=0.000; strong evidence: Mann-Whitney U=1286.500, p=0.000). (Table 10)

Table 10: Treatment and control students' usage of evidence during the poster interview

Evidence Type	Treatment Group Median (n=64)	Control Group Median (n=71)	Significance
Weak Evidence	1.00	0.00	* *(p=.000)*
Strong Evidence	1.00	0.00	* *(p=.000)*

To what extent did participation in the SPP influence students' critical thinking skills within a social context in the museum?

When given time to explore and converse about works of art in the Rafael gallery in the museum, treatment students again had more to say than did control students. Treatment school students averaged 45 lines per small group "untour," while control students averaged 27 lines (Mann-Whitney U=617.500, p=0.000). The following section of the report describes students' individual performance within a social group in the museum, in terms of 1) use of overall Critical Thinking Skills; 2) use of sub-categories of Critical Thinking Skills; and 3) provision of evidence in support of students' assertions.

Students' Use of Overall Critical Thinking Skills during the "Untour"

As can be seen in Table 11, all students (treatment and control) were much more likely in their social groups to use Observation and Interpretation skills than the other CTS, as was the case for the individual poster interviews. Over half of the students used observational skills 28.5 times during the untour and over half used Interpretation skills 19.0 times. The medians for the other CTS ranged between 1 and 2 uses per skill.

Table 11: Students' usage of overall critical thinking skills during the "untour" (N=116)

Critical Thinking Skills	Median*	Range
Observation	28.50	1 – 118
Interpretation	19.00	0 – 97
Evaluation	2.00	0 – 26
Association	1.00	0 – 14
Problem-Finding	1.00	0 – 11
Comparison	1.00	0 – 8
Flexible Thinking	1.00	0 – 12

*The median is the midpoint of the distribution, meaning that half of the students are below this point and half of the students are above this point. Where the median is 0, it means that half or more of students did not use that critical thinking skill.

Comparisons of treatment and control students yielded results similar to those found in the individual poster interviews. As seen in Table 12, treatment students were significantly more likely than control students to use Observation (Mann-Whitney U=675.000, p=0.000) and Interpretation (Mann-Whitney U=836.000, p=0.000).

Table 12: Treatment and control students' usage of overall critical thinking skills during the "untour"

Critical Thinking Skills	Treatment Group Median (n=56)	Control Group Median (n=60)	Significance
Observation	44.00	20.00	* *(p=.000)*
Interpretation	25.00	13.50	* *(p=.000)*
Evaluation	2.50	2.00	Not significant
Association	1.00	1.00	* *(p=.009)*
Problem-Finding	1.00	1.00	Not significant
Comparison	1.00	0.00	* *(p=.006)*
Flexible Thinking	1.00	0.00	* *(p=.001)*

Not so easily seen in Table 12 is the fact that treatment students were also significantly more likely to use Association (Mann-Whitney U=1222.500, p=0.009), Comparison (Mann-Whitney U=1217.000, p=0.006), and Flexible Thinking (Mann-Whitney U=1103.500, p=0.001) (see Table 12). To see how treatment and control students differed in their use of these three CTS, it is necessary to look beyond the medians of each group in order to examine distributions.

Association. Treatment students used significantly greater numbers of Association than did control students during museum "untours." Although both groups have a median of 1, Figures 13 and 14 below show that many more treatment students used Association in higher frequencies.

Figure13: Distribution of treatment students' use of Association during museum "untours."

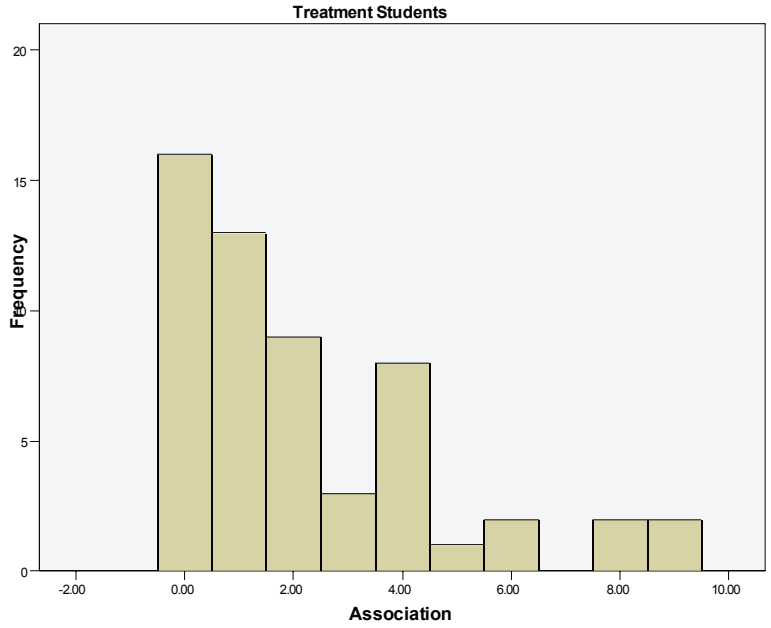
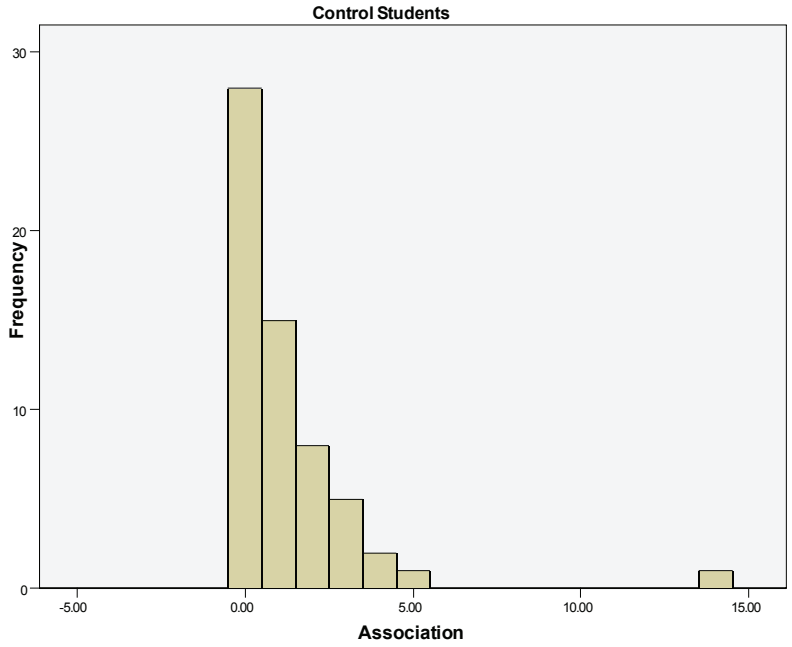


Figure 14: Distribution of control students' use of Association during museum "untours."



Comparison. As with Association, both treatment and control students used Comparison an average of once. The question arises then as to how there can be a significant difference between these two groups. Figures 15 and 16 below answer this question, showing that although both groups have the same median, more treatment students used higher numbers of Comparison.

Figure 15: Distribution of treatment students' use of Comparison during museum "untours."

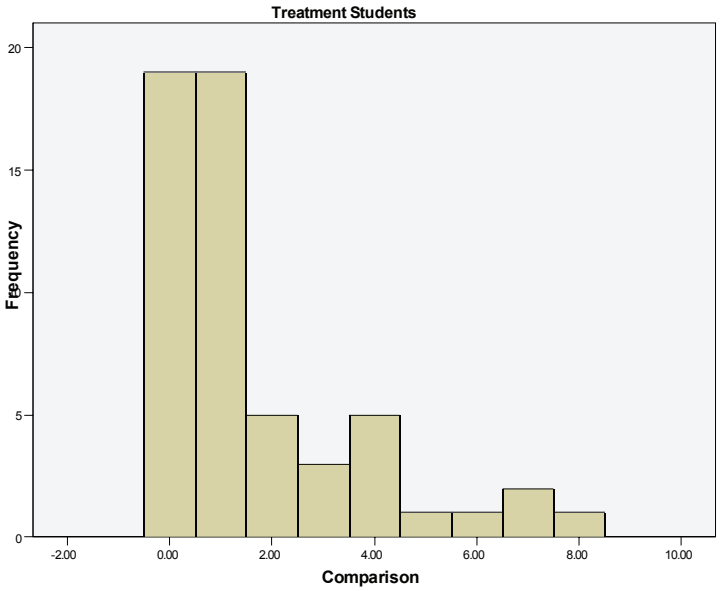
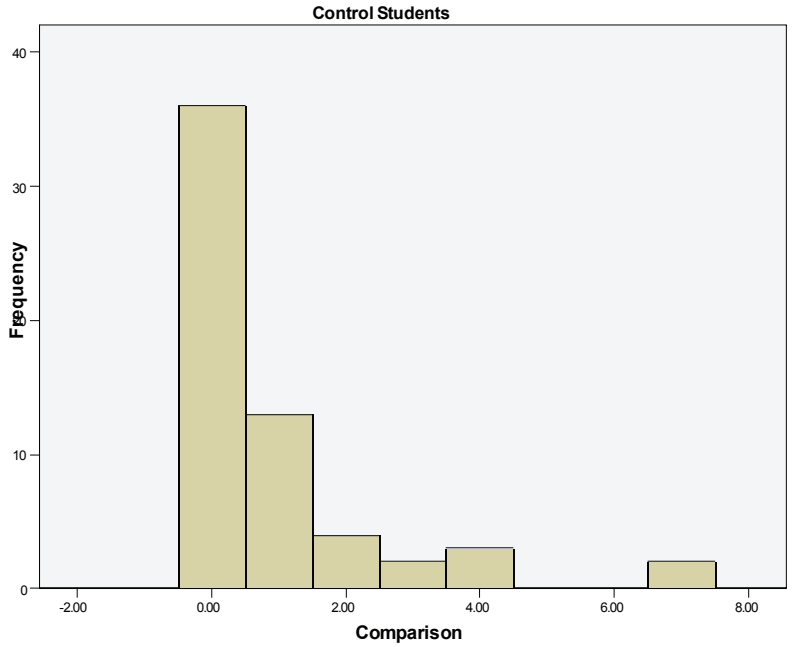


Figure 16: Distribution of control students' use of Comparison during museum "untours."



Flexible Thinking. Treatment students used significantly greater instances of Flexible Thinking during their museum "untour" as compared to control students. Although the median for both groups was 1, Figures 17 and 18 below illustrate that few control students used more than 1-2 instances of this CTS, whereas many treatment students used more than 1-2 instances of Flexible Thinking.

Figure 17: Distribution of treatment students' use of Flexible Thinking during museum "untours"

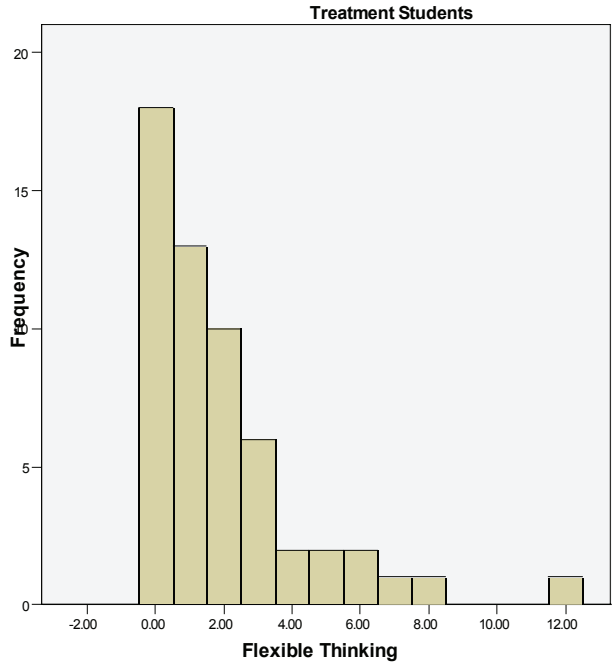
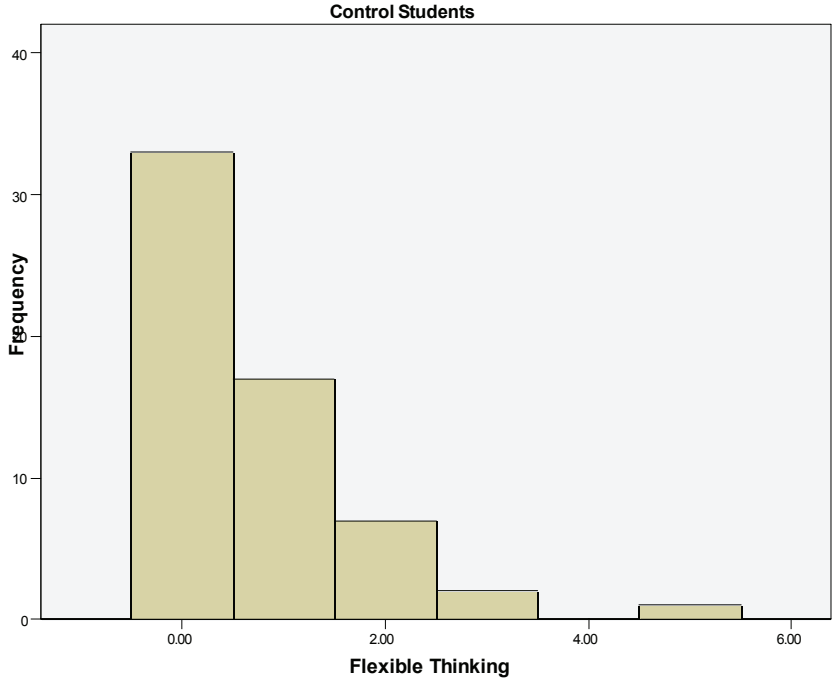


Figure 18: Distribution of control students' use of Flexible Thinking during museum "untours."



Students' Use of Critical Thinking Sub-Categories during the "Untour"

Within the broad groups of CTS, specific sub-categories were identified to more explicitly define how students thought about the works of art in the museum. As shown in Table 13, some sub-categories were used by students more often than others as they talked aloud about objects and works of art in the gallery. As was the case during the poster interviews, both treatment and control students were equally likely to use the Observation CTS (1.3) to note specific details, features or the location of an object in the work (median=10.5). In the gallery space, this included the color, shape, size, or quantity of objects or people; identified particular features or missing features; or situated an object in a physical context. The following examples represent how students used this sub-category:

Look at the design on the red chairs (treatment student, 5th grade)

There's a picture on top (control, 3rd)

It's just like a big candle holder (control, 4th)

It doesn't have any guitar strings (treatment, 5th)

I see a woman with a baby that she's holding (treatment, 3rd)

Table 13: Students' usage of specific critical thinking skills during their "untour" (N=116)

Critical Thinking Skills	Median	Range
Observation (1)		
What something is or is not (1.1)	10.00	0 – 59
Actions (1.2)	3.00	0 – 18
Details, features, locations (1.3)	10.50	0 – 63
What or how something is made (1.4)	0.00	0 – 8
Reading labels (1.5)	2.00	0 – 18
Interpretation (2)		
Use of objects (2.1)	1.00	0 – 13
Implicit Conditions (2.2)	3.00	0 – 16
Identity (2.3)	3.00	0 – 32
Intentions or narrating (2.4)	10.00	0 – 65
Evaluation (3)		
Personal opinion (3.1)	2.00	0 – 4
Merits of the work (3.2)	0.00	0 – 14
Association* (4)	1.00	0 – 14
Problem-Finding (5)		
Isolated	1.00	0 – 11
Synergistic	0.00	0 – 7
Comparison (6)		
Isolated	0.00	0 – 6
Synergistic	0.00	0 – 3
Flexible Thinking* (7)	1.00	0 – 12

* CTS has no sub-categories.

In the museum setting, students were as likely to simply identify what they saw (median=10), as they were to note a specific feature or detail of the artwork. Following are a few typical examples of how students identified what they saw in the gallery (skill 1.1):

Look at the library (treatment student, 3rd grade)

It's a guy (control, 3rd)

There's a baby in this picture (treatment, 5th)

A baby (control, 5th)

I see the sky (treatment, 4th)

That looks like a ukulele (treatment, 5th)

When treatment and control students were compared on the sub-categories, some significant differences emerged (See Table 14). Within the CTS Observation, three of the five sub-categories were used significantly more often by treatment students (1.1: Mann-Whitney U=900.000, p=0.000; 1.2: Mann-Whitney U=1014.500, p=0.000; 1.3: Mann-Whitney U=609.500, p=0.000). The two Observation sub-categories that did not show significant differences (e.g. how a work was made, 1.4, and reading labels, 1.5) were not used very often by either group.

Within the CTS Interpretation, three of the four sub-categories were used significantly more often by the treatment group (2.1: Mann-Whitney U=11327.500, p=0.044; 2.3: Mann-Whitney U=1050.500, p=0.000; 2.4: Mann-Whitney U=772.000, p=0.000). Unlike in the individual poster interviews, however, treatment students talked about how objects were used (2.1) significantly more often than control students did while in the gallery.

Table 14: Treatment and control students' usage of specific critical thinking skills during the "untour"

Critical Thinking Skills	Treatment Group Median (n=56)	Control Group Median (n=60)	Significance
Observation (1)			
What something is or is not (1.1)	14.00	7.00	* *(p=.000)*
Actions (1.2)	4.00	1.00	* *(p=.000)*
Details, features, locations (1.3)	17.50	6.50	* *(p=.000)*
What or how something is made (1.4)	0.00	0.00	Not significant
Reading labels (1.5)	1.50	2.00	Not significant
Interpretation (2)			
Use of objects (2.1)	2.00	1.00	* *(p=.044)*
Implicit Conditions (2.2)	3.50	2.00	Not significant
Identity (2.3)	5.00	3.00	* *(p=.000)*
Intentions or narrating (2.4)	13.50	5.00	* *(p=.000)*
Evaluation (3)			
Personal opinion (3.1)	2.50	2.00	Not significant
Merits of the work (3.2)	0.00	0.00	Not significant
Association* (4)	-	-	-
Problem-Finding (5)			
Isolated	1.00	1.00	Not significant
Synergistic	0.00	0.00	Not significant
Comparison (6)			
Isolated	1.00	0.00	Not significant
Synergistic	0.00	0.00	* *(p=.004)*
Flexible Thinking* (7)	-	-	-

* CTS has no sub-categories.

For the most part, differences between treatment and control are easily seen in Table 14's comparison of medians between the groups. However, in the case of Synergistic Comparison, the significant difference does not show up in the medians, since both group has a median of 0. Figures 19 and 20 below show that more control students tended to not use synergistic Comparison.

Figure 19: Distribution of treatment students' use of synergistic Comparison during museum "untours."

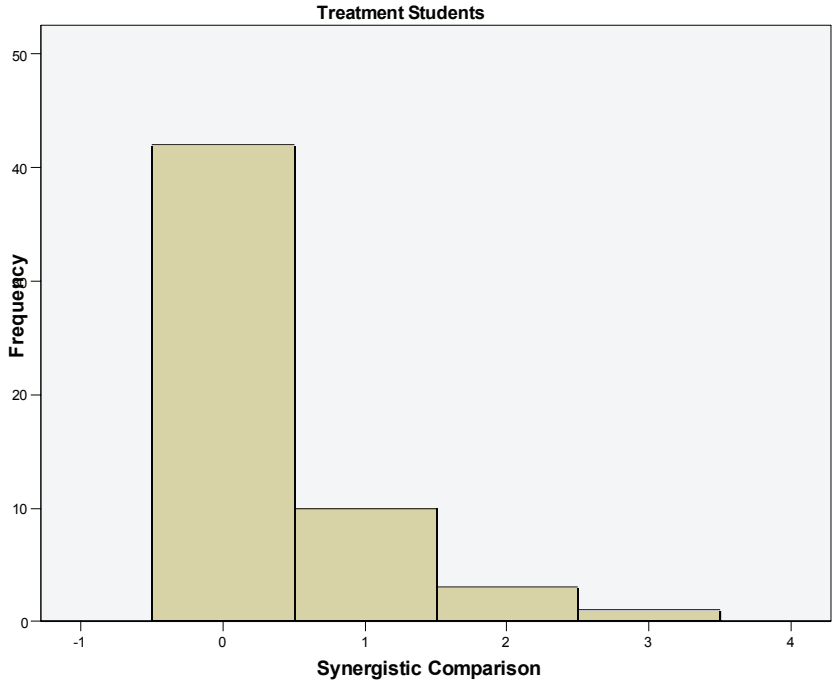
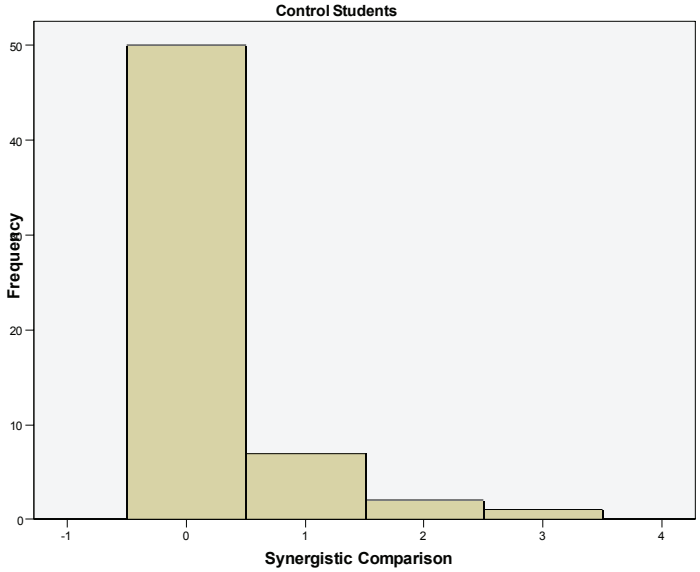


Figure 20: Distribution of control students' use of synergistic Comparison during museum "untours."



Students' Use of Evidence to Support their Assertions during the "Untour"

In the museum, students were asked to provide evidence for their statements. As was the case in the individual poster interview, students were more likely to use weak evidence than strong evidence in the gallery, whether they were in the control or treatment group. The mean usage of weak evidence in interviews was .72 uses per interview and the mean for strong evidence was .66 uses per interview.

The use of evidence to support statements about their assertions about works of art was significantly different between treatment and control groups. Treatment students were significantly more likely than control students to provide both weak and strong evidence as they talked about the works of art in the gallery (weak evidence: Mann-Whitney U=1105.000, p=0.000; strong evidence: Mann-Whitney U=1209.500, p=0.001) (Table 15).

Table 15: Treatment and control students' usage of evidence during the "untour"

Evidence Type	Treatment Group Median (n=56)	Control Group Median (n=60)	Significance
Weak Evidence	1.00	0.00	* *(p=.000)*
Strong Evidence	0.00	0.00	* *(p=.001)*

To what extent did participation in the SPP influence students' critical thinking on standardized tests?

This section of the report presents findings from the 2005-2006 SAT-9 data collected from students; 2005-2006 MCAS data have not yet been analyzed, but will be added to this report in the form of an addendum later this year.

3rd Grade Reading Comprehension

This analysis focuses on the reading comprehension section of the SAT-9 test, which is comprised of five reading sub-scales: specific detail, action/reason/sequence, inference, extended meaning, and critical analysis. When the control and treatment samples for the 3rd grade students were compared, there were no significant differences between the control and treatment samples on any of the reading sub-scales. As Table 16 illustrates, there is very little difference between the weighted means for the control and treatment populations. These differences are too small to be statistically significant.

Table 16: Comparison of 3rd grade treatment and control students on the SAT-9 reading comprehension sub-scales

SAT-9 Sub-scale	Weighted Means		Percent Difference[8]
	Treatment Students	Control Students	
Specific Detail	6.4	6.2	3.1%
Action/Reason/Sequence	4.1	4.0	2.7%
Inference	5.5	5.7	-3.5%
Extended Meaning	6.7	6.5	3.0%
Critical Analysis	2.4	2.2	8.0%

Source: SAT-9 data provided by the Boston Public School System.

4th Grade Reading Comprehension

A comparison of the control and treatment samples for the 4th grade students revealed no significant differences between the control and treatment samples on any of the reading sub-scales. As Table 17 illustrates, there is very little difference between the weighted means for the control and treatment populations. These differences are too small to be statistically significant.

Table 17: Comparison of 4th grade treatment and control students on the SAT-9 reading comprehension sub-scales

SAT-9 Sub-scale	Weighted Means		Percent Difference
	Treatment Students	Control Students	
Specific Detail	6.8	6.2	8.7%
Action/Reason/Sequence	3.7	2.7	27.1%
Inference	9.1	7.9	13.8%
Extended Meaning	8.7	8.0	8.0%
Critical Analysis	5.3	4.9	8.3%

Source: SAT-9 data provided by the Boston Public School System.

5th Grade Reading Comprehension

When the control and treatment samples for the 5th grade students were compared, there were no significant differences between the control and treatment samples on any of the reading sub-scales. As Table 18 illustrates, there is very little difference between the weighted means for the control and treatment populations. These differences are too small to be statistically significant.

[8] Percent difference is calculated by first computing the difference between the two amounts (6.4 – 6.2 = .2) then make a fraction of that difference over the first of the two amounts (.2 / 6.4) = .03125 or 3.1% difference.

Table 18: Comparison of 5th grade treatment and control students on the SAT-9 reading comprehension sub-scales

SAT-9 Sub-scale	Weighted Means		Percent Difference
	Treatment Students	Control Students	
Specific Detail	5.7	5.2	8.3%
Action/Reason/Sequence	4.1	3.4	16.4%
Inference	10.5	9.2	11.9%
Extended Meaning	7.1	6.5	8.5%
Critical Analysis	5.4	5.0	8.6%

Source: SAT-9 data provided by the Boston Public School System.

How does the CTS rubric translate to the larger professional field?

Even though the CTS rubric is a robust, valid, and reliable assessment tool for critical thinking, it is quite cumbersome. It requires substantial training to administer properly. While the value of the rubric in research is not questioned, its value to practitioners is. To make the instrument more user-friendly, we developed a short-form without the sub-categories and tested it with thirteen museum practitioners in six museums across the country (See Appendix F for the Short Form Rubric and Appendix G for an article, Research to Practice: Testing a Tool for Assessing Critical Thinking in Art Museum Programs, submitted to the Journal for Museum Education in February 2007).

The results of this pilot test were encouraging and provide insight into the applicability of such a rubric. The practitioners saw a need to measure critical thinking within their programs, and saw the short-form rubric as a relevant and useful way to do that despite areas that need revision. One area that needs further thinking is the evidence scale and its implementation, as many practitioners had difficulty distinguishing between weak and strong evidence; second, further thought needs to be given to adequate training for practitioners so that they can more confidently code children's comments. Given the limited context provided to these practitioners, and the fact that they are not familiar with data collection methods, it is impressive that they could make sense of the rubric and offer such valuable insights.

Despite the drawbacks in implementing the short-form rubric, it was exciting to see that the exercise successfully stimulated practitioners to think differently about their practice. Clearly the rubric can be a valuable diagnostic and training tool for practitioners as they strengthen the ways that their programs can enhance young people's critical thinking skills.

CONCLUSIONS

There is strong evidence to suggest that the School Partnership Program is quite successful at enhancing students' critical thinking skills in a variety of areas. Equally interesting is that because the design of this study was exceptionally rigorous, we can be reasonably confident that any significant findings are most likely a result of the ISGM School Partnership Program. This is an important point as, like every museum, the ISGM staff underwent the inevitable staff changes and the resulting shifts in the focus of multiple-visit program because of those staff changes. Most museum programs are active, living organisms which, for researchers, constitute a constantly moving target. Still, the core strength of the program revealed itself in every year of research. This section reviews the key findings in the data and concludes with reflections on future research areas that emerged as a result of this study.

Key Findings

The first goal of this research initiative was to "*Develop reliable and valid measures for assessing Critical Thinking Skills (CTS) in and from an art museum that are applicable to a wide range of school/museum educators across the country."*

Developing a responsive and feasible scoring rubric that assessed critical thinking skills related specifically to an art museum multiple visit program was a challenging task during the first two years of this study. The first year was spent in exploratory study to determine what kind of rubric might best reflect student learning in critical thinking. In the second year we reviewed the initial rubric draft developed at the end of Year 1 and revised it to better align with the changing goals of the *SPP* and to create the most valid and reliable measures possible. Reducing something as rich and complex as critical thinking to a handful of categories was a demanding, sometimes frustrating, but ultimately satisfying task for the ISGM education staff as well as the ILI research team. It required us all to articulate what we meant by learning in the *SPP* and to be constantly vigilant that we neither reduced critical thinking skills to meaningless but easily countable bits nor took such a broad perspective that all the richness and intricacy of students' thinking was diluted. Ultimately, by the third year of the research, this rubric proved itself to be a robust, valid, and reliable assessment tool for critical thinking – for students in the ISGM *School Partners Program*. The overall research methodology is quite successful at capturing meaningful data on ways children benefit from this program.

The second part of Goal 1 was to go beyond this specific museum program to create a research methodology and instrument that could yield generalizable data across different school/museum multiple-visit programs. It was not in the scope of this research to replicate the full study in other museum multiple-visit programs, which would be needed to fully generalize results to a larger population. Our approach to the second part of this goal was to focus on making the rubric itself actually useful to museum practitioners.

Reviewing the full rubric and scoring methodology in this light reveals that it is too cumbersome for the average museum practitioner and requires substantial training to administer properly and

score accurately. To make the instrument more user-friendly for practitioners, we developed a short-form without the sub-categories and pilot-tested it with museum educators in other museums. Results from the pilot test strongly suggest that the shortened rubric has great potential for the larger field of museum education. It stimulates museum professionals to be reflective about their own practice, it assists educators in diagnosing the effect of programmatic changes, and it provides them with a highly useful training tool for anyone teaching in the museum.

The second goal for this research was to "*Better understand how an art museum multiple-visit program can facilitate critical thinking amongst a generalizable sample of inner-city elementary students.*"

The data from the performance assessment part of the study consistently showed that students in the ISGM multiple-visit program out-performed a group of comparison students who did not participate in any type of art museum/school program. These findings held true whether students looked at and talked about a poster reproduction or whether they explored a gallery at the museum in company with other students. Students in the ISGM program not only used more critical thinking skills, they used a wider variety of critical thinking categories as measured by the rubric. That all the children in the study used some skills more (Observation and Interpretation) than others, even though the treatment group still used more of the less-used skills (in particular, Association, Comparison, Flexible Thinking, and Evidence), may have two possible explanations.

First, the results may reflect a developmental issue. Most children may not be developmentally able to engage in the other critical thinking skills to the extent that a teenage student might. Certainly the work of Housen (1983, 2002) supports this explanation through her research on aesthetic development. Her theory posits that aesthetic development is not related to age but to one's level of experience in looking at and thinking about art. This theory describes what viewers talk about and categorizes them according to stages of development. Researchers with the Foundation for Critical Thinking, Linda Elder and Richard Paul (1996), also take a development perspective on critical thinking as a skill to be used in all facets of one's life. Their developmental structure is based on *how* children think, such as the unreflective thinkers who are unaware of the role thinking plays in their lives, to the master thinkers who are not only aware of the important role of thinking in their lives but actively seek to be better more cogent thinkers. The fact that the distribution of critical thinking skills in this research, although different in amount, was similar in shape suggests that the developmental theories have merit. The scope and intent of this study did not focus on development as a variable, therefore, there are no data on the effect of age or prior experience on critical thinking, beyond the finding that students who had prior experience in the art museum program engaged in more critical thinking than students who had no prior experience.

Second, the emphasis and focus of a multiple-visit program also might contribute to the ways children engage in critical thinking. The ISGM program was, by the third year, strongly aligned with the Visual Thinking Strategies (VTS) approach[9]. The question "What's going on in this picture?" that the VTS model recommends beginning discussions about art, asks children to

[9] Visual Thinking Strategies is the creation of Abigail Housen and Philip Yenawine. Information about this approach was gathered from the organization's website: www.vue.org.

focus on what they see (Observation) and on what they think is happening (Interpretation) which could explain why treatment students engaged in that skill to a much higher level than did the control students. The model also stipulates that the teacher follows responses to the first question with the question "What do you see that makes you say that?" which requires students to support their assertions with evidence. That treatment students provided more evidence than control students lends credence to the possibility that program focus affects use of specific critical thinking skills. However, the incidence of evidence-giving was, overall, fairly low, especially in comparison with Observation and Interpretation. Students in the SPP are able to give good evidence when in discussion with a skilled teacher but their natural inclination to do so when on their own is quite low. It is conceivable that a multiple-visit program with a focus on making connections, comparing and contrasting, or specifically testing initial interpretations and then practicing changing opinions would result in a different distribution of critical thinking skill use among students. Certainly this position was supported by museum practitioners in the pilot study of the short-form rubric.

That the data from the standardized test did not reveal any differences between treatment and control group was not surprising. Although this measure was required by the U.S. Department of Education, the National Governor's Association Center for Best Practices (2002), in a review or research on the causal relationship between arts experiences and test scores, strongly cautioned against justifying arts education programs solely on the basis of standardized test results. While some studies claim that test results are affected by arts education experiences, just as many find no effect. When an effect is found it is most often associated with regular arts instruction received over three years. Therefore to expect a museum program, that has an average of fifty contact hours over three years in the program, to substantially influence standardized test scores, is unreasonable.

The third goal for this research was to "*Broadly disseminate study results to the museum education profession, and the formal education sector, through channels such as lectures and symposia, publications, website features, and a documentary video.*"

Over the years of the study, the ISGM and ILI have consistently shared the methodology, process of rubric development, and results of each year's study with the larger museum field. This final report, along with the inspiring video that the ISGM made documenting students in the program, will serve as a springboard for continued dissemination to inform the field's understanding of critical thinking in a museum multiple visit program. It is encouraging to have powerful data to support the assertion that such programs can positively influence the quality and quantity of children's critical thinking skills. Further, the most gratifying implication is that museum multiple-visit programs can be extremely effective in working with populations of urban at-risk children, many of whom have few opportunities to enrich their critical thinking through the visual arts or learn how to learn through art and art museums. In addition, the dissemination of the shorter rubric for educators has great potential to influence the larger field.

Reflections

This research provides a solid base for the larger museum field on which to better understand critical thinking skills in an art museum multiple visit program and, as a result, apply these findings to the development of school programs that more consciously and deliberately facilitate critical thinking. As is the case with good art, a good research study often raises as many questions as it answers. Several key questions emerged during the course of this research that can serve as a beginning for a larger professional research agenda on student learning in art museum multiple-visit programs.

Role of age, prior experience, and program focus in critical thinking skill development: This study was not designed to address the effect of developmental differences across students. Yet, it is an intriguing and important question – one that has been studied outside the museum but not within the museum context. Museum practitioners need to better understand the role that age and intellectual development play in the mastery of critical thinking skills in the art museum context. Similarly, research should explore the effect of program focus on students' use of critical thinking skills.

Effect of time-on-task in critical thinking skill development: Another important question is the effect of contact hours on critical thinking skills. While we had this data for the ISGM study, the range of contact hours was so wide and the sample size relatively small that it was impossible to analyze with any degree of confidence. While educators assume that more time spent on-task results in more learning, we do not know how much is sufficient to make a difference in student learning. Are six visits twice as effective as three visits? Is there a difference in learning dependent on *where* most of those visits take place – school or museum?

Effect of regular arts instruction on critical thinking skill development: The students in the control group were selected because they did not have access to in-school sequential art instruction simply because the treatment schools did not offer regular arts instruction. A question that frequently arises after research and evaluation studies on museum/school multiple visit programs centers on the degree to which regular arts instruction might mitigate the effect of the museum program. Would the findings in this study be the same if both groups of students had regular art instruction as part of their general education? A good teacher might focus on students developing skills in looking at and thinking about art and, by virtue of having regular access to student, would have more contact hours with them. It is possible, then that if both treatment and control groups had regular instruction by a skilled teacher that we would see no difference on these critical thinking measures as a result of the museum program. If that was the case then the challenge for educators and researchers would be to identify if and how an art museum experience affects students differently from those who do not have museum experiences. Museum practitioners and school teachers who frequently take students to museums believe there is something unique about the museum experience. Articulating that uniqueness will, we suspect, prove as challenging, frustrating, and satisfying as articulating the critical thinking skills.

A related research area in this category is to look more closely at how students employ evidence for their assertions when guided by a skilled teacher. From the ISGM DVD of students, we see that students can provide sound and reasoned support for their opinions, when facilitated. The study data suggested that the frequency of students giving evidence, while more abundant for treatment than for control students, was still fairly low. More research could better understand the differences between facilitated and un-facilitated critical thinking.

Application of the methodology to other school/museum multi-visit programs: While the samples for the treatment and control groups were randomly selected suggesting that the study results could be generalized to other program, more data is needed in this area. This would require the replication of the full study in one or two other museums with multiple-visit programs with a similar age group. It will also be important that these programs claim to teach children critical thinking. This should not be difficult to identify as a majority of programs in our review of museum multiple-visit programs specifically listed the development of critical thinking skills as the intent of the program (Adams, Luke, & Manuel, 2003).

Usefulness of the short-form rubric with educators: The pilot test of the short-form rubric suggested that the instrument is useful for museum educators on several levels. Testing the instrument more broadly would greatly inform the field as to the ways that the instrument could function. In addition, the instrument might be equally helpful and useful for classroom teachers involved with museum/school partnerships. The more we, as a field, clarify what constitutes learning in the museum, the more likely it will be that the learning happens for students.

References

Adams, M., Luke, J. & Manuel, D. (2003). *Multiple-Visit Museum/School Programs in the Arts: What do students learn?* Technical research report. Annapolis, MD: Institute for Learning Innovation.

Costa, A. & Kallick, B., Eds. (2000). *Assessing and reporting on habits of mind*. Alexandria, VA: Association for Supervision and Curriculum Development.

Elder, L. & Paul, R. (1996) *Critical Thinking Development: A Stage Theory with Implications for Instruction*, At: www.criticalthinking.org.

Housen, A. (2002) Aesthetic thought, critical thinking, and transfer. *Arts and Learning Research Journal*, 18, No. 1.

Housen, A. (1983) *The Eye of the Beholder: Measuring Aesthetic Development*. Unpublished Ed.D. Dissertation. Harvard University.

Institute for Museum & Library Services (2002) *True Needs, True Partners: Museums Serving Schools* 2002 Survey Highlights. Washington, DC: IMLS.

National Governors Association (2002) *Chapter 5 The Impact of Arts Education on Workforce Preparation*. Issue Brief:vEconomic & Technology Policy Studies. Retrieved from http://www.nga.org/cda/files/050102ARTSED.pdf.

Wetterlund, K., & Sayre, S. (2003). *2003 Art Museum Education Programs Survey*. Retrieved from http://www.museum-ed.org.

Appendix A: Year 1 Critical Thinking Skills Rubric