Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: http://www.tandfonline.com/loi/uree20

Measuring Critical Thinking: Results From an Art Museum Field Trip Experiment

Brian Kisida, Daniel H. Bowen & Jay P. Greene

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2016, VOL. 9, NO. S1, 171–187
http://dx.doi.org/10.1080/19345747.2015.1086915

Measuring Critical Thinking: Results From an Art Museum Field Trip Experiment

Brian Kisida[a], Daniel H. Bowen[b], and Jay P. Greene[a]

ABSTRACT
Research shows that participation in school-based arts education has declined over the past decade. A problem for the arts' role in education has been a lack of rigorous scholarship that demonstrates educational benefits. A component of this problem has been a lack of available data. In this study, we use original data collected through a randomized controlled trial to measure the effects of school visits to an art museum. Building on previous work, we find positive effects of art museum visits on students' ability to critically examine a work of art. Importantly, we validate our previous findings with an additional experiment that utilizes a different style of art to assess critical-thinking outcomes, adding extra validity to the assessment instrument and our results. Our research suggests that policymakers should more fully consider the educational benefits of arts education, and scholars should consider broader approaches to measuring student performance in nontested subjects.

KEYWORDS
experimental design
arts education
critical thinking
informal learning

Student achievement data focused on the arts and humanities are particularly rare, and rigorous methodological approaches to the study of arts education are rarer still. Most evaluations of student achievement predominantly focus on outcomes measured by standardized test scores in math and reading (Heckman & Rubenstein, 2001). Most of this work relies on the test score data generated under state accountability systems or other preexisting data sets. This poses a problem for the arts in education. In a policy environment increasingly driven by data and rigorous analytic techniques, unmeasured and understudied subjects face the risk of being deemphasized for failure to demonstrate their value (Gadsden, 2008).

This development is concerning for a number of reasons. First, the efficacy of the mission of educational research depends on the field being driven by academic inquiry, and not operating simply as an extension of state accountability systems. Although there is no question that basic literacy and numeracy are central to education, it is unclear how much of the attention they receive in the research community is simply due to convenience. A lack of available measures of broader components of student achievement at least partially explains their neglected study (Heckman & Rubenstein, 2001). Such a situation is problematic, because emerging research demonstrates that alternative skills are vitally important for

CONTACT Brian Kisida briankisida@gmail.com University of Arkansas, Fayetteville, AR 72701, USA.
[a]University of Arkansas, Fayetteville, Arkansas, USA
[b]Rice University, Houston, Texas, USA
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uree.

determining future life outcomes (e.g., Almlund, Duckworth, Heckman, & Kautz, 2011; Heckman & Rubenstein, 2001; Jackson, 2013; Tough, 2012).

Moreover, the increased emphasis on accountability testing in core subjects has coincided with a notable decline in school-based arts exposure (Gadsden, 2008). A growing body of research is validating the suspected link between educators' increased focus on accountability testing and decreased emphasis on the arts and other nontested subjects (Bassok, Latham, & Rorem, 2016; West, 2007). Additional evidence suggests that the declines are disproportionately affecting disadvantaged students (Chappell & Cahnmann-Taylor, 2013; Rabkin & Hedberg, 2011; Yee, 2014). A recent federal government report found that schools identified as needing improvement under the Federal No Child Left Behind Act and schools with higher percentages of minority students were more likely to experience decreases in the amount of time spent on arts education (Government Accountability Office, 2009). Without research demonstrating the educational benefits of the arts and humanities, practitioners and policymakers who determine where schools focus their resources are ill-equipped to make informed decisions.

Historically, a common method for exposing students to the arts and humanities has been through school-facilitated visits to cultural institutions. Cultural institutions spend more than $2 billion per year on educational activities, and they receive more than 90 million student visits each year from K–12 school groups (National Humanities Alliance, 2012). Yet, similar to school-based arts exposure, there is mounting evidence that school visits to cultural institutions are in decline (AP/*The Cincinnati Inquirer*, 2012; Blair, 2008; Ellerson, 2010; Lewin, 2010; Mehta, 2008; Plummer, 2014). Without evidence demonstrating the potential benefits of arts exposure, there is not a reliable method for policymakers to consider the costs of these declines.

The research we present here helps to address the lack of measured educational outcomes in nontested subjects. First, using original data, we adopt a broader view of educational achievement than what is commonly measured in the study of student outcomes. Second, we expand our consideration of the source of educational enrichment by examining student visits to an art museum. This analysis speaks directly to the policy implications of the increasingly diminished role of the arts in education and the decline in school visits to cultural institutions.

We assess the academic benefits of exposure to the arts using two rounds of original data from a large-scale randomized controlled trial (RCT) involving nearly 8,000 grade 3–12 students assigned by lottery to participate in a facilitated school tour of an art museum. All students in our study completed a follow-up survey, which included a prompt to write essays in response to a work of art that was unfamiliar to them. The essays from the treatment and control groups were coded blindly on a 7-item rubric to assess the students' critical-thinking skills. In a previous study, we examined the first round of data from this project and found that students randomly assigned to visiting an art museum demonstrated stronger critical-thinking skills when analyzing a representational work of art that was unfamiliar to them (Bowen, Greene, & Kisida, 2014). We build on this previous research with a new examination of 3,610 additional students who were part of a second experiment involving a work of abstract art. We find that students who were assigned by lottery to visiting an art museum also demonstrate significantly stronger critical-thinking skills when analyzing an abstract work of art. The pattern of results, however, does not fully align across the two samples.

In the next section we discuss existing research and theory about the potential educational benefits of arts exposure. Next, we describe our sample design, data collection, and the treatment the students received. Then we discuss the critical-thinking skills rubric and our empirical strategy. In the penultimate section we present our results. We conclude with a discussion of our findings, suggestions for future research, and the implications for education research and policy.

Theory and Research on the Educational Benefits of the Arts

Proponents for the inclusion of the arts in education have argued that it helps develop empathy, creativity, and self-expression (Dewey, 1919; Heilig, Cole, & Aguilar, 2010), and serves as a way to strengthen cognitive abilities and critical-thinking skills (Eisner, 2002). A number of studies also claim that exposure to the arts has positive "transfer" effects to core subjects such as math and reading (see, e.g., Baker, 2012; Catterall, Dumais, & Hampden-Thompson, 2012; Deasy, 2002). Critics, however, point out that these studies are typically correlational and unable to demonstrate causal links (Hetland & Winner, 2001; McCarthy, Ondaatje, Zakaras, & Brooks, 2004; Winner & Cooper, 2000). Moreover, the attempts to justify the arts indirectly due to their potential to produce gains in other subjects has been described as self-destructive, and prompted calls to develop theory and gather evidence that demonstrates the direct effects of arts exposure (Hetland & Winner, 2004).

Along these lines, some have suggested that the most immediate effects of the arts on education are those "that pertain to the perception and comprehension of aesthetic features" (Eisner, 1999, p. 147). Similarly, previous research identifies certain "habits of mind" gained by studying art, which include observing, reflecting, envisioning, innovating, stretching and exploring, and engaging and persisting (Hetland, Winner, Veenema, & Sheridan, 2007). Previous studies have also found relationships between arts participation and creativity measures (Luftig, 1994), theory building and reflecting (Heath, 1999), student expressiveness and elaboration (Burton, Horowitz, & Abeles, 2000), tolerance and historical empathy (Greene, Kisida, & Bowen, 2014), attitudes toward cultural consumption (Kisida, Greene, & Bowen, 2014), and critical-thinking skills (Burchenal & Grohe, 2007; Korn, 2007; Lampert, 2006).

Most existing studies, however, have not been able to employ rigorous research designs and have lacked well-developed methods of measuring the types of outcomes theorized to be related to arts exposure. A notable exception resulted from a 2003 evaluation of a school partnership program at the Isabella Stewart Gardner Museum in Boston funded by the U.S. Department of Education's Arts in Education Model Development and Dissemination grant program. The evaluation examined the impact of a curriculum and teaching method, Visual Thinking Strategies (VTS), which was implemented through the critical examination of art. A significant component of the evaluation was the development of a rubric for measuring the kinds of critical-thinking skills theoretically related to learning through the arts (Luke, Stein, Foutz, & Adams, 2007). The final version of the rubric was composed of seven individual critical-thinking skills: observation, interpretation, evaluation, association, problem finding, comparison, and flexible thinking.

In the quasi-experimental evaluation of the program, researchers found that treatment group students demonstrated significantly more instances of these critical-thinking skills (Adams, Foutz, Luke, & Stein, 2007). A separate report found that treatment group students' critical-thinking gains also extended to their writing skills (Desantis, 2009).

In a previous study employing this same rubric, we found that students randomly assigned to visit an art museum demonstrated stronger critical-thinking skills when analyzing a work of representational art, and that the effects were higher for students from smaller towns and students from higher poverty schools (Bowen et al., 2014). We replicate and extend this research by experimentally assessing the critical-thinking outcomes of an additional 3,610 students who examined an abstract work art. As a result, we are able to consider whether the main effects and subgroup effects from our initial experiment are consistent across different forms of art. We also conduct exploratory work that considers how the effects differ between representational and abstract art regarding specific critical-thinking components.

Sample and Data

In November of 2011, the Crystal Bridges Museum of American Art opened in Bentonville, Arkansas. With an endowment greater than $800 million, it is the first major American art museum to open in 50 years (Vogel, 2011). In March 2012, the museum established a program that offered school tours to area students. A substantial portion of the museum's endowment covers expenses related to the school tours, which allows school groups to visit the museum for virtually no cost. The endowment covers admission, transportation, lunch at the museum, substitute teachers, and pre/post-visit curricular materials. Because the establishment of a significant art museum where one did not previously exist was a major event, and the cost of the tours was covered, demand for school tours was much higher than capacity. The museum received applications from 525 school groups representing 38,347 Grade K–12 students during the first two semesters of the program. In the interest of fairness, we conducted a lottery in partnership with the museum to award available slots.

In order to ensure the comparability of the treatment and control groups, we implemented a stratified randomization procedure. The use of a stratified randomization procedure can increase the balance between treatment and control groups while preserving the advantages of random assignment (Schneider, Carnoy, Kilpatraick, Schmidt, & Shavelson, 2007). To ensure that the treatment and control groups were equal on important baseline treatment characteristics, we paired applicant groups with similar demographics (e.g., grade, region, school free- and reduced-lunch status, and school percent minority) and performed isolated lotteries within these pairings. The applicant groups that won each lottery were assigned to the treatment condition, and the corresponding matched applicants that did not win the lottery were assigned to the control group. To incentivize participation in the study, applicants assigned to the control group were guaranteed a spot for the following semester if they participated in data collection.

After the matching process and assignment to treatment, we generated a random number for each applicant pair. This randomly generated number was then used to rank-order the pairs and award spots to the treatment groups until all available spots were filled. As a result, 74 total groups with students in Grades 3–12 were randomly awarded a guided tour of the museum (the treatment groups), while another 74 groups had their tours deferred (the control groups). Forty of the application groups were awarded a tour in the first semester of randomization (spring 2012), while an additional 34 groups were awarded tours in the following semester (fall 2012). Applicant pairings not selected received apologetic letters and encouragement to apply in future rounds.

Trained members of the research team visited the students in their classrooms and administered surveys to both the treatment and its paired control group on average three weeks ($M = 21.8$ days, $SD = 12.1$) after the treatment group's visit to Crystal Bridges. Seven matched pairs that were originally part of the lottery were excluded from the study because of tour cancellations or erroneous application information. Because participation in data collection was packaged as a mandatory component of receiving an immediate or deferred school tour, all of the remaining treatment groups that visited the museum and their matched control groups completed surveys. In total, 67 matched treatment and control group pairs (35 in the spring and 32 in the fall) completed a critical-thinking assessment, representing a total of 134 applicant groups, over 7,500 students, and 111 different schools.

The Treatment

The treatment conditions were consistent across the spring and fall experiments. Before they visited the museum, treatment group teachers were sent a packet containing a 5-minute orientation video for students and teachers to watch. The video emphasized the student-driven nature of the tours, and established that student discussions were a central component of the tour process. Teachers were also provided with information about the themes of the tour they had selected, a sample of three images that the students would see, and discussion questions to ask their students prior to the visit. The questions were intended to introduce students to the types of themes they would learn about and to prepare them for the dialogue-driven nature of the tour. The museum also provided teachers with post-visit materials that included suggestions for classroom activities and factual information about the artworks.

The museum tours were led by paid museum educators, tailored for specific grade levels, and aligned with Common Core Curriculum Standards. During a typical tour, students were placed into small groups of 10 to 15 that focused intensively on four or five paintings or sculptures in the museum. The goal of museum educators was to facilitate an open-ended, student-centered approach to discuss the works of art, encourage a deep level of engagement, and motivate students to seek out their own unique interpretations. When relevant, museum educators supplied sociological and historical information about the art in order to enhance student understanding. Guiding student-driven discussion, however, was the main goal of the museum educators.

Critical-Thinking Assessment

The student surveys we administered contained questions regarding student demographics, art consumption and production, attitudes toward cultural institutions, and knowledge of art. After completing the survey items, students were shown a painting they had not previously seen—a relatively unknown work of art that was not part of the museum's collection. In the spring semester, students were asked to analyze Bo Bartlett's *The Box* (Figure 1). As a result of piloting multiple images with students prior to data collection, this example of representational art was chosen because students seemed to respond to the younger subjects in the painting, and the somewhat ambiguous nature of the painting provided a lot of opportunities for students to provide unique interpretations. Additionally, the painting has a number of objects that students were able to incorporate into their analysis.

In the following fall semester, we chose a more abstract painting that was also unfamiliar to the students—Marsden Hartley's *Eight Bells Folly: Memorial to Hart Crane* (Figure 2). By selecting a more abstract painting, we are able to broaden our application of the critical-

Figure 1. Bo Bartlett's *The Box*, 2002, oil on linen, 82 × 100 in.

thinking rubric and determine if the museum experience improves the students' ability to critically examine different styles of art.

Once presented with a copy of the painting, students were given exactly five minutes to write an essay in response to the following two questions: (a) What is going on in this painting? (b) What do you see that makes you think that? These questions are often used as prompts by art educators when facilitating student-driven discussions about art and are part of the VTS learning approach. The first prompt asks students to construct a narrative about the work, while the second question "subtly asks the viewer to supply evidence to back up his answer to the first question" (Housen, 2001, p. 7).

Figure 2. Marsden Hartley, *Eight Bells Folly: Memorial to Hart Crane,* 1933, oil on canvas, 30 5/8 × 39 3/8 in., Collection of the Frederick R. Weisman Art Museum at the University of Minnesota, Minneapolis. Gift of Ione and Hudson D. Walker. 1961.4.

Some essays were largely observational, while other essays provided deeper and more complex interpretations of the paintings. For example, in the following passage about *The Box* from a 10th-grade girl, the student analyzes the objects placed throughout the painting and provides observations, interpretations, and associations:

> I believe the children are reminiscing on the loss of their father. The look on the children's faces is very mournful. In the open bucket you can see things that would be sent home if a loved one was lost in war. The Popeye doll seems like he would represent the father's strength. There is a wedding photo, probably for remembrance and what looks like communion, which represents religion that maybe the family was close to. Also, to me, the empty chair in the foreground shows where the father would be sitting if he were present.

In this passage written about *Eight Bells Folly,* an 11th-grade boy makes numerous observations and interpretations and assigns an overall narrative to the painting:

> I see the ship as a life making decisions. I see that this ship is in a storm, it is trying to navigate through the storms of life. The waves are crashing high above the small ship's sails, the wind is blowing, and it is a dark night. The eyes represent loved ones who are concerned with the outcome of the ship, and are there to give counsel when needed.

Four researchers independently coded the student's written responses using Luke et al.'s (2007) critical-thinking skills checklist, and then tallied the number of observations, interpretations, evaluations, associations, instances of problem finding, comparisons, and instances of flexible thinking. In order to minimize bias, coders were not made aware of any student characteristics, including whether they were in the treatment or control group. To test for inter-rater reliability, the researchers coded a set of overlapping essays—750 in the

Table 1. Descriptive statistics and inter-coder reliability for critical thinking items.

	Spring sample: *The Box* (Representational art)			Fall sample: *Eight Bells Folly* (Abstract art)		
Item	Average (*SD*)	Percent agreement	Weighted kappa	Average (*SD*)	Percent agreement	Weighted kappa
Composite (sum of 7)	8.16 (3.85)	99.4	0.84	8.69 (4.28)	99.3	0.89
Observation: identifying something; what something is; what is happening; locations; counts	3.97 (2.40)	99.4	0.78	6.08 (3.25)	99.3	0.91
Interpretation: relationships; feelings related to work of art; ascribing meaning	3.90 (2.35)	98.7	0.56	2.42 (2.00)	98.6	0.70
Evaluation: personal preferences and perceived merits of the work of art	0.02 (0.18)	99.6	0.40	0.07 (0.31)	98.9	0.48
Association: connecting art with previous knowledge or experience	0.06 (0.25)	96.8	0.37	0.03 (0.18)	99.5	0.53
Problem Finding: noting information needed to form a conclusion; requesting information	0.01 (0.12)	99.7	0.13	0.03 (0.22)	99.7	0.77
Comparison: noticing relationships; noticing patterns; noting similarities or differences	0.02 (0.15)	99.1	0.69	0.00 (0.05)	—	—
Flexible Thinking: seeing things from multiple perspectives; revising thoughts	0.17 (0.43)	98.7	0.84	0.05 (0.24)	98.8	.09

spring sample and 250 in the fall sample. Descriptive statistics for both samples, as well as the percent agreement and the more conservative Cohen's weighted kappa for the overlapping items (Cohen, 1968) are provided in Table 1.

Similar to the federally funded study that developed the rubric, observations and interpretations were the most common elements in the students' essays. Notably, observations were much more likely for the abstract work of art, while observations and interpretations both factored heavily in the scores for the more literal painting. It is possible that students found the abstract image more difficult to relate to, and thus harder to interpret.

The combination of all seven items, which is the dependent variable used in our main outcome analyses, displays a high rate of reliability between coders in both samples (weighted kappa = .84, .89, respectively). When the items of the critical-thinking scale are examined separately, most of the individual items also show high levels of intercoder reliability. The item "problem finding" is an exception in the spring sample, which is explained by the fact that occurrences in student essays were particularly rare. The same is true for "flexible thinking" in the fall sample. Additionally, instances of "comparisons" were too rare among the 250 overlapping essays in the fall sample for adequate reliability tests.

Empirical Strategy

Comparisons between the treatment and control groups on key variables show that the stratified randomization procedure largely achieved the goal of producing comparable groups, although as is often the case in randomized controlled trials, they are not perfectly identical

Table 2. Treatment/control balance on key characteristics.

	Spring sample				Fall sample			
Characteristic	Treatment (n = 1,720)	Control (n = 1,890)	Difference	p value	Treatment (n = 1,747)	Control (n = 2,211)	Difference	p value
Percent females	53.37	52.28	1.10	0.24	50.37	50.47	−0.10	0.86
Percent White	61.45	58.94	2.51	0.18	54.67	60.74	−6.08	0.47
Percent Hispanic	20.58	20.26	0.32*	0.01	19.12	17.87	1.25	0.79
Percent Black	3.02	4.92	−1.90	0.25	2.69	2.99	−0.29	0.05
Percent other	14.94	15.87	−0.93	0.70	23.53	18.41	5.12	0.16
Average grade	6.15 (2.56)	6.21 (2.71)	−0.07	0.12	5.76 (2.07)	5.65 (2.18)	0.11	0.31
Cultural activities	0.77 (1.00)	0.81 (1.04)	−0.05	0.67	1.04 (1.16)	0.92 (1.08)	0.12**	0.00
School FRL	50.94 (21.67)	53.00 (21.83)	−2.06	0.24	58.00 (23.73)	57.87 (20.40)	0.13	0.81
School size (100s)	6.83 (4.36)	8.10 (5.39)	−1.27	0.08	6.19 (2.57)	6.03 (2.73)	1.65	0.43
Town size (1000s)	37.94 (28.27)	55.19 (36.58)	−17.26*	0.04	38.91 (30.37)	30.16 (29.22)	8.75	0.85

**$p < .01$, *$p < .05$, two-tailed.

Note. School FRL, school size, and town size are measured at the applicant group level, other demographic variables are measured at the student level. School FRL = percentage of students receiving free or reduced-price lunch. Cultural activities are the sum of self-reported previous cultural activities students participated in outside of school, including dance lessons, music lessons, art classes, and theater participation. The reported p value is from the coefficient on the treatment indicator when each covariate is regressed on the treatment indicator and the matched pair dummies (as in the base model for the impact model). Treatment and control group means and differences are actual. Standard deviations of continuous variables are shown in parentheses. A joint F test from a model regressing the treatment indicator on the full list of covariates failed to reject the null hypothesis that the effects of the covariates are jointly equal to zero (spring sample p value = 0.43, fall sample p value = 0.12).

(Table 2). Raw means and differences are displayed for individual student-level, school-level, and community-level characteristics. The displayed *p* values are from the coefficient on the treatment indicator when each covariate is regressed on the treatment variable and the matched pair indicators. Three of the differences are statistically significant. In the spring sample, the treatment group is slightly more likely to identify as Hispanic. The magnitude of the difference, however, is not substantial. The spring sample treatment group is also more likely to come from a town with a smaller population. Some slight imbalance on these measures makes sense, as pairs were matched based on a school's overall school percent minority, region, school percent FRL, and grade. Percent Hispanic and town population were not explicitly incorporated into the creation of matched pairs. In the spring sample, the only significant difference between the treatment and control groups is a slight difference in the number of previously reported cultural activities. For both samples, a joint *f* test failed to reject the null hypothesis that the effects of the covariates on treatment are jointly equal to zero (spring sample *p* value = 0.43, fall sample *p* value = 0.12).

Because randomization generated comparable treatment and control groups, we can use straightforward analytic techniques to estimate the impacts of the school tour of an art museum. In its simplest form, this technique could estimate simple mean differences using the following equation for outcome *CTS*, the standardized critical-thinking score, of student *i* in matched pair *m*:

$$\mathrm{CT}S_{\mathrm{im}} = \alpha + \beta_1 \mathrm{Treat}_i + \mathbf{Match_i}\boldsymbol{\beta_2} + \mathbf{Grade_i}\boldsymbol{\beta_3} + \varepsilon_{\mathrm{im}} \tag{1}$$

The indicator variable $Treat_{\mathrm{i}}$ is equal to 1 if the student was in the treatment group randomly assigned to visit the museum for a school tour and is equal to 0 otherwise. Because we used a stratified randomization procedure within matched applicant group pairs, $\boldsymbol{Match_{\mathrm{im}}}$ is also included in the model as a vector of dummy variables that have the statistical effect of estimating within, as opposed to across, matched pairs. Moreover, we include dummy variables to control for student grade level because in some cases matched pairs were composed of adjacent grades within the same school. Finally, $\varepsilon_{\mathrm{im}}$ is a stochastic error term clustered at the applicant group level to account for the spatial correlation of students nested within applicant groups.

Although proper randomization generated comparable groups, they are not perfectly identical. The basic regression model may be improved by adding additional controls for observable characteristics to increase the precision of the estimated impact by accounting for minor differences between the treatment and control groups. Moreover, by adding observable characteristics to the regression model, we can examine the relationship between these characteristics and the outcome measures. This yields the following equation to be estimated:

$$\mathrm{CT}S_{\mathrm{ims}} = \alpha + \beta_1 \mathrm{Treat}_i + \mathbf{Match_i}\beta_2 + \boldsymbol{X}_{\mathbf{i}}\beta_3 + \boldsymbol{Z}_{\mathbf{s}}\beta_4 + \varepsilon_{\mathrm{ims}} \tag{2}$$

where $\boldsymbol{X_i}$ is a vector of student characteristics and $\boldsymbol{Z_s}$ is a vector of school and community characteristics. Important student characteristics are gender and grade level. We include gender in our models as a binary measure equal to 1 if the student is female, and grade level is controlled for using a series of dummy variables. Additionally, we are able to include a measure of students' prior cultural activities, which could potentially moderate the effects of the museum tour. Students reported if, outside of their school, they had ever taken dance lessons (21% responded yes), music lessons (28% responded yes), art classes (20% responded yes), or participated in theater (23% responded yes). We sum the number of affirmative responses to these questions into a baseline measure of cultural activities. School characteristics are school level percent free- and reduced-price lunch (FRL) and school size. Finally, we use the population of the children's town of residence as an indicator of rural status.

In addition to overall impacts, we also test for heterogeneous effects across particular types of students. We test for heterogeneous effects by modifying equation (2) to include interactions between the binary treatment variable and student and school characteristics. For our analysis, we explore potential interaction effects using gender, our baseline measure of cultural activities, school FRL levels, school size, and town size.[1] These measures serve as potential indicators of students' disadvantaged status and students' previous exposure to the arts and cultural activities, which may moderate responses to the treatment. Our measure of previous cultural activities measures this directly. Additionally, students from higher FRL schools, students in smaller schools, and students in smaller towns have likely had fewer opportunities to participate in enriching cultural activities.

Results

Regression estimates for the spring sample who analyzed a work of representational art are shown in Table 3. In the most parsimonious model (column 1), the impact of the treatment is 10% of a standard deviation. Adding student, school, and community characteristics does not substantially change the overall effect size, which is to be expected with experimental data (column 2). Descriptively, female students, on average, score higher on the critical-thinking measure than do male students. Our measure of previous cultural activities is also positive and significant. Each additional cultural activity a student reported having previously done (dance classes, music classes, etc.) was associated with a 9% of a standard deviation increase in our critical-thinking measure. The size of a student's town, which serves as a measure of how urban or rural a student's community is, is significant and negatively related to the outcome.

When we interact the treatment variable with the other covariates (columns 3–7), one of the five interaction terms are significant. Students attending schools with higher proportions of FRL-eligible students have critical-thinking outcomes that are significantly higher as a result of the treatment. The interaction of school size falls just short of statistical significance ($p = .05$). Because the inclusion of each interaction term represents a distinct way of comparing the treatment and control groups, we use Benjamini & Hochberg's (1995) correction for multiple comparisons to test for the potential that the significant interaction of school FRL is a false discovery due to chance. The significant treatment interaction for students at higher FRL schools survives this adjustment.

The overall effect for the fall sample, who analyzed a more abstract work of art, is similar in magnitude to the spring results (Table 4). The effect size in the parsimonious model (column 1) is 13% of a standard deviation, while the effect size in the model that includes covariates is 8% of a standard deviation (column 2). Similar to the spring results, female students

[1] For continuous variables (school size, town size, school FRL levels, and cultural activities), our tests for heterogeneity impose a linear structure on the relationship between the treatment and the treatment impact. To investigate nonlinearity, we also examined heterogeneity by collapsing the continuous variables into roughly equal subgroups. The results were largely consistent, except in the case of FRL levels, which showed more consistent evidence of heterogeneity when reducing model dependence. As a result, when examining heterogeneity along FRL levels, we collapse FRL into a binary measure of higher FRL ($>50\%$ FRL $= 1$, $\leq 50\% = 0$).

Table 3. Regression estimates of treatment effects on critical thinking about representational art.

	(1) Effect size (SE)	(2) Effect size (SE)	(3) Effect size (SE)	(4) Effect size (SE)	(5) Effect size (SE)	(6) Effect size (SE)	(7) Effect size (SE)
Treatment	0.10**	0.10**	0.07	0.12**	0.24**	−0.02	0.19
	(0.05)	(0.03)	(0.04)	(0.04)	(0.07)	(0.05)	(0.11)
Female		0.38**	0.35**	0.38**	0.38**	0.38**	0.38**
		(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Cultural activities		0.09**	0.09**	0.10**	0.09**	0.09**	0.09**
		(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
School size		0.61	0.62	0.60	0.62	0.49	0.65
		(0.35)	(0.35)	(0.35)	(0.33)	(0.26)	(0.36)
School FRL		0.68	0.68	0.65	0.43	0.05	0.31
		(0.72)	(0.72)	(0.71)	(0.62)	(0.13)	(0.71)
Town size		−0.03*	−0.03*	−0.03*	−0.02	−0.04**	−0.03
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Treat × female			0.06				
			(0.05)				
Treat × activities				−0.02			
				(0.03)			
Treat × school size					−0.17		
					(0.09)		
Treat × high FRL						0.22**	
						(0.07)	
Treat × town size							−0.02
							(0.02)
Observations	3,610	3,610	3,610	3,610	3,610	3,610	3,610
R^2	0.22	0.29	0.29	0.29	0.29	0.29	0.29

**$p < .01$, *$p < .05$, two-tailed.

Note. Estimates are obtained from ordinary least squares regression models with robust standard errors clustered by applicant group. All models control for grade level and lottery pair. Effect sizes are in terms of standard deviation units. School FRL = percentage of students receiving free or reduced-price lunch, except in column 6, where it is a coded as a binary measure of lower (≤50%) or higher (≤50%) proportions of students who are FRL-eligible. Cultural activities are the sum of self-reported previous cultural activities students participated in outside of school, including dance lessons, music lessons, art classes, and theater participation. School size is expressed in 1,000s. Town size is expressed in 10,000s.

score higher, as do students who have had more experience with cultural activities. In this case, students from larger towns score higher.

The effects across subgroups, however, are less clear than the spring results (columns 3–7). When examining a more representational work of art, the treatment had a larger effect on students at higher FRL schools. In the fall sample, however, this interaction falls just short of statistical significance ($p = .05$). That is, when analyzing a more abstract work of art, we cannot be as confident of differential effects across schools with higher FRL levels as a result of the treatment.

It is possible that the different pattern of heterogonous findings between the two experiments is related to issues of statistical power. It could be the case that the second experiment had sufficient power to detect an overall effect similar to the first experiment, but lacked the power to detect similar heterogeneity. In terms of magnitude, the coefficients on the higher FRL school interaction terms appear nearly identical across both experiments. Pooling the data from both experiments and estimating the triple interaction of treatment, higher FRL, and the outcome image confirms this. The triple interaction is not significant, indicating that the magnitude of the differences between higher and lower FRL schools is not distinguishable across the two samples. This suggests that a lack of power in the second experiment may be the issue.

Table 4. Regression estimates of treatment effects on student critical thinking about abstract art.

	(1) Effect size (*SE*)	(2) Effect size (*SE*)	(3) Effect size (*SE*)	(4) Effect size (*SE*)	(5) Effect size (*SE*)	(6) Effect size (*SE*)	(7) Effect size (*SE*)
Treatment	0.13 *	0.08*	0.09	0.09	0.18*	−0.06	0.05
	(0.05)	(0.04)	(0.05)	(0.05)	(0.09)	(0.09)	(0.08)
Female		0.31**	0.32**	0.31**	0.31**	0.31	0.32**
		(0.04)	(0.05)	(0.04)	(0.04)	(0.04)**	(0.04)
Cultural activities		0.07**	0.07**	0.07**	0.07**	0.06**	0.07**
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
School size		−0.06	−0.06	−0.06	−0.00	−0.03	−0.03
		(0.13)	(0.13)	(0.13)	(0.13)	(0.11)	(0.12)
School FRL		0.55	0.56	0.55	0.58	0.06	0.63*
		(0.34)	(0.34)	(0.34)	(0.36)	(0.13)	(0.34)
Town size		0.07**	0.07**	0.07**	0.07**	0.08**	0.07**
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Treat × female			−0.01				
			(0.07)				
Treat × activities				−0.01			
				(0.03)			
Treat × school size					−0.17		
					(0.12)		
Treat × high FRL						0.22	
						(0.11)	
Treat × town size							0.01
							(0.01)
Observations	3,958	3,958	3,958	3,958	3,958	3,958	3,958
R-squared	0.12	0.16	0.16	0.16	0.16	0.16	0.16

$^{**}p < .01$, $^{*}p < .05$, two-tailed.

Note. Estimates are obtained from ordinary least squares regression models with robust standard errors clustered by applicant group. All models control for grade level and lottery pair. Effect sizes are in terms of standard deviation units. School FRL = percentage of students receiving free or reduced-price lunch, except in column 6, where it is a coded as a binary measure of lower (≤50%) or higher (≤50%) proportions of students who are FRL-eligible. Cultural activities are the sum of self-reported previous cultural activities students participated in outside of school, including dance lessons, music lessons, art classes, and theater participation. School size is expressed in 1,000s. Town size is expressed in 10,000s.

A second reason for weaker evidence of heterogeneity in the fall sample could be due to the differences in the chosen images. We know that descriptively, students tended to make more observations than interpretations when analyzing *Eight Bells Folly*, while observations and interpretations were nearly evenly represented when analyzing *The Box* (Table 2). To shed some additional light on this aspect, we examined impacts on observation and interpretation separately within each experiment. Across both samples, the effect sizes for students at higher FRL schools are larger for observations and interpretations. However, in the spring sample they are only significantly larger for observations ($p = .00$), and in the fall sample they are only significantly larger for interpretations ($p = .00$). Thus, though the descriptive findings suggest that abstract art generates fewer instances of interpretations, the isolated results indicate that interpretations are the main source of heterogeneity. Due to the inherent noisiness in looking at these items in isolation, however, it is difficult to draw firm conclusions.

Discussion

Students have traditionally received exposure to the arts from their schools and through visits to cultural institutions. In both cases, however, growing evidence suggests the amount of exposure has been declining. Moreover, this decline is occurring without an adequate consideration of the academic benefits that these experiences might provide. Amid a lack of data and rigorous research approaches, the costs of reduced exposure to the arts are unknown to parents and policymakers. In our analysis, we find that a relatively modest amount of arts exposure produced modest but significant effects. Students were briefly exposed to curricular materials in their classrooms, and they spent a day of their schooling at an art museum with museum educators. For many of these students, this was their first school-facilitated visit to an art museum. Because Crystal Bridges is the first major art museum to be built within a reasonable travel distance, schools in the area had previously been unable to provide this experience. In this environment, even a minimal intervention produced significant changes in the students' ability to think critically about a work of art they had not seen previously. Our two samples analyzed very different works of art, yet our overall results are consistent. Because museum visits were randomly assigned, we can be particularly confident that the museum exposure received by the treatment group caused the effects.

When analyzing a more representational work of art, we find that the treatment effect is greater for students attending schools with higher proportions of FRL-eligible students. For the students who analyzed a more abstract work of art, however, the evidence of differential effects was less pronounced. Across both samples, female students and students with higher previous levels of cultural activities demonstrated a significant advantage when asked to critically examine a work of art, though there was no differential treatment effect across those particular variables.

It is not entirely clear why the effects are more concentrated among students at poorer schools when examining a representational image. These students may have had an easier time relating to the image of *The Box* after they were exposed to the treatment. The painting depicts two children in a realistic setting, surrounded by literal representations of objects. Examining the more abstract *Eight Bells Folly* may have been more difficult, resulting in modest gains across the full sample that were not as concentrated among students at poorer schools. Our analysis also suggests, however, that the lack of clear evidence of differential effects when analyzing abstract art may be attributable to study power. The coefficient on the higher-FRL interaction term was similar in magnitude across both samples, but fell slightly short of statistical significance for students who analyzed the abstract image. Additionally, an exploratory examination of key components of our critical-thinking outcome suggests that disadvantaged students may receive a greater benefit in the number of observations when examining literal art, and a greater benefit in the number of interpretations when examining abstract art. Future research should further examine the issue of heterogeneous effects on specific critical-thinking outcomes regarding different styles of art.

There are important limitations to this study. It is important to note that the main source of the variation in our critical-thinking outcome is generated by observations and interpretations. It may be an overreach to conclude that these two items are sufficient to demonstrate improved critical thinking about art generally. Although others in the field have theorized that observation and interpretation "may often serve as building blocks for other skills, such as comparison and flexible thinking," this is an area where future research is needed (Adams et al., 2007).

Additionally, because our study examines an area where few cultural institutions exist, our findings may only generalize to students with little prior exposure to such experiences. It is possible that students living in areas with more cultural opportunities would not experience similar benefits, although even in culturally rich areas it is likely that disadvantaged students lack access. Additionally, we were only able to assess students a short time after the experience.

Future research could evaluate whether the benefits of an educational arts experience endure over a longer period of time. Moreover, this research does not establish which components of the art museum experience were essential for increases in critical-thinking skills, or if these same effects could be generated from school-based arts exposure. Finally, data limitations prevent us from directly testing to determine if there are spillover effects in other academic subjects. Winner and Cooper (2000), however, suggest that enhancements in critical thinking produced by arts experiences may not be limited to the arts. They suggest that skills such as observation, critical thinking, and problem solving could transfer to other academic areas.

Rigorous study of the arts, humanities, and other untested subjects is woefully lacking in education policy research. With noted declines in these subjects coinciding with a policy environment increasingly driven by data and quantitative research, more empirical studies of their educational benefits are needed. Such efforts will often require the generation of original data, as state accountability and administrative data sets are typically insufficient for examining subjects beyond reading, math, and science. Advocates make numerous claims about the benefits of the arts and humanities—increased student engagement, increased social responsibility, increased creativity, increased empathy, and increased tolerance, to name a few. Such outcomes seem paramount and fundamental to the mission of education. Yet, surprisingly, few of these claims have been empirically examined with rigorous research designs.

Our findings have important policy implications. To the extent that academic research influences policymakers, it is crucial that policymakers receive information about the broad spectrum of educational benefits available to students. Here, we have established that an arts experience can have a significant impact on critical-thinking skills. This suggests that there are measurable, negative consequences when the arts are reduced in schools. Our results also suggest that in some instances, disadvantaged students may reap the greatest benefits from arts exposure facilitated by their school. Because disadvantaged students receive fewer arts experiences outside of school, public education plays a crucial role in providing those students with access to art. With this and additional research, policymakers and educators may be able to make better informed decisions about where and how to concentrate school resources that extend beyond the task of maximizing performance in core subjects.

Acknowledgments

The authors gratefully acknowledge Anne Kraybill of the Crystal Bridges Museum of American Art and Michelle Grohe of the Isabella Stewart Gardner Museum for their valuable advice and assistance. We are also extremely grateful to University of Arkansas research assistants Josiah Foster, Charles Denny, and Kimberly Torres.

ARTICLE HISTORY

Received 24 November 2014
Revised 13 June 2015
Accepted 22 August 2015

EDITORS

This article was reviewed and accepted under the editorship of Carol McDonald Connor and Spyros Konstantopoulos.

References

Adams, M., Foutz, S., Luke, J., & Stein, J. (2007). *Thinking through art: Isabella Stewart Gardner Museum School Partnership Program summary of final research results*. Annapolis, MD: Institute for Learning Innovation. Retrieved from http://www.gardnermuseum.org/microsites/tta/links/Year_3_Report.pdf

Almlund, M., Duckworth, A. L., Heckman, J. L., & Kautz, T. D. (2011). *Personality psychology and economics* (NBER Working Paper No. 16822). Retrieved from http://www.nber.org/papers/w16822

AP/*The Cincinnati Enquirer*. (2012, October 13). School field trips in decline amid standardized testing. *The Columbus Dispatch*. Retrieved from http://www.dispatch.com/content/stories/local/2012/10/13/school-field-trips-in-decline-amid-standardized-testing.html

Baker, R. A., Jr. (2012). The effects of high-stakes testing policy on arts education. *Arts Education Policy Review*, *113*, 17–25.

Bassok, D., Latham, S. & Rorem, A. (2016). Is kindergarten the new first grade? *AERA Open*, *1*(4), 1–31.

Benjamini, Y., & Hochberg, Y. (1995). Controlling for the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*(1), 289–300.

Blair, E. (2008, December 22). Museum field trips tailored to teach to the test. *National Public Radio*. Retrieved from http://www.npr.org/templates/story/story.php?storyId=98593843

Bowen, D. H., Greene, J. P., & Kisida, B. (2014). Learning to think critically: A visual art experiment. *Educational Researcher*, *42*(1), 37–44.

Burchenal, M., & Grohe, M. (2007). Thinking through art: Transforming museum curriculum. *The Journal of Museum Education*, *32*(2), 111–122.

Burton, J. M., Horowitz, R., & Abeles, H. (2000). Learning in and through the arts: The question of transfer. *Studies in Art Education: A Journal of Issues and Research*, *41*(3), 228–257.

Catterall, J. S., Dumais, S. A., & Hampden-Thompson, G. (2012). *The arts and achievement in at-risk youth: Findings from four longitudinal studies* (Research Report No. 55). Washington, DC: National Endowment for the Arts.

Chappell, S. V., & Cahnmann-Taylor, M. (2013). No child left with crayons: The imperative of arts-based education and research with language "minority" and other minoritized communities. *Review of Research in Education*, *37*(1), 243–268.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychologocial Bulletin*, *70*(4), 213–220.

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

Desantis, K. (2009). *Report to the education department of the Isabella Stewart Gardner Museum on the 8th grade School Partnership Program Visual Thinking Strategies Adaptation 2008–2009*. Retrieved from http://www.vtshome.org/research/research-studies

Dewey, J. (1919). Imagination and expression. *Teachers College Bulletin*, *10*(10), 7–15.

Eisner, E.W. (1999). Does experience in the arts boost academic achievement? *The Clearing House*, *72*(3), 143–149.

Eisner, E. W. (2002). *The arts and the creation of mind*. New Haven, CT: Yale University Press.

Ellerson, N. M. (2010). A cliff hanger: How America's public schools continue to feel the impact of the economic downturn. *American Association of School Administrators*, Retrieved from http://www.aasa.org/uploadedFiles/Policy_and_Advocacy/files/CliffHangerFINAL%281%29.pdf

Gadsden, V. L. (2008). The arts and education: Knowledge generation, pedagogy, and the discourse of learning. *Review of Research in Education*, *32*(1), 29–61.

Government Accountability Office. (2009). *Access to arts education: Inclusion of additional questions in education's planned research would help explain why instruction time has decreased for some students*. Retrieved from http://www.gao.gov/assets/290/286601.pdf

Greene, J. P., Kisida, B., & Bowen, D. H. (2014). The educational value of field trips. *Education Next*, *14*(1), 78–86.

Heath, S. B. (1999). Imaginative actuality: Learning in the arts during the nonschool hours. In E. Fiske (Ed.), *Champions of change: The impact of the arts on learning* (pp. 19–34). Washington, DC: The Arts Education Partnership and the President's Committee on the Arts and the Humanities.
Heckman, J. J., & Rubenstein, Y. (2001). The importance of noncognitive skills: Lessons from GED testing. *The American Economic Review*, *91*(2), 145–149.
Heilig, J. V., Cole, H., & Aguilar, A. (2010). From Dewey to No Child Left Behind: The evolution and devolution of public arts education. *Arts Education Policy Review*, *111*(4), 136–145.
Hetland, L., & Winner, E. (2001). The arts and academic achievement: What the evidence shows. *Arts Education Policy Review*, *102*(5), 3–6.
Hetland, L., & Winner, E. (2004). Cognitive transfer from arts education to non-arts outcomes: Research evidence and policy implications, In E. Eisner & M. Day (Eds.), *Handbook on research and policy in art education* (pp. 135–162). Mahwah, NJ: Lawrence Erlbaum Associates.
Hetland, L., Winner, E., Veenema, S., & Sheridan, K. (2007). *Studio thinking: The real benefits of arts education*. New York, NY: Teachers College Press.
Housen, A. (2001). Voices of viewers: Iterative research, theory, and practice. *Arts and Learning Research Journal*, *17*(1), 1–12.
Jackson, C. (2013). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina* (NBER Working Paper No. 18624). Retrieved from http://www.nber.org/papers/w18624
Kisida, B., Greene, J. P., & Bowen, D. H. (2014). Creating cultural consumers: The dynamics of cultural capital acquisition. *Sociology of Education*, *87*(4), 281–295.
Korn, R. (2007). *Teaching literacy through art: Final report—Synthesis of 2004–05 and 2005–06 studies*. Alexandria, VA: Randi Korn and Associates.
Lampert, N. (2006). Critical-thinking dispositions as an outcome of art education. *Studies in Art Education*, *47*(3), 215–228.
Lewin, T. (2010, April 21). Museums take their lessons to the schools. *The New York Times*. Retrieved from http://www.nytimes.com/2010/04/22/us/22fieldtrips.html?pagewanted=all
Luftig, R. (1994). *The schooled mind: Do the arts make a difference?* Hamilton, OH: Fitton Arts Center.
Luke, J. L., Stein, J., Foutz, S., & Adams, M. (2007). Research to practice: Testing a tool for assessing critical thinking in art museum programs. *Journal of Museum Education*, *32*(2), 123–136.
McCarthy, K. F., Ondaatje, E. H., Zakaras, L., & Brooks, A. (2004). *Gifts of the muse: Reframing the debate about the benefits of the arts*. Santa Monica, CA: RAND Corporation.
Mehta, S. (2008, May 19). Schools can't spare time or dimes for field trips. *Los Angeles Times*. Retrieved from http://www.latimes.com/local/la-me-fieldtrips19-2008may19,0,473819,full.story#axzz2zGK0YU9n
National Humanities Alliance. (2012). *Public witness testimony on National Endowment for the Humanities FY 2013 funding* (Hearing before the House Interior Appropriations Subcommittee, House of Representatives, 113th Cong. [testimony of James Grossman]). Retrieved from http://www.nhalliance.org/bm~doc/testimony_grossman.pdf
Plummer, M. (2014, October 10). Long Beach museum looks to buck trend of declining student field trips. *Southern California Public Radio*. Retrieved from http://www.scpr.org/blogs/education/2014/10/10/17383/long-beach-museum-looks-to-buck-trend-of-declining/
Rabkin, N., & Hedberg, E. C. (2011). *Arts education in America: What the declines mean for arts participation* (Research Report No. 52). Washington, DC: National Endowment for the Arts.
Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (Report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.
Tough, P. (2012). *How children succeed: Grit, curiosity, and the hidden power of character*. Boston, MA: Houghton Mifflin Harcourt.
Vogel, C. (2011, June 16). A billionaire's eye for art shapes her singular museum. *The New York Times*. Retrieved from http://www.nytimes.com/2011/06/17/arts/design/alice-walton-on-her-crystal-bridges-museum-of-american-art.html?pagewanted=1&_r=1&ref=arts&

West, M. (2007). Testing, learning, and teaching: The effects of test-based accountability on student achievement and instructional time in core academic subjects. In C. E. Finn, Jr. & D. Ravitch (Eds.), *Beyond the basics: Achieving a liberal education for all children* (pp. 45–61). Washington, DC: The Fordham Institute.

Winner, E., & Cooper, M. (2000). Mute those claims: No evidence (yet) for a causal link between arts study and academic achievement. *Journal of Aesthetic Education*, *34*(3/4), 11–75.

Yee, V. (2014, April 7). Arts education lacking in low-income areas of New York City, report says. *The New York Times*. Retrieved from http://www.nytimes.com/2014/04/07/nyregion/arts-education-lacking-in-low-income-areas-of-new-york-city-report-says.html