UNIVERSITY OF ARKANSAS | College of Education & Health Professions
Education Reform

WORKING PAPER SERIES

An Experimental Evaluation of Arts Field Trips

Heidi H. Erickson*
Kennesaw State University

Angela R. Watson
Johns Hopkins University

Jay P. Greene
University of Arkansas

Last Revised September 2, 2020

EDRE Working Paper 2020-03

* Corresponding author. hholme11@kennesaw.edu

Acknowledgements: We thank The Woodruff Arts Center for their partnership on this project along with the National Endowment for the Arts for the grant support that made this research possible. We also thank Laura Florick for managing data collection and for research assistance. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the views of The Woodruff Arts Center, The National Endowment for the Arts, Kennesaw State University, Johns Hopkins University, or the University of Arkansas.

This paper was supported in part by a grant award from the Research Labs program at the National Endowment for the Arts (Grant #: 17-3800-7016).

Abstract

This paper presents results of a multi-visit, longitudinal experiment on the academic and social-emotional effects of arts-based field trips. We randomly assign fourth and fifth grade students to receive arts-based field trips throughout the school year or to serve as a control. Treatment students express greater tolerance for people with different opinions and a desire to consume arts. Additionally, treatment students have fewer behavioral infractions, attend school more frequently, score higher on their end-of-grade exams, and receive higher course grades. Effects are strongest when students enter middle school. We find no effect on students' desire to participate in the arts, empathy, or social perspective taking.

I. Introduction

There is a substantial literature on unintended consequences of test-based accountability polices. Previous studies find that schools facing accountability pressure, at times, narrow the curriculum (Hout & Elliott, 2011; Stecher, 2002), direct resources away from non-tested subjects (West, 2007), or artificially boost or manipulate test scores (Cullen & Reback, 2006; Figlio & Getzler, 2006; Figlio & Winicki, 2005; Jacob & Levitt, 2003). Another, although less studied, unintended consequence is a decline in the number of field trips students attend. For generations, K-12 students across America have loaded onto buses and headed off on field trips. Field trips offer students new and diverse experiences in a larger world than they may have access to otherwise. In recent decades, institutions such as arts venues, science museums, and zoos have reported a decline in field trip attendance (McCord & Ellerson, 2009). Teachers and students also report a decline in school sponsored field trips, particularly for minority students in low academically performing schools (Government Accountability Office, 2009; Keiper et al., 2009). This decline also likely affects families with limited resources more than middle-class families with flexible resources as they are more likely to take their children to cultural institutions outside of school field trips (Kornrich, 2016). Under pressure to improve students' math and reading test scores, schools have reconsidered the costs and benefits of traditional educational field trips and have opted for increased classroom instruction (Gadsden, 2008; Rabkin & Hedberg, 2011). While many educators maintain that field trips have value not captured by common measures of learning such as test scores (Student & Youth Travel Association, 2016), district and school administrators face pressure to maximize easily measured metrics of learning.

Despite the century long tradition of school field trips, there is limited evidence on the extent to which there are educational and social emotional benefits of this practice. There is

evidence that single-visit field trips to culturally enriching institutions boost educational outcomes such as social-emotional learning (SEL) (Greene, Kisida, & Bowen, 2014; Greene et al., 2015; Greene et al., 2018; Kisida, Goodwin, & Bowen, 2020; RK & Associates, 2018). We expand this literature by conducting, to our knowledge, the first-ever longitudinal, multi-visit field trip experiment. It is possible that multiple field trips over a period of years could have different effects on student outcomes relative to one-year, single-visit field trips. We randomly assign fourth and fifth grade students in fifteen elementary schools in a large urban school district to receive three arts-based field trips throughout the school year or to serve as a control group. We observe students in the first year of treatment and continue following them as they enter middle school, even after they stop receiving treatment. In this paper, we estimate the causal effect of attending three arts-based field trips in one year, six arts-based field trips over two years, and the effect up to two years following treatment on students' academic performance, school engagement, and social-emotional skill acquisition.

Our current study adds to the existing literature on the effects of arts-based field trips in four primary ways. First, we use an experimental design that allows us to capture the causal effects on students from attending multiple arts-based field trips. Second, where most of the previous literature focuses on the effects from attending one field trip, treatment students in this study attend three different arts field trips: an art museum, a live theater performance, and a symphony. Third, treatment students not only receive three field trips in one year, but a subsample of students receive two doses of treatment for a total of six field trips over two years. Fourth, this study takes place in a large urban city, and the participating schools consist primarily of students of color who are from economically disadvantaged backgrounds. Fifth, this study is

the first longitudinal experiment allowing us to estimate the cumulative and persistent effects, if any, students experience from arts exposure years after receiving treatment.

We focus on the effects of arts-based field trips. Arts field trips not only provide students the opportunity of attending museums and theaters, but they connect students to a larger world outside that of their own schools and neighborhoods by exposing students to different people, places, and ideas. Such exposure may help students develop social emotional skills such as tolerance, empathy, and social perspective taking. Additionally, exposing students to different art forms can increase their desire to consume the arts in the future (Greene et al., 2018; Greene, Kisida, & Bowen, 2014), which is particularly important for arts institutions.

While field trips can expose students to a broader world and provide a unique learning environment, in theory, it is possible that taking students away from traditional classroom instruction multiple times throughout the school year may harm student test scores. On the other hand, arts exposure for students is associated with modest academic gains (Ludwig, Boyle, & Lindsay, 2017; Jægar & Møllegarrd, 2017; Ruppert, 2006). However, it seems unlikely that three arts-based field trips will significantly improve or harm students' academic performance, particularly on math and reading test scores.

Our findings show significant educational and school engagement benefits for students who attend multiple arts-based field trips. We find that treatment students exhibit higher levels of school engagement as well as increased tolerance and conscientiousness compared to control students. Surprisingly, we find that treatment students perform significantly better on their end-of-grade standardized tests and receive higher course grades than control students. These effects appear strongest in years following treatment. These findings have significant implications for

educators and policy makers as they allocate resources, including students' time, and consider accountability polices.

II. Previous Literature

Despite the educational tradition of fieldtrips, limited rigorous research exists on the effects of such activities on students. In this section, we summarize the empirical literature evaluating the outcomes of field trips and arts education. We group the literature based on research design, including both non-experimental and experimental studies, and by social-emotional and academic outcomes.

A. Non-Experimental Studies

Research results suggest that students experience social-emotional and academic benefits from exposure to arts instruction in school. A meta-analysis of drama-based pedagogy finds overall increased academic achievement as well as favorable attitudes toward school for participating students (Lee et al., 2015). Similarly, using a matching design, researchers find that students have increased levels of empathy and theory of mind shortly after exposure to drama activities (Goldstein & Winner, 2012). Theory of mind is defined as the ability to understand that people have differing emotions and beliefs and is closely related to the notion of social perspective taking (SPT) (Gehlbach, Brinkworth, & Wang, 2012) we use in our study.

There is also evidence that single-visit, arts-based field trips benefit students' social emotional abilities. A recent study, using a quasi-experimental design, of single-visit art museum field trips finds that students who attend the field trip experience increases in critical thinking, creative thinking, and human connection (RK & Associates, 2018). Human connection, a related construct to SPT and empathy, is defined as an awareness or sense of connection to others and the self. Additionally, this study compares the outcomes of students who attend the museum

field trip to the outcomes of students who receive a similar arts program in a classroom instead of in the museum. The authors find that the in-gallery experience appears to be more impactful than simply seeing and discussing identical art content at school (RK & Associates, 2018).

The benefits to students from arts exposures may be affected by the amount and consistency of the experiences. Painter, Lacoe, and Williams (2015), using a matching design, evaluated the School in the Park program, a museum-based educational program for low income-students in San Diego. The authors find that participating students show small positive gains on math and ELA test scores as well as reduced absences and suspensions. Students who participate in the program for an extended period in elementary school exhibit benefits into high school and are more likely to take AP courses, score higher on the SAT, graduate from high school, and enroll in college.

Furthermore, longitudinal studies of long-term exposure to the arts also find positive correlations between arts exposure and academic outcomes (Ruppert, 2006). Jægar and Møllegarrd (2017), comparing identical twins, find that children who frequent museums, theaters, and musical performances when they are younger also perform better in school when they are teenagers. Notably, a meta-analysis on the effects of student achievement from arts integration programs finds a four-percentile-point increase in student academic achievement (Ludwig, Boyle, & Lindsay, 2017). While a four-percentile-point point increase reflects modest academic gains, the authors warn against causal interpretation as only one of the studies in the meta-analysis was able to establish a causal connection between arts activities and academic performance.

B. Experimental Studies

There is a growing, yet still limited, body of literature on the causal effects of cultural field trips, arts integration and specifically, arts-related field trips for students. Recent studies find academic, school engagement, and social emotional learning benefits from arts integration programs. A study of a districtwide arts enrichment program where, due to budget constraints, schools are randomly chosen to participate, shows positive outcomes on students' compassion for others, school engagement, as well as increased standardized test scores (Bowen & Kisida, 2019). In another study, students in schools that are randomly assigned to participate in a theater-based program on state history demonstrate increased empathy as well as increased content knowledge and interest in the arts (Kisida, Goodwin & Bowen, 2020).

There is also experimental evidence of similar positive social emotional and academic outcomes from attending single-visit, arts-based field trips. Greene, Kisida, and Bowen (2014) evaluate the effects of a single visit to an art museum and find that students who tour an art museum demonstrate a greater desire to consume art in the future and actually visit the same art museum on their own following the field trip (also see Kisida, Greene, & Bowen, 2014). In addition, treatment students demonstrate increased levels of critical thinking skills, as well as increased tolerance, content knowledge, and historical empathy (Bowen, Greene, & Kisida, 2014). Further, these benefits, measured a few weeks after the intervention, appear stronger for students from low socioeconomic backgrounds.

In a similar study evaluating the effects of attending a field trip to see live theater performances, treatment students demonstrate higher levels of tolerance, social perspective taking, and evidence of increasing desire to consume theater in the future (Greene et al., 2015; Greene et al., 2018). Greene et al. (2018) adds a second treatment condition wherein some

students receive a field trip to a live theater performance, some receive a field trip to see a movie of the same play, and the control group remains at school and receives neither the play nor the movie treatment. Students who view the live theater performance demonstrate higher levels of tolerance, social perspective taking, and content knowledge compared to the students who viewed a movie of the same play.

We add to the literature on the effects of arts-based field trips by using an experimental design to identify the causal effect of these activities on students. Additionally, we estimate the effects, if any, of attending multiple field trips throughout the school year, the compounding effect of attending arts-based field trips two school years in a row, and the persistent effect up to two years following treatment.

III. Empirical Approach

A. Description of the Treatment

In partnership with The Woodruff Arts Center in Atlanta, Georgia and a large urban school district, we randomly assign fourth and fifth grade classes within fifteen elementary schools to receive a field trip to each of the three Woodruff arts partners, the Alliance Theatre, the Atlanta Symphony Orchestra, and the High Museum of Art, or to serve as a control group.[1] The Woodruff Arts Center also provides one day of professional development for teachers in treatment grades where teachers experience the content of the field trips before the school year begins. Treatment students attend the three field trips throughout the course of the school year. The Woodruff provided all field trips free of charge to the participating schools.

[1] The fifteen elementary schools were selected by The Woodruff Arts Center to participate in the study and are not necessarily representative of the larger school district.

The three field trips, all part of the existing educational programming at each venue, are carefully designed for elementary students and cultural relevancy.[2] The hour-long Alliance Theatre performance is designed for children and families, is performed by a professional cast, and is of the highest artistic quality. A trained volunteer docent leads the hour-long High Museum of Art's program featuring several works of art followed by an hour-long hands-on studio experience led by a teaching artist. Finally, the Atlanta Symphony Orchestra fills their 1,700-seat facility for an hour-long concert with a full symphony performing music carefully selected for younger audiences and accented with large-screen video descriptions and images.

Control group students receive "business as usual" during the school year. Absent the three field trips provided by The Woodruff, students typically receive one field trip during the year to a location in Atlanta. The one field trip control students attend could be to one of The Woodruff arts partners or other institutions such as the Jimmy Carter Presidential Library, The Chic-fil-A College Football Hall of Fame, or the Georgia Aquarium.[3] Given that control students receive one field trip throughout the year, we estimate the effect of attending three arts-based field trips compared to attending one fieldtrip in a year.

B. Sample and Randomization

The data contains three cohorts of students. In the first year of the study, school year 2016-17, the first cohort consists of fourth and fifth grade students at four participating elementary schools. In the second year of the study, school year 2017-18, we add a second cohort of students which consists of new fourth grade students at the four original schools along

[2] Details of the field trips were provided by the Woodruff Arts Center as well as our research teams' observations of the field trips. For more information about The Woodruff and the three arts partners see https://www.woodruffcenter.org/.

[3] This is not a comprehensive list of field trips that students in the control group attended, as the specific field trip varied by school and grade.

with fourth and fifth grade students at six additional elementary schools. In year three, school year 2018-19, we add our final cohort of students which includes new fourth grade students at the six schools added in year two along with fourth and fifth grade students at an additional five new elementary schools. In total, our sample includes fifteen elementary schools and just over 2,100 students. The fifteen elementary schools are geographically near each other and feed into three middle schools.

It is logistically difficult for schools to take a mix of fourth and fifth grade students from different classes and schools on three field trips throughout the year. To minimize the administrative burden on the schools and create minimal disruption to normal school schedules, we randomly assign the fourth or fifth grades within each elementary school to create our treatment and control groups. Fourth and fifth grade students in the same schools are likely to be very similar to each other. Arts-related field trips are unlikely to affect fourth grade students in a significantly different way than fifth grade students in the same school.

Through the randomization process, we ensure that we have a balance of fourth and fifth grades that are assigned to the treatment and control groups. In the first year of the study with four participating schools, two of the four schools have fourth grade receive treatment and fifth grade serve as a control group, while the other two schools have fifth grade receive treatment and fourth grade serve as a control group. In the second year, three of the six new schools had fourth grade receive treatment and fifth grade serve as control with the other three schools having the opposite treatment and control assignments. In the third year, three of the five schools had fourth grade receive treatment and fifth grade serve as a control with the other two schools having the opposite assignments.

Students retain their original treatment assignment over the three years as they advance to the next grade level. Table 1 shows treatment assignment by each school and cohort in the third year of the intervention. Treatment students who were in fourth grade in their first year remained treatment students in the next year receiving an additional dose of treatment, three more field trips, in fifth grade. Treatment students who were in fifth grade in their first year remained treatment students in their second year but did not receive an additional dose of treatment in sixth grade.

Within a given elementary school, if fourth grade was assigned treatment in the first year, then in the second year, the new cohort of fourth grade students were control students as the previous fourth grade treatment students were now in fifth grade. Conversely, if fourth grade was assigned control in the first year, then in the second year, the new cohort of fourth grade students were treatment students as the previous fourth grade control students were now in fifth grade. This process of assigning treatment and control groups ensures that we can always compare treatment and control students within the same schools.

By rolling out the study over three years, we are able to estimate the effect for students receiving treatment in one year, receiving treatment for two consecutive years, the effect one year following treatment, and the effect two years following treatment. For example, cohort one treatment students, who are in seventh grade in the third year, entered the study when they were in fifth grade (Table 1). These students received treatment in fifth grade but did not in sixth or seventh grade, so in the second year of the study we consider them one-year post treatment and in the third year we consider them as two-years post treatment. Cohort one treatment students who are in sixth grade in the third year entered the study when they were in fourth grade. These students received treatment in fourth and fifth grade but did not receive treatment in sixth grade;

as such, we consider them as receiving one dose of treatment in the first year, double treatment in the second year, and as one-year post treatment in the third year. Therefore, in year three, we estimate the effect of three field trips in one year, six field trips in two years, and the effect up to two years following treatment.[4]

(Table 1 about here)

C. Data and Outcome Measures

We use two sources of data. First, student surveys were collected before and after treatment. Following randomization, our research team surveyed all students at the beginning of the school year in students' first year of the study to collect demographic and pre-treatment measures of the outcomes. We then administered follow up surveys at the end of the year for post-treatment measures. Students in our sample, on average, perform below grade level on the Georgia Milestone end-of-grade ELA exam; therefore, to help students complete the survey, our research team read the survey aloud while students filled in their answers. We collected survey data for students' first year in the program. As such, in our analysis we only look at the effect of survey outcome measures following one year of treatment.

The survey included similar constructs that have been used in previous research to measure students desire to participate and consume the arts[5] (Greene et al., 2018; Greene, Kisida, & Bowen, 2014), tolerance for different people (Bowen & Kisida, 2019; Greene, et al., 2018;

[4] Cohort one treatment students who are in sixth grade in year three, received treatment in both fourth and fifth grade. As such, in year three, these students are one-year post a double dose of treatment. We do not estimate the effect of one-year post one treatment and one-year post two treatments because there is a limited number of students who fall into these subcategories. We only estimate the effect of one-year post treatment regardless of if students received one or two dosages of the treatment in previous years.

[5] Our survey included separate constructs for art participation and consumption for each art institution, an art museum, theater, and symphony. We combine the three subcategories into one overall art consumption scale and one art participation scale. Refer to Appendix A for the complete constructs.

Greene, Kisida, & Bowen, 2014), political tolerance (Peterson, Campbell, & West, 2001), social perspective taking (Gehlbach, 2004; Gehlbach et al., 2008; Gehlbach, Brinkworth, & Wang, 2012; Greene et al., 2018), empathy (Greene, Kisida, & Bowen, 2014; Bowen & Kisida, 2019), and school engagement. Appendix A contains the specific survey questions. The Cronbach's alphas for tolerance, political tolerance, and school engagement were below 0.6 suggesting low reliability, and, as such, we do not report results for these constructs. We believe the low Cronbach's alphas are in part due to the difficulty students had in understanding and completing the survey, especially on questions about tolerance that included more advanced vocabulary. Despite reading the survey aloud, it was apparent to our research team that many students struggled to accurately complete sections of the survey containing difficult vocabulary. We present results from one question from the larger tolerance construct where students marked how much they agreed or disagreed on a five-point scale with the statement, "I believe people can have different opinions about the same thing" because we believed students could more easily grasp the meaning of this item. This question is part of a larger construct that has been used in similar research to measures students' tolerance for different people (Bowen & Kisida, 2019; Greene, et al., 2018; Greene, Kisida, & Bowen, 2014). However, the results on tolerance presented in this paper should be interpreted cautiously given that it represents only one question from a larger scale.

We also construct measures of survey effort including item non-response (Hitt, Trivitt, & Cheng, 2016; Zamarro et al., 2016) and careless answering (Hitt, 2015; Zamarro et al., 2016). These effort measures have been used as proxies for students' conscientiousness and motivation in completing school tasks. Item non-response is simply the proportion of survey items a student leaves blank. Careless answering captures how consistent students answer survey questions

within a survey construct. When constructing our careless answering measure, we only use survey constructs that have sufficiently high Cronbach's alphas: art consumptions (0.91), art participation (0.85), and SPT (0.76).

For our second data source, we use administrative data provided by the school district for all participating students. We received students' baseline data for the year prior to entering the study as well each subsequent year's data. As such, in our analyses using administrative data, we estimate the effect of treatment in one year, treatment two years in a row, and the effect one and two years following treatment.

The administrative data includes a robust set of student characteristics including demographics, students with disability (SWD) designations, and limited English proficiency (LEP) indicators. For our outcome measures we use a combined standardized score for students' English Language Arts (ELA) and math tests scores on the Georgia Milestone end-of-grade exams, course grades, the number of behavioral infractions a student receives in a year, and the proportion of time a student is absent in the year.[6]

Including administrative data greatly enhances our analysis. First, it provides robust data on student characteristics that we do not get from student surveys including baseline measures for the outcomes. Second, even if students move schools following treatment, if they remain in the large school district, we still receive their data. Third, administrative data contains many behavioral measures that are not affected by the potential self-report biases and measurement

[6] The Georgia Milestones end-of-grade exams are used as a significant part of the state's school accountability program and have both norm-referenced and criterion-referenced items. Students in grades three through eight are required to take both the math and ELA exam. Students in fifth through eighth grade also take a science and social studies exam. As not all student in our sample take the science and social studies exams, we only focus on math and ELA scores. We standardize all exam scores within grade level by year.

error of the student survey constructs. Despite the poor survey measures for school engagement, we can estimate the treatment effect on school engagement as proxied by school attendance and disciplinary infractions. We believe these are good proxies because if a child enjoys school, they are more likely to be engaged, attend more often, and act-out less. Additionally, we have two measures of academic achievement using the Georgia Milestones end-of-grade exam scores as well as course grades.

D. Identification Strategy

We use an experimental design to estimate the causal impact of the arts-based field trips on student outcomes. An experimental design provides the best potential to capture causal impacts as it accounts for selection bias which could be created by schools or classes selecting to attend field trips (Mosteller & Boruch, 2002; Pirog et al., 2009; Rossi, Lipsey, & Freeman, 2004). Given the randomization of treatment, we use a straightforward analytic approach to estimate the intent to treat (ITT) effect of the intervention on students' academic, school engagement, and social emotional outcomes.[7] We estimate the treatment effects using the following equation where *i* denotes student and *s* denotes school:

$$Y_{is} = \delta_0 + \delta_1 OneTreat_{is} + \delta_2 TwoTreat_{is} + \delta_3 OneYearPost_{is} + \delta_4 TwoYearsPost_{is} + \beta Baseline_i + X_i\tau + \theta_s + \alpha_i + \varepsilon_{is}$$

- Y_{is} is the given outcome of interest.
- $OneTreat_{is}$ equals one if a student received one dosage of treatment the first year the student entered the study.

[7] We do not estimate any treatment on the treated effect as we do not have data on which students actually attended the various field trips.

- $TwoTreat_{is}$ equals one if a student received a second dose of treatment in the second year the student was in the study.
- $OneYearPost_{is}$ equals one if a student received treatment in the prior year.
- $TwoYearsPost_{is}$ equals one if a student received treatment two years prior.
- $Baseline_i$ is the baseline or pre-treatment measure of the given outcome.[8]
- X_i is a vector of student characteristics that includes baseline test scores as well as indicators for students' gender, SWD designation, whether the student is in middle school, and for the cohort in which the student entered the study.
- θ_s is a vector of fixed effects for each school. Including θ_s means our approach effectively compares treatment and control students in the same school.
- α_i are student random effects, which allows for correlation between a student's error over multiple years of the program.
- ε_{is} is the error term clustered at the teacher level.

The estimated causal treatment effect of receiving one dose of treatment is δ_1, δ_2 represents the effect of receiving two doses of treatment in two consecutive years, δ_3 represents the effect one year following treatment, and δ_4 represents the effect two years following treatment.

E. Baseline Equivalence and Attrition

Experimental designs rely on randomization to create similar treatment and control groups. While we cannot know if our treatment and control groups are similar on unobservable characteristics, looking at baseline equivalence of observable characteristics gives some evidence if randomization worked as expected. Our treatment and control groups appear very similar on

[8] We use the term baseline measures when referring to administrative data, but when referring to survey data, we use the term pre-treatment measures as pre-treatment surveys were collected after randomization.

demographics and pre-treatment outcome measures. Table 2 includes regression adjusted mean differences between treatment and control students on student characteristics and pre-treatment outcome measures. All baseline and pre-treatment measures are standardized except for the number of infractions, number of suspensions, proportion of enrolled days absent, and proportion of students who report previously attending The Woodruff. [9] We observe no statistically significant difference between treatment and control group students on any demographics nor on any baseline measures from district administrative data including, test scores, course grades, attendance, and disciplinary records. We observe small, statistically significant differences on pre-treatment survey measures including students' desires to consume the arts in the future and how much they agree/disagree with the statement "People can have different opinions about the same thing." Given these small differences, we control for student demographics and pre-treatment measures in our analysis. Overall, the results of the treatment control comparison give us confidence that the randomization process produced similar groups allowing us to identify the causal impact of the intervention on student outcomes.

(Table 2 about here)

In addition to baseline equivalence, consent to and attrition from the study can affect the similarities between the treatment and control groups. Table 3 details the number of treatment and control students who consented to the study and the number of students we have year three outcome data. Consent forms were given to all enrolled fourth and fifth grade students at the beginning of the school year in which they first entered the study. We received consent to participate in the study from 79 percent of all enrolled fourth and fifth grade students at the 15

[9] Refer to appendix B, Table 1B for summary statistics on all variables.

participating elementary schools, with 90 percent of treatment students and 69 percent of control students consenting (Column 3).

While we have a high differential consent between the treatment and control groups, 21 percentage points, we have reason to believe that treatment and control students remain similar to each other. First, our population of students likely come from very similar backgrounds as they all attend schools geographically near each other and live in the same urban school district. Additionally, we randomized fourth and fifth grades within the same school further ensuring students live in similar neighborhoods and share similar experiences. Second, we demonstrate, at least on observable characteristics, that despite the differential consent rates, our treatment and control groups are similar at baseline. If treatment students were more motivated to consent to the study, we might expect measures of student engagement such as test scores, survey effort, and disciplinary records to be more favorable for treatment students, but we find no evidence of this. It is likely that there was a difference in the rate of consent between treatment and control groups because teachers in treatment grades may have just been more diligent about distributing and collecting the forms. Furthermore, differential consent is particularly concerning when the intervention likely affects attrition. We believe that it is highly unlikely that students would leave their school due to their treatment status, given that treatment consists only of three field trips.

We also consider attrition of students from the study by year three. We define students as leaving the study by year three if we do not receive administrative data for them. Of all students who were eligible to enter the study in their first year (Fall FTE), we do not receive year three administrative data for 34 percent of them (Column 3). There is a lower attrition rate when considering the number of students who have year three data and consented to the study. Of students who consented to the study, we do not receive year three administrative data for 17

percent of them, 16 percent of the control group and 17 percent of the treatment group (Column, 6). So, while there is large differential consent between the treatment and control group, once students consented to the study, we do not lose data from the treatment and control groups at significantly different rates.

(Table 3 about here)

IV. Results

Using our analytic sample of three cohorts of students, we estimate the treatment effect for both our survey and administrative data outcomes. First, we present the effect of receiving one year of treatment on our survey outcomes measures. We only estimate the effect of receiving one year of treatment, as surveys were only collected for each cohorts' first year in the study. All survey outcomes are measured in the same year students received their first dose of treatment. Table 4 presents treatment coefficients from separate regressions for each outcome. All treatment effects are expressed in standard deviation terms. We find that treatment students report a greater desire to consume arts in the future, by 9.2 percent of a standard deviation, and are more likely to agree with the statement, "I believe people can have different opinions about the same thing", by 13.8 percent of a standard deviation, than control students. We find no statistically significant effect on students' desire to participate in the arts, empathy, or on social perspective taking.

(Table 4 about here)

Additionally, we find an increase in treatment students' conscientiousness as measured by careless answering on the survey. Treatments students are less careless completing the surveys than control students by 12.1 percent of a standard deviation. We cannot detect a difference between treatment and control students in how often they skip survey questions. The

two measures of survey effort, item non-response and careless answering, should be taken jointly. Both measures are designed to proxy for students' conscientiousness while completing a school task. As such, these findings provide some evidence that the treatment affected student's conscientiousness in school.

Next, we present results using outcomes from administrative data. Table 5 presents treatment coefficients from separate regressions for each outcome. For these outcomes, we estimate the effect of receiving one year of treatment, two years of treatment, and the effect one and two years post treatment. We find that treatment students score higher on their Georgia Milestones end-of-grade exams by 10.5 percent of a standard deviation two years after treatment[10], earn higher course grades by 24.6 percent of a standard deviation two years after treatment, are less absent by 0.5 percentage points one year post treatment, and have 0.19 fewer behavioral infractions one year post treatment than control group students.

These effect sizes are both statistically significant and substantially large when considering that we estimate the effect of receiving, at most, six arts-based field trips across two years. To better understand the magnitude of these effects, the regression adjusted average number of infractions for the control group was 0.24 compared to 0.04 infractions for treatment students, which is an 83 percent decrease. Similarly, the regression adjusted percent of days students are absent from school within a year for the control group is 1.5 percent compared to 1 percent for the treatment group, which is a 33 percent decrease. However, the treatment effect of

[10] We combine students' math and ELA test scores for an overall measure of performance on the Georgia Milestones end-of-grade exams. The treatment effect remains positive and significant for both math and ELA exam scores when we run separate regressions for each subject.

the proportion of days a student was absent from school was not robust to multiple specification of the model.[11]

Interestingly, these significant effects only appear one to two years post-treatment. This is also the time when students enter middle school. Middle school differs in many ways from elementary schools, particularly when considering student discipline policies and course grades. There is limited variation in the number of infractions recorded in elementary schools. However, once students enter middle school, we observe an increase in the number of students with recorded disciplinary infractions and greater variation in the number of infractions reported. There is also greater variation in students' course grades once they enter middle school. Given these differences, it is possible that the treatment affected students' behavior at school in the first and second years of treatment, but we lacked variation in the outcomes to detect an effect in when students were in elementary school. Another possible explanation is that the treatment helped facilitate a smoother transition for students between elementary and middle schools. The field trips may have exposed students to a broader world and helped them adjust to experiences that were unfamiliar to them.

Also, contrary to what we expected, treatment students scored higher on their end-of-grade exams two years after treatment. We believed that three to six days out of the classroom for field trips were unlikely to negatively affect test scores, but at the same time, three to six arts-based field trips were unlikely to provide enough math or ELA content to significantly improve scores. There are a few possible explanations for this unexpected result. First, treatment may

[11] We excluded just under five percent of our sample due to a number of outliers in the attendance data for whom we observed very high rates of absences, with some students missing 20 to 75 percent of the school days. We originally excluded students from our sample whose absent rate was two standard deviations above the mean. On further robustness checks, the results are not robust to different exclusion rules.

have affected students' academic performance through school engagement. Given the positive treatment effects on various school engagement measures including survey effort measures, school attendance, and discipline records, treatment students could have been more engaged than control group students in a variety of ways including the effort they put towards learning academic content and demonstrating that learning on end-of-grades exams. Second, students may have learned skills or content from the field trips that assisted them on their exams. A few of The Woodruff arts partners considered the Georgia state standards when designing their programming with the goal of connecting students' experiences to classroom content. However, this explanation seems less probable given that not all of the field trips were geared toward state standards. In addition, the field trips were only three days in a school year and were unlikely to include enough content that overlapped with a significant portion of the standardized tests to account for the observed difference, particularly when the test score increase only appears in years following treatment.

(Table 5 about here)

V. Discussion

The findings from this study add to our knowledge about the effects of arts field trips for students. We find that treatment students report a greater desire to consume the arts in the future, express greater tolerance for people with different opinions, and exhibit increased conscientiousness in the same year as treatment. Treatment students also score higher on end-of-grade exams, earn higher course grades, are absent less often, and have fewer behavioral infractions than control students. These effects appear one to two years after treatment as students leave elementary school and enter middle school.

Contrary to similar research on field trips, we find no effect on students' social perspective taking (Greene et al., 2018) or empathy (Greene, Kisida, & Bowen, 2014; Kisida, Goodwin, & Bowen, 2020). While, similar to other studies, we find some evidence of increased tolerance, the treatment effect on tolerance should be interpreted cautiously as it includes only one question on the survey, due to the full tolerance scale having low reliability.

Overall, this study provides compelling evidence that arts-based field trips can benefit students' academic performance and social emotional wellbeing, and that exposure to the arts through field trips can have a compounding and persistent effect even after treatment has ceased. Given that the control group students attended at least one field trip in a school year, which could have been to one of The Woodruff art partners, our analysis effectively estimates the difference between receiving one or two field trips over two years to receiving three or six arts-based field trips over two years. Additionally, most students, 69 percent of the control group and 71 percent of the treatment group, had attended at least one of The Woodruff arts partners before the study began, meaning that many of these students had previously experienced these art forms. As such, we believe the benefits we find are the effect of multiple experiences with the arts and not simply the impact of attending a first art field trip.

There are some important limitations to our findings. First, the experimental design, while the best method to produce causal results, is, unfortunately, a black box and is not designed to give evidence of mediating mechanisms. We can only hypothesize potential explanations as to why students who attended multiple arts-based field trips exhibit greater academic performance, social emotional learning, and school engagement. There is great potential for future research to consider the possible mechanisms that contribute to the benefit of

arts-based field trips as well as explore students, teachers, and administrators experiences with such field trips.

Second, when comparing this study to previous research or when using this study's findings to inform school practices, it is important to consider the unique population of participating students. All students in our sample live in a large urban school district in the Atlanta metro area and are a racially homogenous group with over 90 percent of the entire sample identifying as black or African American. Students in our sample perform below average on state standardized tests. At baseline, most students, 78 percent, perform below proficient on the ELA Georgia Milestones end-of-grade exam, with 48 percent classified as "beginning learner" which is the lowest achievement category. It is possible that the treatment would impact a racially and academically heterogeneous group of students differently than it did for this relatively homogeneous one. While we believe this study has high internal validity, it likely has limited external validity.

Additionally, given the overall low academic performance of students in our sample, surveys are a weak instrument to measure the participating elementary students' attitudes on constructs such as tolerance, empathy, and school engagement. Working with the education team at The Woodruff along with some teachers from the participating schools, our research team designed a survey instrument that we hoped would accommodate the students' academic levels. However, while administering the survey, we noticed many students struggled to accurately complete sections of the survey, despite reading the survey aloud to all students. As such, our survey measures likely suffer from significant measurement error and should be interpreted cautiously. Fortunately, the administrative data provided by the school district do not suffer from these same limitations, and we are able to measure students' academic performance and school

engagement using their end-of-grade test scores, course grades as well as their attendance and behavioral records. Furthermore, these measures of student performance and engagement are difficult for schools to affect given that many other student, family, and school factors influence these outcomes. Given the relatively low-touch intervention students in our study experienced, it is remarkable that we detect significant effects on such outcomes.

Finally, the estimated treatment effects vary across the three cohorts. Our analytic approach combines the three cohorts to estimate the overall effect of the treatment on the various outcomes. The three cohorts of students are all in the same school district and are enrolled in neighboring elementary schools. These cohorts do not significantly differ from each other on observable characteristics. However, when looking at the estimated treatment effects by cohort, the results vary.[12] Most notably, cohort one has a strong positive treatment effect on student test scores in all treatment conditions, ranging from a 14 percent standard deviation increase in the first year of treatment to a 20 percent standard deviation increase two years following treatment. Cohort two treatment students showed a negative test score effect one year post treatment, and cohort three treatment students showed no effect in the first year of treatment. These differences across cohorts suggest that the positive test score effects we observe in our combined model may be driven by cohort one students. Cohort one is also the only cohort that has been in the study long enough for us to observe the effect two years post treatment.

The treatment effects on students' course grades also varies by cohort, with a positive effect for cohort 1 students two years post treatment, a negative effect in the first year of treatment for cohort 2 students, and a positive effect in the first year of treatment for cohort 3 students. Part of the variation we observe in treatment effects across cohorts could be due to

[12] Refer to Appendix B, Table 2B for the estimated treatment effect on all administrative outcomes by cohort.

major disruptions to students' normal school schedules in the second year of the study. In 2017, the second year of our study, Atlanta faced serious storms from Hurricane Irma in the fall followed by heavy ice storms in the winter. Each of these storms resulted in some schools in our sample closing for multiple weeks during the year. As a result, many of the field trips in the second year of the study were postponed, rescheduled, and packed into the remaining school days once students returned to school. These storms affected our cohort two students the most as it impacted their first dose of treatment. While these disruptions affected students in both the treatment and the control group, it is possible that packing field trips into an already hectic school year does not benefit student learning, or that the benefits students experience from field trips do not outweigh the negative effects from missing multiple weeks of school. Whatever the reason is for the differences in treatment effects we observe across cohorts, in our primary analysis combining all three cohorts, we include a fixed effect for each cohort that should account for unobserved differences between the cohorts as it compares treatment and control students within the same cohort. The estimated treatment effects do not significantly differ when we include or exclude the cohort fixed effect from the analysis.

VI. Conclusion

In this paper, we present the results of, to our knowledge, the first-ever multi-visit longitudinal field trip experiment. We estimate the causal effects of students receiving three arts-based field trips in one year, six field trips over two consecutive years, and the effect up to two years post treatment. The findings presented here suggest that continued exposure to the arts through field trips can benefit students' academic performance and increase their engagement in school. One of most significant policy implications comes from our findings that treatment students score higher on end-of-grade exams and receive higher course grades than control

students. These gains are strikingly significant given that the elementary schools in our sample are generally low performing schools. In part due to accountability pressures for schools to increase test scores, schools have reduced the number of field trips students attend and opted for increased seat time in core subjects. However, the evidence presented here questions the necessity to trade field trips for additional classroom instruction. While quality classroom instruction is important for student academic progress, there are other valuable ways to enhance student learning while also providing opportunities for a broader curriculum. Furthermore, students who experienced multiple field trips also attended school more often and had fewer behavioral infractions. School attendance and student discipline records are corelated with many other important outcomes for students such higher academic achievement (Anderson, Ritter, & Zamarro, 2019; Gottfried, 2010; Noltemeyer, Ward, & Mcloughlin, 2015), lower probability of grade retention (Swanson, Erickson & Ritter, 2017), and increased social and educational engagement (Gottfried, 2014). So, even in the absences of positive treatment effects on test scores or course grades, our findings suggest that arts-based fieldtrips hold value above purely student academic performance.

Another important consideration for educators and policymakers is the role of field trips in providing equitable access to cultural institutions for all students. Field trips may play a more critical role in schools where students from economically disadvantaged families attend, as their families may not have the resources to expose their children to cultural institutions outside of school at a similar rate as do higher income families. Schools can provide access to cultural institutions for all students. Moreover, schools that serve a large population of disadvantaged students also face greater accountability pressures and, as such, may further reduce the number

of field trips. Educators and policymakers should consider the multidimensional benefits from arts-based field trips when they are deciding how to allocate time and resources.

References

Anderson, K. P., Ritter, G. W., & Zamarro, G. (2019). Understanding a vicious cycle: The relationship between student discipline and student academic outcomes. *Educational Researcher*, *48*(5), 251-262.

Bowen, D., Greene, J., & Kisida, B. (2014). Learning to think critically: A visual art experiment. *Educational Researcher*, 43(1), 37-44.

Bowen, D., & Kisida, B. (2019). Investigating Causal Effects of Arts education Experiences: Experimental Evidence from Houston's Arts Access Initiative. Rice Kinder Institute for Urban Research. Retrieved from https://kinder.rice.edu/research/investigating-causal-effects-arts-education-experiences-experimental-evidence-houstons-arts

Cullen, J. B. & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. In T. J. Gronberg & D. W. Jansen. (Eds.), *Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics.* Vol. 14 (pp. 1-34), Amsterdam: Elsevier Science.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy.

Figlio D. N. & Getzler L. S. (2006). Accountability, ability, and disability: Gaming the system? In T. J. Gronberg & D. W. Jansen. (Eds.), *Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics.* Vol. 14 (pp. 35-49), Amsterdam: Elsevier Science.

Figlio, D. N., & Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of public Economics*, *89*(2-3), 381-394.

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

Gehlbach, H. (2004). A new perspective on perspective taking: A multidimensional approach to conceptualizing an aptitude. *Educational Psychology Review*, 16(3), 207-235.

Gehlbach, H., Brown S., Loannou, A., Boyer, M., Hudson, N., Niv-Solomon, A., Maneggia, D., & Janik, L. (2008). Increasing interest in social studies: Social perspective taking and self-efficacy in stimulating simulations. *Contemporary Educational Psychology*, 33(4), 894-914.

Gehlbach, H., Brinkworth, M., & Wang, M. (2012). The social perspective taking process: What motivates individuals to take another's perspective? *Teachers College Record*, 114, 1-29.

Goldstein, T., & Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of Cognition and Development*, 13(1), 19-37.

Gottfried, M. A. (2010). Evaluating the relationship between student attendance and achievement in urban elementary and middle schools: An instrumental variables approach. *American Educational Research Journal*, *47*(2), 434-465.

Gottfried, M. A. (2014). Chronic absenteeism and its effects on students' academic and socioemotional outcomes. *Journal of Education for Students Placed at Risk*, *19*(2), 53-75.

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. Report to Congressional Requesters. February, GAO-09-286. Retrieved from https://www.gao.gov/new.items/d09286.pdf

Greene, J., Erickson, H., Watson, A., Beck, M. (2018). The play's the thing: Experimentally examining the social and cognitive effects of school field trips to live theater performances. *Educational Researcher*, 47(4), 246-254.

Greene, J., Hitt, C., Kraybill, A., & Bogulski, C. (2015). Learning from live theater: Students realize gains in knowledge, tolerance, and more. *Education Next*, 15(1), 54-61.

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

Hout M., & Elliot S. W. (2011). Incentives and test-based accountability in education. National Research Council of National Academies, Washington DC.

Hitt, C. (2015). Just filling in the bubbles: Using careless answer patterns on surveys as a proxy measure of noncognitive skills. EDRE Working Paper 2015-06. Fayetteville, AR: Department of Education Reform, University of Arkansas.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, 52, 105-119.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*(3), 843-877.

Jægar, M., & Møllegarrd, S. (2017). Cultural capital, teacher bias, and educational success: New evidence form monozygotic twins. *Social Science Research*, 64, 130-144.

Keiper, S., Sandene, B., Persky, H., & Kuang, M. (2009). *The nation's report card: Arts 2008 -- music and visual arts*. National Assessment of Educational Progress at Grade 8. NCES 2009-488. Retrieved from https://nces.ed.gov/nationsreportcard/pubs/main2008/2009488.asp

Kisida, B., Goodwin, L., & Bowen, D. (2020) Teaching history through theater: The effect of arts integration on students' knowledge and attitudes. *AERA Open*, 6(1).

Kisida, B., Greene J., & Bowen, D. (2014). Creating cultural consumers: The dynamics of cultural capital acquisition. *Sociology of Education*, 87(4), 281-295.

Kornrich, Sabino (2016). Inequalities in parental spending on young children: 1972 to 2010. *AERA Open,* 2(2), 1-12. Retrieved from https://journals.sagepub.com/doi/pdf/10.1177/2332858416644180

Lee, B., Patall, E., Cawthon S., & Steingut, R. (2015). The effect of drama-based pedagogy on preK-16 outcomes: A meta-analysis of research from 1985-2012. *Review of Educational Research*, 85(1), 3-49.

Ludwig, M., Boyle, A., & Lindsay, J. (2017). Review of evidence: Arts integration research through the lens of Every Students Succeeds Act. American Institutes for Research. Washington, DC. Retrieved from http://www.air.org/resource/review-evidence-arts-integration-research-through-lens-every-student-succeeds-act

McCord, R., Ellerson, N. (2009). Looking Back, Looking Forward: How the Economic Downturn Continues to Impact School Districts. American Association of School Administrators Arlington, VA. Retrieved from http://www.aasa.org/uploadedFiles/Resources/files/LookingBackLookingForward.pdf

Mosteller, F. & Boruch, F. (Eds.) (2002). *Evidence matters: Randomized trials in education research*. Washington, D.C.: The Brookings Institution.

Noltemeyer, A. L., Ward, R. M., & Mcloughlin, C. (2015). Relationship between school suspension and student outcomes: A meta-analysis. School Psychology Review, 44(2), 224–240.

Painter, G., Lacoe, J., & Williams, D. (2015). Evaluating the academic and behavioral impacts of "School in the Park." Social Innovation Policy Brief. Sol Price School of Public Policy, University of Southern California. Retrieved from https://socialinnovation.usc.edu/files/2013/01/SITP.pdf

Peterson, P.E., Campbell, D. E., & West, M. R. (2001). An evaluation of the basic fund scholarship program in the San Francisco Bay area, California. Working Paper 01-01. Program on Education Policy and Governance, Harvard University. Retrieved from https://www.innovations.harvard.edu/evaluation-basic-fund-scholarship-program-san-francisco-bay-area-california

Pirog, M. A., Buffardi, A. L., Chrisinger, C. K., Singh, P., & Briney, J. (2009). Are alternatives to randomized assignment nearly as good? Statistical corrections to nonrandomized evaluations. *Journal of Policy Analysis and Management*, 28(1), 169–172.

RK & Associates (2018). Impact study: The effects of facilitated single-visit art museum programs on students grades 4-6. Prepared for the National Art Education Association and Association of Art Museum Directors. Retrieved from https://www.arteducators.org/research/articles/377-naea-aamd-research-study-impact-of-art-museum-programs-on-k-12-students

Rabkin, N., & Hedberg, E. (2011). Arts education in America: What the declines mean for arts participation. National Endowment for the Arts, Research Report #52. Retrieved from https://www.arts.gov/publications/arts-education-america-what-declines-mean-arts-participation

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. Seventh edition. Sage.

Ruppert, S. (2006). Critical evidence: How the ARTS benefit student achievement. National Assembly of State Arts Agencies. Retrieved from https://nasaa-arts.org/wpcontent/uploads/2017/05/critical-evidence.pdf

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practices. In L. A. Hamilton, B. M. Stecher Finn & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 80-100), Arlington, VA: RAND.

Student Youth & Travel Association (2016). Student youth and travel digest: A comprehensive survey of the student travel market. Mclean, VA. Retrieved from https://syta.org/wp-content/uploads/2017/08/SYTD-Social-Impact-8.11.pdf

Swanson, E., H. Erickson, H., & Ritter, G. W. (2017). Examining the Impacts of Middle School Disciplinary Policies on Ninth-Grade Retention. *Educational Policy*, 0895904819843600.

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. F. Finn & D. Ravitch (Eds.), *Beyond basics: Achieving a liberal education for all children* (pp. 46-62), Washington DC: Thomas B. Fordham Institute.

Zamarro, G., Cheng, A., Shakeel, M., & Hitt, C. (2016). Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics*, 72, 51-60.

Tables and Figures

Table 1: Treatment Assignment in Year 3 by School and Cohort

School 1		**School 5**		**School 11**	
4th	-	4th	Treatment	4th	Treatment
5th	Treatment	5th	Control	5th	Control
6th	Control	6th	Treatment- 1yr post		
7th	Treatment- 2yr post	**School 6**		**School 12**	
School 2		4th	Treatment	4th	Treatment
4th	-	5th	Control	5th	Control
5th	Treatment	6th	Treatment- 1yr post		
6th	Control	**School 7**		**School 13**	
7th	Treatment- 2yr post	4th	Treatment	4th	Treatment
School 3		5th	Control	5th	Control
4th	-	6th	Treatment- 1yr post		
5th	Control	**School 8**		**School 14**	
6th	Treatment- 1yr post	4th	Control	4th	Control
7th	Control	5th	Treatment- double dose	5th	Treatment
School 4		6th	Control		
4th	-	**School 9**		**School 15**	
5th	Control	4th	Control	4th	Control
6th	Treatment- 1yr post	5th	Treatment- double dose	5th	Treatment
7th	Control	6th	Control		
		School 10		**KEY**	**Cohort 1**
		4th	Control		**Cohort 2**
		5th	Treatment- double dose		**Cohort 3**
		6th	Control		

Notes: Table shows treatment assignments by cohort and grade levels in year 3. Randomization occurred within schools between 4th and 5th grades. Students in 6th or 7th grade in year three were randomly assigned to the treatment or control group when they were in 4th or 5th grade in their first year. Students enter middle school in grade 6; however, we show students in 6th and 7th grade in the school they were randomized. Cohort 1 treatment students who are in 6th grade in year 3 entered the study in year 1 as 4th graders. These students received treatment in 4th grade and another dose in 5th grade.

Table 2: Baseline Equivalence of Treatment and Control Groups

Variables	Treatment (mean)	Control (mean)	Difference (T-C)	N
	(1)	(2)	(3)	(4)
Demographics:				
Female	0.51	0.52	-0.01	2159
Black or African American	0.99	0.99	0.00	2148
Students with Disabilities (SWD)	0.16	0.17	-0.01	2047
Baseline Academic Performance				
ELA	-0.32	-0.36	0.03	1904
Math	-0.27	-0.34	0.07	1902
Combined Tests	-0.32	-0.36	0.04	1907
Course Grades	0.17	0.12	0.05	1922
Baseline Discipline Measures				
Infractions	0.13	0.10	0.02	2163
Suspensions	0.02	0.04	-0.02	2163
Baseline Proportion of Days Absent	0.04	0.04	0.00	1939
Pre-treatment Survey Measures				
Desire to Consume Art	0.12	0.01	0.11*	1947
Desire to Participate in Art	0.05	0.03	0.02	1947
Social Perspective Taking	0.09	0.05	0.04	1933
Empathy	0.00	-0.07	0.07	1946
"Different opinions about the same thing"	0.09	-0.03	0.13**	1924
Previously attended The Woodruff				
Previously attended Alliance Theater	0.31	0.32	-0.01	1910
Previously attended Atlanta Symphony	0.46	0.42	0.04	1915
Previously attended High Museum of Art	0.53	0.50	0.03	1832
Pre-treatment Survey Effort				
Careless Answers	0.06	0.15	-0.09	1947
Item Non-response	0.13	0.13	-0.01	1936

Notes: *** p<0.01, ** p<0.05, * p<0.1. The treatment and control group means are regression adjusted controlling for school fixed effects with standard error clustered at the teacher level. All baseline and pre-treatment measures are standardized except for the number of infractions, number of suspensions, proportion of enrolled days absent, and proportion of students who report previously attending The Woodruff. All test scores are standardized Georgia Millstone end-of-grade exams and are standardized within grade level by year.

Table 3: Consent and Attrition by Treatment Assignment

	(1)	(2)	(3)	(4)	(5)	(6)
	Fall FTE	**Consent**	**Attrition**	**Yr. 3 Data**	**Attrition**	**Attrition**
	# of Students	*# of Students*	*FTE to Consent*	*# of Students*	*FTE to Yr. 3 Data*	*Consent to Yr. 3 data*
Total Sample	2767	2196	0.21	1831	0.34	0.17
Control	1398	967	0.31	813	0.42	0.16
Treatment	1369	1229	0.10	1018	0.26	0.17
Difference (C-T)	29	-262	0.21	-205	0.16	-0.01

Notes: Fall FTE comes from the Georgia Department of Education and is the best estimate of the number of students who were eligible to participate in the study. As we randomized by grade level within a school, the Fall FTE represents the sum of all 4th and 5th grade students enrolled in the participating 15 schools in the years each school entered the study. Schools distributed consent forms to all enrolled 4th and 5th grade students at the beginning of the school year. We consider students as having year 3 data if we received district administrative data for them in school year 2018-19.

Table 4: Estimated Treatment Effect on Survey Outcome Measures

	1st Treatment	Controls	N
	(1)	(2)	(3)
Art Consumption	0.092*	X	1451
	(0.053)		
Art Participation	0.025	X	1451
	(0.052)		
Tolerance "Different Opinions"	0.138***	X	1422
	(0.046)		
Social Perspective Taking	0.054	X	1435
	(0.056)		
Empathy	-0.067	X	1449
	(0.056)		
Non-Response	-0.069	X	1446
	(0.063)		
Careless Answering	-0.121**	X	1450
	(0.057)		

Notes: *** p<0.01, ** p<0.05, * p<0.1. All effects are expressed in standard deviation terms. Standard errors clustered at the classroom level are in parentheses. All models include school and cohort fixed effects along with controls for students' gender, SWD status, baseline standardized test scores, and the pre-treatment measure of the given outcomes. Student random effects are not included as we only estimate the effect in the first year of treatment.

Table 5: Estimated Treatment Effect on Test Scores, Course Grade, Attendance, & Infractions

	1st Treatment	2nd Treatment	1 Yr. Post Treatment	2 Yrs. Post Treatment	# observations	# of students
	(1)	(2)	(3)	(4)	(6)	(7)
Combined Test Score	0.031	0.030	0.002	0.105***	3107	1825
	(0.030)	(0.040)	(0.039)	(0.034)		
Course Grades	0.015	0.029	0.094	0.246**	3157	1842
	(0.052)	(0.060)	(0.087)	(0.097)		
Proportion Absent	-0.001	0.002	-0.005**	0.006	3071	1825
	(0.002)	(0.002)	(0.003)	(0.004)		
# of Infractions	0.043	0.093	-0.193**	0.043	3359	1929
	(0.049)	(0.067)	(0.097)	(0.153)		

Notes: *** p<0.01, ** p<0.05, * p<0.1. Estimated treatment effects for each outcome are from separate regressions. Standard errors clustered at the classroom level are in parentheses. All models include school and cohort fixed effects and student random effects, along with controls for students' gender, SWD status, baseline combined standardized test scores, and the baseline measure of the given outcome. The combined test score is a standardized score of students' Georgia Milestone ELA and math exams. Test scores are standardized within grade by year. We removed a small number of outliers in our models estimating the treatment effect on the number of infractions and the proportion of days absent, accounting for less than 0.5 percent of the sample in the infraction sample and 5 percent of the sample in the attendance data.

Appendix 1: Survey Instrument

Our research team administered surveys at the beginning and end, following treatment, of each students' first year. The surveys were administered on paper and a member of our research team read aloud each question along with answer options while students completed their surveys. The survey also included demographic questions on students' age, race/ethnicity, and gender, which are not included in this appendix.

Art Consumption (Greene et al., 2018; Greene, Kisida, & Bowen, 2014) **-** students indicate whether they disagree a lot, disagree a little, do not agree or disagree, agree a little, or agree a lot with each statement.

Visual Arts

- Visiting art museums is fun.
- I plan to visit art museums when I am an adult.
- Art is interesting to me.
- I feel like I don't belong when I'm at an art museum.
- I feel comfortable talking about art.
- I would tell my friends that they should visit an art museum.
- How interested are you in visiting an art museum?
  Students had different answer options for this item.
  Not interested / Slightly interested / Somewhat interested / Interested / Very interested

Symphony

- Listening to orchestra music is interesting to me.
- I feel comfortable talking about orchestra music.
- I would tell my friends that they should hear an orchestra music concert.
- I plan to go to orchestra music performances when I am an adult.
- Orchestra music concerts are fun.
- How interested are you in going to an orchestra music performance?
  Students had different answer options for this item.
  Not interested / Slightly interested / Somewhat interested / Interested / Very interested

Theater

- Trips to see live theater are fun.
- Live theater is interesting to me.
- I feel comfortable talking about theater performances.
- I would tell my friends that they should see a live theater performance.
- I feel like I don't belong when I'm in a theater.
- I plan to see live theater performances when I am an adult.
- How interested are you in seeing live performances in a theater?

Students had different answer options for this item.
Not interested / Slightly interested / Somewhat interested / Interested / Very interested

Art Participation (Greene et al., 2018; Greene, Kisida, & Bowen, 2014) **-** students indicate whether they are not interested, slightly interested, somewhat interested, interested, or very interested to each statement.

Visual Arts

- How interested are you in making a work of art?
- How interested would you be in entering your work of art in a contest?
- How interested are you in taking an art class?
- I would be interested in joining an art club if my school had one.

Symphony

- If your school had an orchestra or band, how interested would you be in playing a musical instrument in it?
- How interested are you in taking music class?
- How interested are you in learning to play a musical instrument?
- I would be interested in joining an orchestra music club if my school had one.

Theater

- How interested are you in being in a theater performance?
- How interested are you in taking a drama class?
- If your school were having auditions for a play, how interested would you be in trying to get a role in that play?
- I would be interested in joining a drama club if my school had one.

Empathy- students indicate whether they disagree a lot, disagree a little, do not agree or disagree, agree a little, or agree a lot with each statement.

- It upsets me when another child is being shouted at.
- When I see someone suffering, I feel bad too.
- It makes me sad to see a child who can't find anyone to play with.

Fantasy Empathy Items (Davis, 1980).[13]

- After seeing a play or movie, I have felt as though I were one of the characters.
- When I watch a good movie, I can very easily put myself in the place of the leading character.
- When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me.

Political Tolerance (Peterson, Campbell, & West, 2001) - students indicate whether they disagree a lot, disagree a little, do not agree or disagree, agree a little, or agree a lot with each statement. [14]

[13] The fantasy empathy subscale was added in the second year of the study.
[14] The political tolerance scale was added in the second year of the study.

- Some people have views that you oppose very strongly. Do you agree that these people should be allowed to come to your school and give a speech?
- Some people have views that you oppose very strongly. Do you agree that these people should be allowed to live in your neighborhood?
- Some people have views that you oppose very strongly. Do you agree that these people should be allowed to run for president?

Tolerance (Greene, et al., 2018; Greene, Kisida, & Bowen, 2014)**-** students indicate whether they disagree a lot, disagree a little, do not agree or disagree, agree a little, or agree a lot with each statement.

- I think people can have different opinions about the same thing.
- Women are equally able to do the same jobs that men can do.
- I am interested in learning about people different than me.

School Engagement- students indicate whether they disagree a lot, disagree a little, do not agree or disagree, agree a little, or agree a lot with each statement.[15]

- Sometimes school is a waste of time.
- I feel proud being a part of this school.
- Getting good grades is important to me.
- School is boring.

Social Perspective Taking (Gehlbach, 2004; Gehlbach et al., 2008; Gehlbach, Brinkworth, & Wang, 2012; Greene et al., 2018)- students had the following answer options, almost never, once in a while, sometimes, often, or almost all the time

- How often do you attempt to understand your friends better by trying to figure out what they are thinking?
- How often do you try to think of more than one explanation for why someone else acted as they did?
- Overall, how often do you try to understand the point of view of other people?
- When you are angry at someone, how often do you try to "put yourself in his or her shoes"?
- How often do you try to figure out what motivates others to behave as they do?
- How often do you try to figure out what emotions people are feeling when you meet them for the first time?
- In general, how often do you try to understand how other people view the situation?

[15] In the first year of the study, only the item "School is boring." was included on the survey. In the second year of the study we added the remaining items.

Appendix 2: Additional Tables

Table A1: Summary Statistics

	Observations	Mean	SD	Minimum	Maximum
Treatment Variables					
Ever treatment	3,908	0.56	0.50	0	1
First treatment	3,908	0.31	0.46	0	1
Second Treatment	3,906	0.10	0.30	0	1
One-year post treatment	3,906	0.11	0.32	0	1
Two-years post treatment	3,908	0.03	0.16	0	1
Demographics					
Female	3,896	0.53	0.50	0	1
SWD	3,782	0.14	0.34	0	1
Black or African American	3,889	0.93	0.26	0	1
Pre- and Post- Treatment Measures					
Pre-combined test score	3,425	0	1	-2.78	4.15
Post-combined test score	3,473	0	1	-2.50	3.56
Pre-course grades	3,461	0	1	-5.06	2.71
Post-course grades	3,547	0	1	-5.88	3.06
Pre-proportion days absent	3,141	0.03	0.03	0	0.13
Post-proportion days absent	3,141	0.04	0.03	0	0.15
Pre-number of infractions	3890	0.14	0.61	0	7
Post-number of infractions	3890	0.35	0.98	0	7
Pre-art consumption	3,463	0	1	-3.47	1.62
Post-art consumption	2,061	0	1	-2.99	1.88
Pre-art participation	3,463	0	1	-3.31	1.92
Post-art participation	2,061	0	1	-2.98	1.97
Pre "different opinions"	3,418	0	1	-3.61	0.55
Post "different opinions"	2,049	0	1	-3.73	0.56
Pre- SPT	3,436	0	1	-2.79	2.00
Post- SPT	2,050	0	1	-2.61	2.11
Pre-empathy	3,462	0	1	-3.47	1.02
Post-empathy	2,060	0	1	-3.20	1.11
Pre-careless answers	3,463	0	1	-2.40	2.96
Post-careless answers	2,060	0	1	-2.61	3.19
Pre-item non-response	3,451	0	1	-1.60	16.02
Post-item non-response	2,054	0	1	-3.10	19.01

Notes: Table includes summary statistics for dependent and independent variables. All student observations over the three years are included. There is a total of 2,197 individual students with 968 as control and 1,229 as treatment. Most outcomes variables are in standard deviations, except for the number of infractions and the proportion of days absent from school. All test scores are standardized Georgia Millstone end-of-grade exams and are standardized within grade level by year. We removed a small number of outliers in the number of infractions and the proportion of days absent, accounting for less than 0.5 percent of the sample in the infraction outcome analysis and 5 percent of the sample in the attendance outcome analysis.

Table A2: Treatment Effect by Cohort on Test Scores, Course Grade, Attendance, and Infractions

	1st Treatment	2nd Treatment	1 Yr. Post Treatment	2 Yrs. Post Treatment	# observations	# of students
	(1)	(2)	(3)	(4)	(6)	(7)
Panel A: Cohort 1						
Combined Test Score	0.141**	0.211***	0.141***	0.201***	1166	467
	(0.058)	(0.062)	(0.054)	(0.050)		
Course Grades	-0.082	-0.008	0.158	0.272**	1195	469
	(0.103)	(0.120)	(0.111)	(0.110)		
Proportion Absent	-0.001	0.001	-0.004	0.008*	1131	459
	(0.003)	(0.003)	(0.004)	(0.004)		
# of Infractions	0.043	-0.068	-0.252	-0.244	1341	494
	(0.085)	(0.105)	(0.226)	(0.217)		
Panel B: Cohort 2						
Combined Test Score	-0.040	-0.070	-0.128**	-	1281	687
	(0.052)	(0.063)	(0.065)	-		
Course Grades	-0.169*	-0.135	-0.150	-	1298	696
	(0.096)	(0.126)	(0.157)	-		
Proportion Absent	0.002	0.006*	-0.009***	-	1293	647
	(0.002)	(0.003)	(0.003)	-		
# of Infractions	0.096	0.305***	-0.188	-	1359	727
	(0.105)	(0.102)	(0.186)	-		
Panel C: Cohort 3						
Combined Test Score	0.034	-	-	-	660	660
	(0.049)	-	-	-		
Course Grades	0.271***	-	-	-	664	664
	(0.097)	-	-	-		
Proportion Absent	-0.007***	-	-	-	647	647
	(0.002)	-	-	-		
# of Infractions	-0.062	-	-	-	686	686
	(0.132)	-	-	-		

Notes: *** p<0.01, ** p<0.05, * p<0.1. Estimated treatment effects for each outcome are from separate regressions. Standard errors clustered at the classroom level are in parentheses. All models include school fixed effects along with controls for students' gender, SWD status, baseline combined standardized test scores, and the baseline measure of the given outcome. Models for cohort 1 and 2 also include student random effects. Combined test score is a standardized score of a student's Georgia Milestone ELA and math exam. Test scores were standardized within grade by year. We removed a small number of outliers in our models estimating the treatment effect on the number of infractions and the proportion of days absent, accounting for less than 0.5 percent of the sample in the infraction sample and 5 percent of the sample in the attendance data.