The Best Paintings in Life Are Free: A study of the effects of pricing variance on the enjoyment of fine art paintings and perceptions of artists' abilities.

John Pette, Kyle Chuang, Andrew Larimer, Emily Rapport

12/16/2018

I. Research Question

In the world of professional art, the monetary value of a painting is determined by a variety of factors including the historical sales of the artist, trends in popularity of subjects or styles, and, presumably, the actual quality of the work. To the layperson, it can be difficult to untangle how much of one's perception of a piece of art is influenced by the work itself, as opposed to by the contextual details around it, including its price. We began this project interested in this phenomenon, wondering specifically if and how the price of a piece of art plays a role in the layperson's perception of that art.

II. Experimental Design

II.A. Overall Design

Our experiment attempts to answer this question: does the price of a painting impact 1) the subject's enjoyment of it and 2) the subject's perception of its quality?

A key consideration in our design is that monetary value, conceptually, can best be understood relatively. One cannot understand whether a painting worth $1,000 is expensive or inexpensive without having other priced art to compare it to.

Additionally, people have their own inherent gauges of whether a price is expensive or inexpensive, based on their own relationships to money and their life experiences, so without context there is likely high variation in how individuals perceive the same prices. For this reason, if we wanted to measure how a person's ratings of a piece of art are related to that piece's price, we thought it was important that the subject saw a variety of prices, so that a given piece's price was either high or low within the context of prices on other works of art.

In order to measure the difference in how a subject rates higher-priced and lower-priced art, and to control for any inherent differences in how people would rate the art assigned each price, we chose a difference-in-differences design. We compared the difference in the subjects' average ratings for low priced paintings and their average ratings for high priced paintings to the difference in ratings between the same sets of paintings for subjects who did not see any variation in prices.

Our experiment included one treatment group and one control group. Subjects from each group took surveys that showed them paintings one by one. The pieces of art were accompanied by a few pieces of information: the title, the artist's name, and the price. The pieces were relatively obscure, such that we did not expect subjects to have preconceived notions of the works or the artists. There were 20 pieces of art in total, divided into two groups we called the Constant Price Set and the Varied Price Set. Both the treatment and control groups saw both sets of art, but the prices they saw for the Varied Price Set diverged.

For the 10 pieces in Constant Price Set, both treatment and control saw the same prices, in dollar values that ranged from $1,000-5,000; for the 10 pieces in Varied Price Set, the control group saw prices in that same range, whereas the treatment group saw prices in the $10,000-15,000 range.

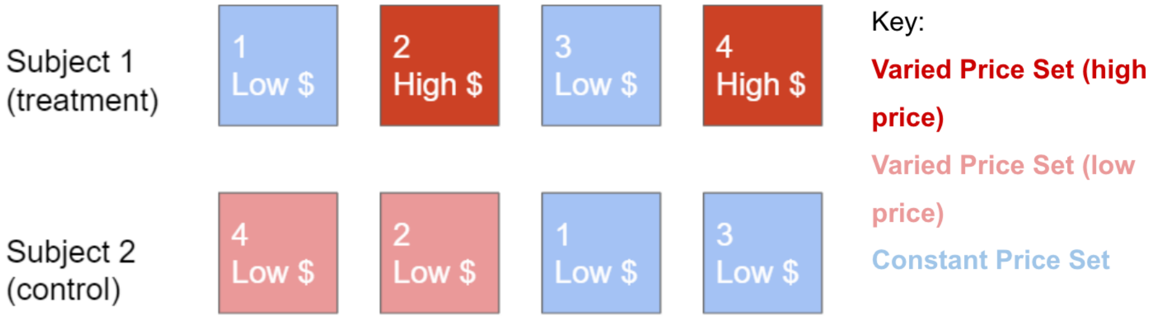

Figure 1: In our design, treatment and control see the same base-price paintings and prices, but on a second set of paintings, the control sees low prices and the treatment sees high prices.

The subjects reviewed the pieces of art one-by-one, with the ordering of the pieces determined randomly, such that different subjects saw the art from the two groups interspersed in random sequences. For instance, subject 1 might have seen 1. Constant, 2. Varied, 3. Constant, 4. Varied, and so on, whereas subject 2 saw 1. Varied, 2. Varied, 3. Constant, 4. Constant, etc., as shown in Figure 1.

We displayed the paintings in random order because we did not want to exclusively study the effect of anchoring on low prices and then seeing higher prices or vice versa anchoring on high prices and then seeing lower prices; we relied on the fact that different anchoring prices would be distributed randomly across the groups, and thus we would be studying whether a gap emerged between prices diverging in either a high or low direction.

For each piece of art, the subjects answered two questions on a 5-star scale (with rating precision at the half-star level):

- How would you rate the painter's ability?
- How much do you enjoy this piece?

The potential outcome of interest, for each of the two questions above, is the difference between subjects' average ratings for the 10 Constant Price Set paintings and their average ratings for the 10 Varied Price Set paintings. For the control group, this is the difference in means between the ratings for two groups of art that have the same family of prices; we would expect that the difference here is trivial, but might be nonzero due to the variation in the art included in the two groups (people could on average prefer the art in one set to the art in the other set for any number of idiosyncratic reasons). For the treatment group, this is the difference in means between the set of art with lower prices and the set of art with higher prices.

This design allowed us to understand the treatment group's differences in means as the impact of seeing different prices, since the control group's difference in means measure allowed us to control for the differences in the average ratings caused by general preferences to one set of art and its other metadata. When we compared the control group's average difference in ratings between sets to the treatment group's average difference in ratings, we were isolating the effect of the price manipulation and controlling for different overall ratings of the paintings. We also included a variety of demographic questions at the end of the survey so that we could control for different covariates that we thought might be related to perception of art and prices, such as age, income level, and education. For another control variable, we also asked whether subjects had engaged in one of the following arts-related activities in the last year: visiting an art museum or gallery, taking a visual art class, or purchasing a piece of art.

II.B. Sourcing Art and Subjects

In selecting the art we showed to our subjects, we limited ourselves to a single medium: paintings. This controlled for natural disparities of opinions that might exist between media. We wanted to avoid complicating people's subjective opinions by forcing them to judge a painting alongside another medium, such as a sculpture.

We sourced the photos of the paintings and associated metadata from Paddle8.com, an art auction website that primarily features modern and contemporary art. We chose Paddle8 because it contained professional photos of paintings of consistent quality. All of the paintings were products of lesser-known artists. This was intentional, as we wanted to reduce the chance of a subject being influenced by prior familiarity with an artist's work. If a subject recognized an artist, it could bias their opinions of that work. We also made sure to vary the art itself; no two pieces were by the same artist. Again, we did this to minimize potential bias.

In sourcing subjects, we had two main goals: obtain a representative sampling of the U.S. population and obtain enough survey responses to have sufficient statistical power. We first considered using friends and family via Facebook or hiring people through Mechanical Turk, but we were not satisfied that these options would give us a representative sampling. Our team identified a good alternative: a market research company called Lucid, which offered exactly what we wanted: a sample of people representative of U.S. demographics. For academic research, they were able to provide results for $1 per survey response, and once we launched, we had survey results within a day. This cut down our time constraints substantially and helped us with goal #2: statistical power.

When we were initially looking at recruiting subjects ourselves, we thought that the maximum number of responses we could realistically expect to receive over a 2-3 week period would be 150. We used this figure in our initial power calculations. We assumed a large standard deviation for our calculations, because we thought it was reasonable to assume that a large number of respondents could be at either end of the 5-point rating range. With a standard deviation assumption of 1.1, we calculated that we could detect an average treatment effect of 0.5 with 80% power.

Once we had the Lucid sample, we realized we could revise our calculations. We adjusted the power calculation to include 350 responses – we wanted to leave room in our $500 budget to repeat our experiment if something went horribly wrong. With 350 responses and the same assumptions, we would have 99% power (or alternately, we would have an 80% chance of detecting an effect of 0.33 or larger). Ultimately, our sample size was 381 and the power for the collected dataset was 81% and 86% for Questions 1 and 2, respectively.

II.C. User Testing Takeaways

We tested the initial draft of the survey on several volunteer subjects prior to launch and uncovered come unforeseen issues. Our most troubling concern was that some people were not noticing the prices. Initially, we had tried to incorporate enough metadata that the nature of our experiment would not be obvious, but we overextended this approach. Following user testing, we cut down the metadata and made the prices bold to maximize chances that users would see them.

The other key problem we encountered was with our first question for each painting. We had phrased it as, "How would you rate the technical quality of the piece?", but some people did not understand the question, and those who did found it difficult to answer. We rephrased it to be, "How would you rate the painter's ability?"

Our initial concern about survey length was ultimately unfounded. We had worried that, with 20 different images and demographic questions, our survey was too long, as Lucid had imposed a timing cap for survey completion of 15 minutes. User testing showed this to be a non-issue; subjects generally required between 5 and 10 minutes to finish.

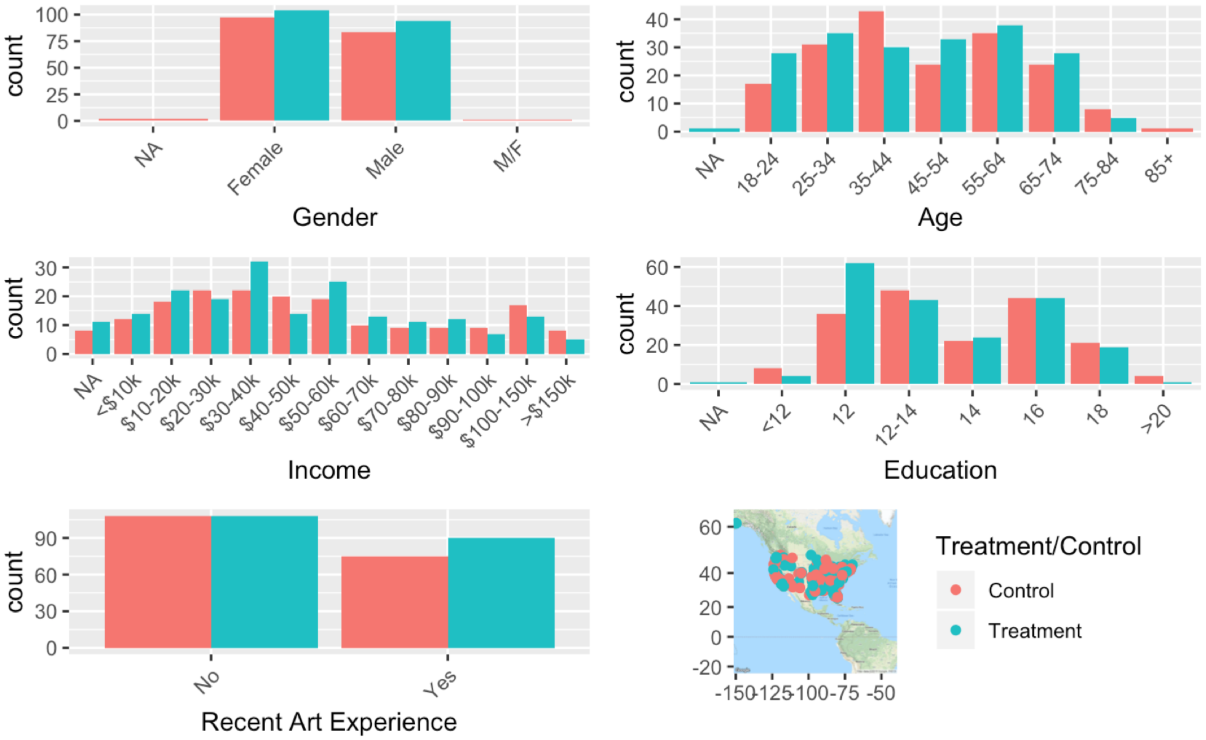

Figure 2: Graphing our control and treatment groups shows balance accross the groups.

III. Data Exploration

There were a total of 381 subjects in the experiment. 198 subjects were treated while 183 served as control, for a treatment/control ratio of 51/49. The balanced experiment helped minimize the standard error of the results.

We examined the data to ensure that randomization had been done properly and no clustering occurred. We studied the following variables to ensure proper randomization: gender, age, income, education, recent art experience, and geographic spread. Plots of these distributions across our treatment and control groups are displayed in Figure 2.

There are a few things to note in the distributions. In the gender categories, NA and non-binary each had one data point. In the age category, 75+ had very few data points and we realized we may have been better off combining the top two age groups, as 85+ had only a single response. In education, despite the few data points between the "<12" and ">20" Years of Education categories, we chose to keep them separate.

The variables indicated a roughly balanced split across all variables, as well as geographic spread. We performed the Wilcoxon sum rank test on all of the variables except geographic spread to ensure the similarities between the control and treatment groups. We did not reject the null hypothesis that the two groups were similar.

We performed a final check by comparing the mean ratings within the Constant Price set for both groups, the results of which are displayed in Figure 3.

The histograms indicate very similar ratings. Since the distribution of the ratings was not normal, we performed a randomization inference between the two groups. The difference between the two means was -0.03. From a simulation of 1,000 trials, the p-value for that difference in means was 0.17. The 95% confidence interval for the simulation was -0.0667 and +0.0666. Since 0 is well within the 95% confidence interval for

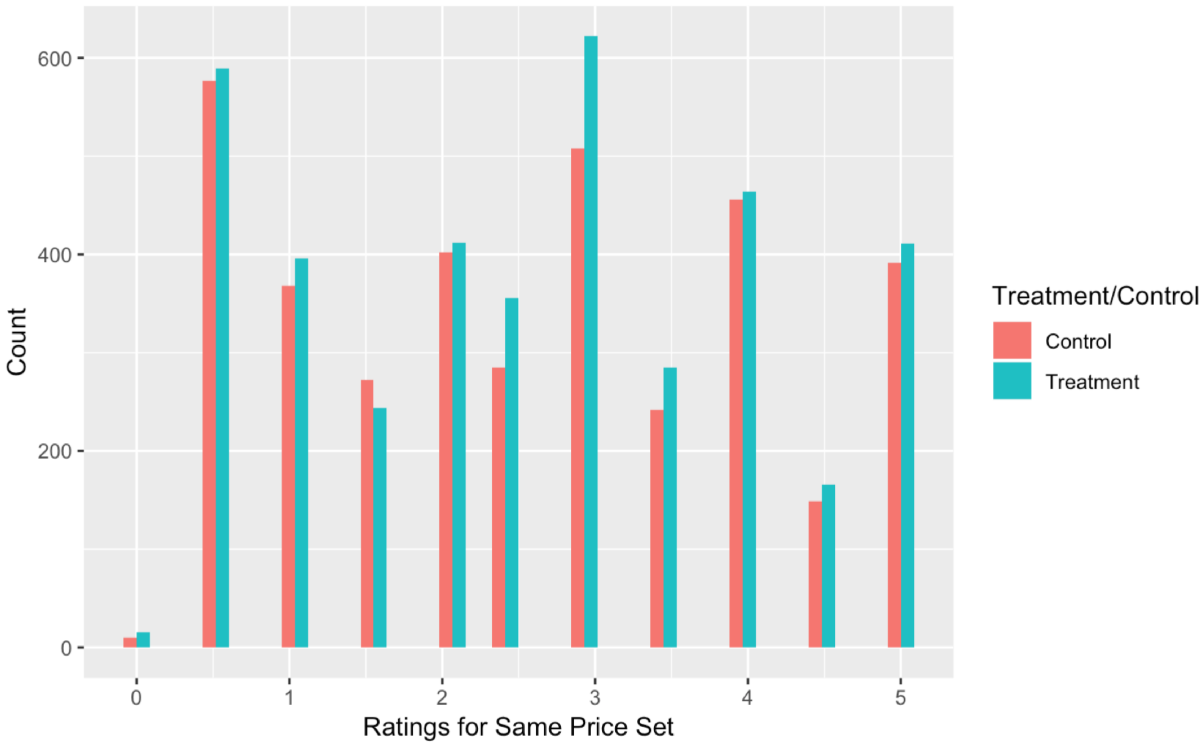

Figure 3: The mean ratings within Constant Price set for both groups continues to show balance.

the difference in means, that suggests that from a perspective of statistical significance, the mean difference between the average ratings is effectively zero. The zero mean confirms proper randomization of the subjects, as well as the apples-to-apples comparison of the control and treatment groups.

III. Main Findings

In keeping with our Difference-in-Difference design, our primary regression model is shown here:

$$Rating \sim Treatment + VariedPriceSet + Treatment: VariedPriceSet$$

Variable definitions

- Rating: The subject's rating on the given Question (we ran models separately for Question 1 and Question 2)
- Treatment: an indicator for whether the subject is in the treatment group.
- VariedPriceSet: an indicator for whether a particular painting is in the set of paintings that had different prices for treatment and control.
- Treatment:VariedPriceSet: an interaction term that equals 0 when both of the indicators above are true.

We are primarily interested in the coefficient on the interaction term "Treatment:VariedPriceSet." The VariedPriceSet paintings are those which were displayed with high prices to the treatment group; therefore, this coefficient can be interpreted as the change in average rating for high-priced paintings vs other paintings.

We ran separate models for the ratings on Question 1 (“How would you rate the painter’s ability?”) and Question 2 (“How much do you enjoy this piece?”). We evaluated each question’s data with both a Short Model, which is the regression shown above, as well as a Long Model that included indicator variables for standard intervals of our demographic covariates: gender, age, education level, household income, and the subject’s response to our question about art activity.

If we were unable to detect any treatment effect with the Short Model, it was possible that we might detect an effect in the Long Model, which controls for possible variance associated with the covariates.

The results of our regression are displayed in Table 1, which shows our main treatment coefficients, with standard errors in parentheses. The covariates in the Long Model have been omitted to save space. We note that all standard errors have been adjusted to use robust standard errors to be conservative.

Table 1: Short Regression and Long Regression With All Demographic Indicators (Omitted)

	Dependent variable:			
	rating			
	Question 1 (Short)	Question 2 (Short)	Question 1 (Long)	Question 2 (Long)
Intercept	2.851^{***}	2.501^{***}	3.441^{***}	3.272^{***}
	(0.034)	(0.035)	(0.459)	(0.437)
Treatment	0.026	0.042	−0.014	0.006
	(0.047)	(0.049)	(0.047)	(0.047)
Price Varies	-0.087^{*}	-0.127^{**}	-0.087^{*}	-0.127^{***}
	(0.048)	(0.050)	(0.046)	(0.047)
Treatment:Price Varies	0.005	−0.014	0.005	−0.014
	(0.066)	(0.068)	(0.064)	(0.064)
Observations	7,620	7,620	7,620	7,620
Adjusted R^2	0.001	0.002	0.082	0.112
Residual Std. Error	1.448 (df = 7616)	1.479 (df = 7616)	1.388 (df = 7585)	1.396 (df = 7585)
Note:				*p<0.1; **p<0.05; ***p<0.01
				Parentheses denote robust standard errors.

We see that the coefficient of the interaction term is not statistically significant, even in the models with full covariate coverage, and thus we are unable to reject the null hypothesis that the price of paintings have no effect on viewers’ assessment of art quality, or of their enjoyment of the art.

The coefficient for treatment is 0 as expected, indicating proper randomization and a good basis for the experiment. A zero effect of the subject being in the treatment group suggests no issues with our randomization. Interestingly, the VariedPrice coefficient is statistically significant for Question 2 in both versions of our models. This shows that, on average, the subjects rated their enjoyment of the VariedPrice paintings lower than the other set of paintings, regardless of whether they saw those paintings having a lower price. By using this coefficient to detect the effect of the consistent elements of the paintings (the work itself, the title, the name of the artist) on the ratings, we ensure that what is left over in the interaction term captures only the effect of the higher prices on ratings.

IV. Secondary Findings

Many of the covariates were not significant individually but jointly significant under an F-test. Tables 2 and 3 display the F-Statistics for Questions 1 and 2 respectively.

	Df	F	Pr(>F)
Treatment	1	0.12	0.7328
Varied Price Set	1	7.05	0.0079
Gender	3	11.16	0.0000
Age	8	22.10	0.0000
Income	12	13.25	0.0000
Education	7	5.14	0.0000
Art Activity	1	191.50	0.0000
Treatment:Varied Price Set	1	0.01	0.9330
Residuals	7585		

Table 2: F-Statistics for Question 1

	Df	F	Pr(>F)
Treatment	1	0.00	0.9759
Varied Price Set	1	17.66	0.0000
Gender	3	63.11	0.0000
Age	8	36.91	0.0000
Income	12	13.77	0.0000
Education	7	5.26	0.0000
Art Activity	1	245.58	0.0000
Treatment:Varied Price Set	1	0.04	0.8324
Residuals	7585		

Table 3: F-Statistics for Question 2

These F-test results suggest that Gender, Age, Income, Education and Art Activity all have statistically significant impacts on subjects' ratings.

Close examination of Gender shows that the significance primarily came from categories NA and non-binary. Those 2 categories contained a total of three subjects. Otherwise, the coefficients for Male and Female are almost exactly the same, indicating no true difference.

The fact that many of the covariate indicators were jointly significant, but not individually significant, motivated us to binarize the variables in an attempt to understand the actual effects of the covariates on ratings.

We replaced the full set of age indicator covariates with a single “55 and older” covariate. The age 55 was chosen due to evidence in our data that the ratings began to decline in our data set starting around that age for question one (the ratings for question 2 begin getting lower at a younger age, but we wanted to just choose one binarizing variable). We note that Ages 74-85 only contained 13 data points and Ages 85+ only had one; combining these three groups into a single covariate helped us avoid making it look as though something was significant just because there were very few respondents in that category.

Income was split into $60,000 and over. The $60,000 number was chosen from the approximate U.S. median household income.

Education was split into finishing 16 years of schooling (finishing a bachelor's degree) or more. Finally, we performed a regression to see if any of the covariate terms themselves had significant coefficients, indicating a relationship between that covariate and price, and to see if any of the interaction terms with the covariates had significant coefficients, indicating heterogeneous treatment effects.

As seen in Table 4 below, found that answering yes to the question about participation in an art activity had a significant positive association with ratings (.490 and .557 stars for Question 1 and Question 2, respectively), but it did not have a statistically significant relationship with how respondents' ratings were influenced by price. Being over age 55, however, was associated with both participants' overall ratings and how their ratings changed for higher-priced paintings.

Table 4: Exploring Covariate Interactions with Treatment:Price Varies

	Dependent variable:	
	rating	
	Question 1	Question 2
Intercept	2.794***	2.463***
	(0.041)	(0.042)
Treatment	−0.0005	0.011
	(0.046)	(0.047)
Price Varies	−0.087*	−0.127***
	(0.047)	(0.048)
Art Activity	0.490***	0.557***
	(0.038)	(0.039)
Over 55	−0.387***	−0.510***
	(0.039)	(0.039)
Treatment:Price Varies	0.158**	0.118
	(0.079)	(0.080)
Treatment:Price Varies:Art Activity	−0.141*	−0.098
	(0.073)	(0.073)
Treatment:Price Varies:Over 55	−0.247***	−0.243***
	(0.077)	(0.074)
Observations	7,620	7,620
Adjusted R^2	0.054	0.077
Residual Std. Error (df = 7612)	1.409	1.423

Note: *p<0.1; **p<0.05; ***p<0.01
Parentheses denote robust standard errors.

This means that subjects over age 55 gave lower ratings when they saw higher prices, even though the overall sample did not. This is possibly due to these individuals developing their personal aesthetic tastes in a period before contemporary art developed the styles popular today, so they may react against art in these styles receiving such high valuations. Alternatively, it could mean that the value of money changes as individuals approach the age of retirement or live off retirement savings. Both of these suggestions are simply theories, and it is hard to know why age would interact with the treatment in this way. It does suggest interesting grounds for further future research regarding the value of money and the attitudes of people of different ages.

Another interesting finding is that Question 1 & Question 2 both have the same sign with regards to the treatment effect. Upon further examination, the two ratings are indeed predictors of each other. We observed a positive relationship between technical quality and enjoyment ratings, indicating that the higher a subject rates the technical ability of an artist, the more likely it is they will enjoy their work, as displayed in Table 5. While this is not a full experiment on its own, it also indicates possible grounds for future research and experimentation.

Table 5: Regressing Question 1 Ratings on Question 2 Ratings

	Dependent variable:
	rating_Q1
Intercept	0.865^{***}
	(0.020)
Q2 Rating	0.797^{***}
	(0.006)
Observations	7,620
R^2	0.663
Adjusted R^2	0.663
Residual Std. Error	0.841 (df = 7618)
F Statistic	$15{,}016.580^{***}$ (df = 1; 7618)
Note:	*p<0.1; **p<0.05; ***p<0.01

V. Conclusion

In this experiment, we did not find evidence overall to suggest that people were influenced by price when determining their opinions on the ability of a painting's creator or their enjoyment of that piece.

If we suspect that people failed to notice the prices in our experiment, as they did during our initial user testing, we would have to be more creative with our design. One possible approach could be to start each survey with a control image and then, only for the control image, ask whether they thought the stated price was fair. This should make certain that subjects paid attention to the prices as they completed the surveys.

While it was not our primary research goal, we did find some evidence to suggest that people over the age of 55 seem to react negatively to high prices when determining their opinions on a piece of contemporary art. Further investigation of this claim could be grounds for future experiments.

Our research suggests other possible follow-up experiments as well in order to further explore the relationship between art and its value. Here are some suggestions for further research:

- To further explore the question of whether price impacts people's perceptions of art, we might try a version of the same experiment that narrows the question to something more specific: if a subject starts by seeing several prices in a certain price range, then sees a higher-priced painting, does the subject perceive that painting to be more enjoyable or of higher quality? If we formulated an experiment that way, we would change our randomization scheme and look for a more focused effect.
- Alternatively, we could focus further research on a particular demographic of individuals likely to purchase a particularly genre of art.
- We might decide to run a similar experiment that tries to simulate an art-viewing experience, by bringing people through an exhibit-like space in partnership with a local gallery, to see if the digital presentation interfered with subjects' enjoyments or perception of value.
- As we did find a statistically significant relationship between people who had taken an art class or participated in an art activity in the last year and their higher-than-average ratings of paintings on

> both questions (as shown in the "Art Activity" row of Table 4). While not a causal finding, it does suggest a potential experiment if we suspect that those art activities increase enjoyment of art. We could explore this further by randomizing subjects into groups and bringing a treatment group to an art lesson or on a visit to a museum, while a control group is provided a placebo experience unrelated to art, then asking both groups to rate their enjoyment of different works of art. This would help us glean whether participation in art experiences actually causes enjoyment of art, versus the highly plausible alternative that folks who are predisposed to enjoy art are the ones who seek out art experiences.

In the end it is interesting, and perhaps inspiring, to note that individuals do not seem to be overly dependent on experts' valuations when determining their feelings towards works of art.