A Causal Reanalysis of "Improving analytical reasoning and argument understanding: a quasi-experimental field study of argument visualization"

Matthew Boggess, Noa Bendit-Shtull

December 11, 2019

Introduction

The 2018 paper *Improving analytical reasoning and argument understanding: a quasi-experimental field study of argument visualization* studies the impact of training in argument visualization on analytical reasoning. As the original study notes, analytical reasoning is a foundational skill for higher level reasoning, yet traditional college educations do not directly cultivate this ability. The authors hypothesized that training students to visualize the logical structure of presented arguments would lead to improvements in argument parsing and generalized analytical reasoning. To investigate this hypothesis, they created a 12 week seminar course for freshman students at Princeton consisting of group and individual practice of visualizing arguments. They recruited additional freshman who did not take the seminar to serve as controls, creating a "quasi-experiment" with a structure much like a randomized experiment, but where treatment assignment was not randomized and instead determined by the class enrollment procedure.

The authors conclude that the seminar in argument visualization had a significant positive impact on students' generalized analytical reasoning skills as measured by an increase in performance on the logical reasoning section of the LSAT. However, a major limitation in the causal interpretation of this outcome is that the study was not a fully randomized experiment. As a result, it is subject to both observed and unobserved potential confounders that could instead explain the results. In this project report, we present a reanalysis of the data from the original study using causal inference techniques learned in class to account for the lack of randomization in the original design.

Study Overview

Study Design

The population studied was made up of 161 freshman at Princeton University between 2013 and 2017; 105 of these students were enrolled in the argument visualization seminar, which was the treatment. Formally, let $Z_i \in \{0, 1\}$ indicate whether unit i receives the treatment (i.e. is enrolled in the seminar).

At Princeton, enrollment in Freshman Seminars is determined via an application process. Students rank their preferences, and submit a short essay explaining why they are interested in their top choice. According to the Princeton website, "All seminar placements will be made through an automated system that gives priority to your preferences... If you are not assigned to any of your preferences, you will be placed on the waiting list for your top-ranked seminar." Faculty who select students from the wait list may consider the students' essays in their decision. In order to minimize selection bias, the 56 control units were recruited from among "individuals who expressed interest in the seminar but were not enrolled due to limited space in the class" (Princeton 2019).

Outcomes

The outcome variable was the change in performance between the LSAT logical reasoning (LR) pre-test and post-test, which was used as a measure of improvement in analytical reasoning ability. Formally, the potential outcomes are $Y_i(1)$ and $Y_i(0)$, the change in LSAT LR performance with and without participation in the argument visualization seminar. Figure 1 replicates Figure 2 in the original paper, which shows the

Table 1: Comparison of pre-test and post-test scores

Seminar	Pre-Score	Post-Score	Change in Scores
0	17.38	17.86	0.48
1	15.90	18.48	2.57

distribution of the change in LSAT LR scores in the treatment and control groups. Table 1 shows the pre, post, and differences in scores between groups.

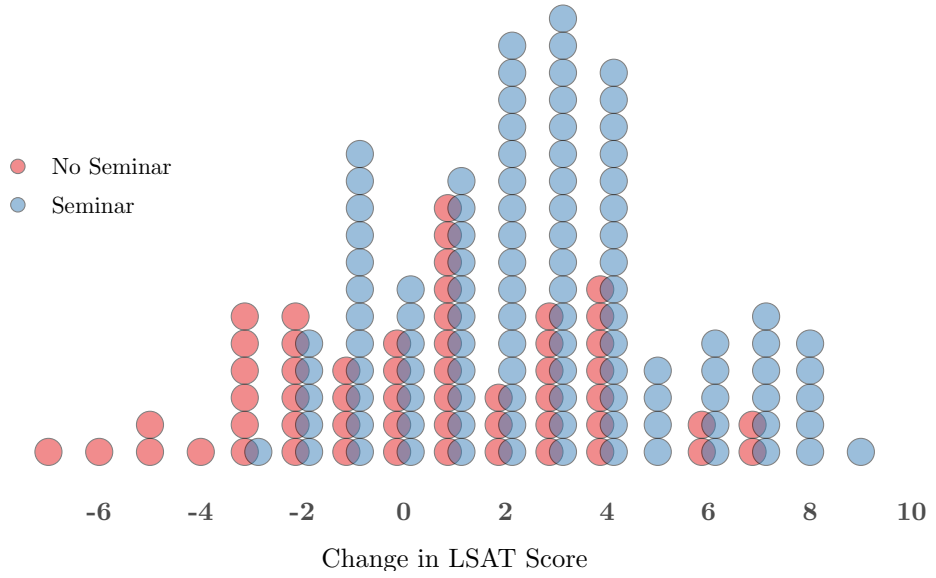

Figure 1: Distribution of Change in LSAT LR scores for seminar and non-seminar students.

Causal Estimand and Key Assumptions

The causal estimand is the average treatment effect of the seminar,

$$\tau^{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0) = \overline{Y}(1) - \overline{Y}(0)$$

Given that treatment assignment was not randomized, we assume strong ignorability in order to conduct our analyses. Specifically, we assume that assignment to the seminar is probabilistic, individualistic, and unconfounded. That is,

$$0 < P(Z_i = 1|X, Y_i(0), Y_i(1)) < 1$$
$$\text{and} \qquad P(Z_i = 1|X, Y_i(0), Y_i(1)) = \pi(X_i)$$

The unconfoundedness assumption means that there are no unobserved covariates that influence the likelihood of enrolling in the seminar and the outcome.

Table 2: Replication of ANCOVA with pre-test score

term	estimate	std.error	statistic	p.value	conf.interval
(Intercept)	7.316	0.946	7.734	0	[5.46 , 9.17]
preLR	-0.393	0.051	-7.725	0	[-0.49 , -0.29]
seminar	1.511	0.421	3.585	0	[0.68 , 2.34]

Replication and Limitations of Original Analysis

The students in the seminar, on average, performed 2.6 points better on the post-test than on the pre-test (t = 9.56, p = 6.41e-16, d = 0.77), while the students in the control group only improved by 0.5 points on average (t = 1.11, p = 0.27, d = 0.11). Finally, the study found that the improvement of seminar students was significantly greater than that of control students (t = -4.31, p = 2.91e-05, d = 0.72).

Next, the authors performed an ANCOVA in order to assess whether controlling for pre-test scores affected the results. The results of the linear model are replicated in Table 2; after controlling for the effect of the pre-test score, the seminar has a highly significant effect. (Note that the pre-test score is highly significant, since a higher score means less room for improvement.)

An F-test (F(1, 158) = 12.85, p = 4.48e-04) confirms that participation in the seminar does explain some of the variance in scores.

The original analysis that we have replicated here suffers from several shortcomings when trying to assess the ATE due to the non-randomized treatment assignment. First, while the ANCOVA adjusts for pre-treatment scores, there are multiple other covariates that were observed that would ideally be accounted and adjusted for to ensure comparability between the treated and control units. Furthermore, there is also a possibility that the effect could be due to some unobserved confounder. While there is no way to control for this, we can use sensitivity analyses to assess how sensitive the results are to a potential unobserved confounder.

Exploratory Data Analysis

As discussed previously, treatment assignment was not random and instead determined by the Princeton Freshman Seminar application process. Thus, in order to make causal claims, we will want to ensure that the covariate distributions between treatment groups are adequately balanced so that the treated and control units are comparable. Additionally, the treatment schedule itself was rather complicated and so we investigate to ensure we restrict to cohorts that received comparable versions of the treatment.

Missing Covariate Data

Unfortunately, we were not able to obtain any covariate data for the 20 control units obtained in the Fall 2013 quarter. Thus we exclude them from all further analysis here leaving us with 36 control units from the Fall 2014 quarter.

One important covariate is the student's score on high school standardized tests. Given that the LSAT is also a standardized test, one could easily imagine a relationship beween SAT scores and improvement on the LSAT LR. Either SAT or ACT scores were provided for all but one of the students. In order to standardize (no pun intended) all of the test scores to a single scale, we converted ACT scores for students who did not provide SAT scores to the median of their corresponding SAT score range using the official conversion table (Fraccia 2016). For the single student who did not provide either, we imputed their score with the mean for their corresponding treatment group.

In addition, one student was missing data indicating whether English was their first language or not and another was missing their age. For both of these, we filled their value with the most common data value from other students in their corresponding treatment group.

Table 3: Covariate imbalance in subsample of data

	std diff	**z-stat**	**p-value**
preLR	-0.2164	-0.9680	0.3330
preFormA	-0.0769	-0.3457	0.7296
preFormB	0.0769	0.3457	0.7296
preSelf	-0.2300	-1.0280	0.3039
preOther	0.0364	0.1638	0.8699
major_type	-0.2851	-1.2700	0.2041
preAge	0.2352	1.0512	0.2931
sex	-0.3340	-1.4825	0.1382
ESL	-0.1861	-0.8341	0.4042
SATfilled	-0.5856	-2.5280	0.0115
preSelf.NATRUE	-0.3593	-1.5911	0.1116
preOther.NATRUE	-0.3593	-1.5911	0.1116

Variation in Exam Forms & Seminar Offerings

The treatment in this study was a rather complicated intervention consisting of an entire quarter of instruction with multiple homework assignments and in-class activities. Furthermore, this seminar was repeated across a span of 4 years with seven different offerings. Thus in order to treat this as a single homogenous level of treatment, we need to assume that this seminar is roughly comparable across all offerings. The top panel of Figure 2 shows the change in LSAT LR scores across seminar offerings. While the effect looks reasonably stable across quarters, there are a few important things to note. First, all of the controls were collected in Fall 2014. This means that we cannot attempt to match within semester and will have to assume comparability across different quarters. Second, there was one seminar that was taught by a different instructor ("Fall 2015 Sub"). While we don't have any quantitive data to verify any differences between this seminar and the others, we were warned by the study author that students spent less time practicing the techniques during this offering. Thus we have excluded the students who took this version of the seminar from further analysis to better ensure homogeneity of the treatment.

Another complication in the study design was that different cohorts used different versions of the LSAT LR assessment. All quarters prior to Fall 2016 took versions A & B of the exam (pre and post ordering was counter-balanced within treatment groups). All quarters from Fall 2016 on took versions C & D. The shift in test versions was made due to the observation that version B appeared to be substantially more difficult than version A. The bottom panel of Figure 2 reflects this observation. Notably, students improve much less when taking the easier version first (version A) and the effect appears to disappear with the other order. Versions C & D seem more comparable to version B based on the outcome. However, since none of the controls received versions C and D, we made the conservative decision to exclude all treated subjects from Fall 2016 onwards from further analysis to ensure we could completely balance the assessment versions.

Covariate Imbalance

Table 3 shows the covariate balance between treated and control students. We see a significant difference ($p = 0.02$) in SAT scores between the treatment group and control group; on average, seminar participants have lower SAT scores. This is an important imbalance, since it's reasonable to expect a positive correlation between scores on the SAT and LSAT.

Figure 3 shows the distributions for the four covariates with the most imbalance between groups. Visually, there is also some evidence of imbalance in sex and major type, where seminar participants are more likely to be female and in non-stem majors.

These imbalances can also be captured in a propensity score, calculated using the covariates pre-seminar

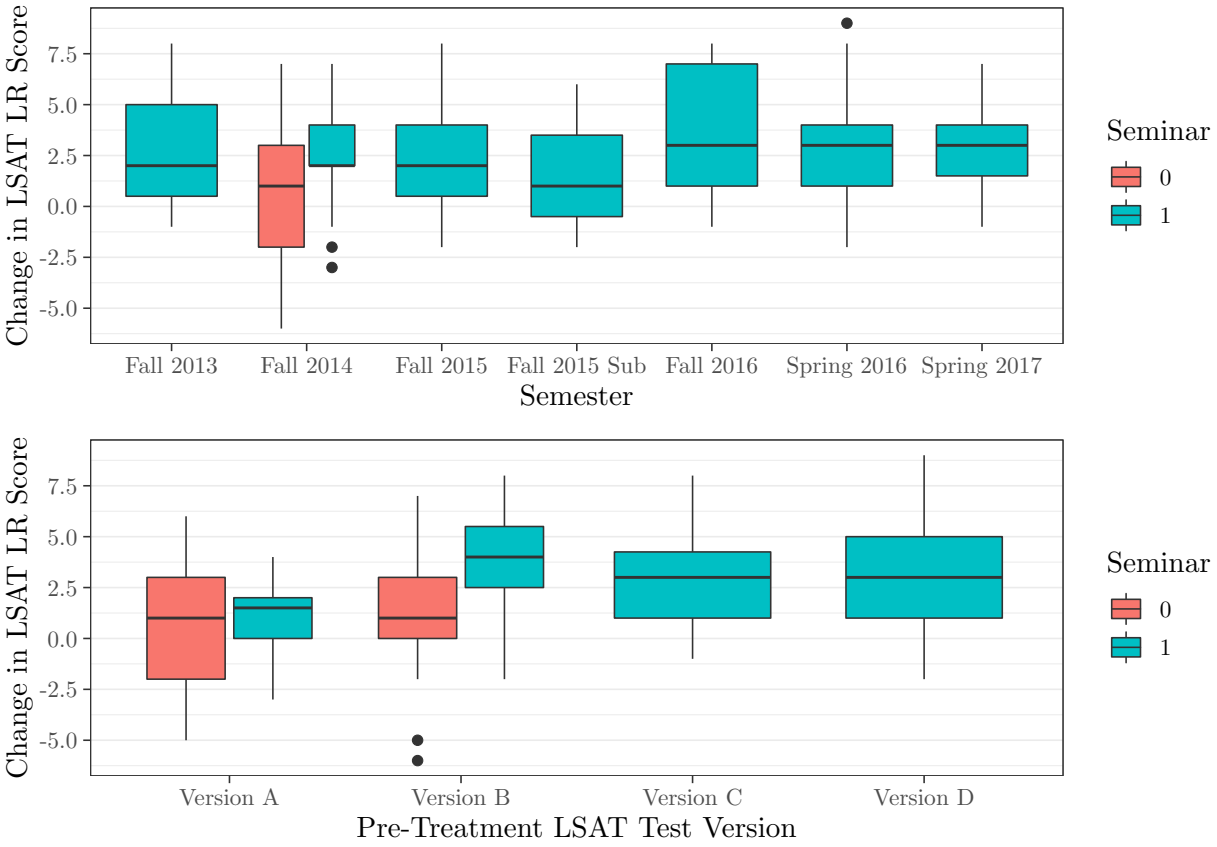

Figure 2: Top Panel: Change in LSAT LR score acros seminar offerings. Fall 2015 Sub refers to an offering taught by a secondary instructor different from the instructor for the other offerings. Bottom Panel: Change in LSAT LR score based on which version of the test they took prior to treatment. Versions A and B were counter-balanced together as well as versions C and D.

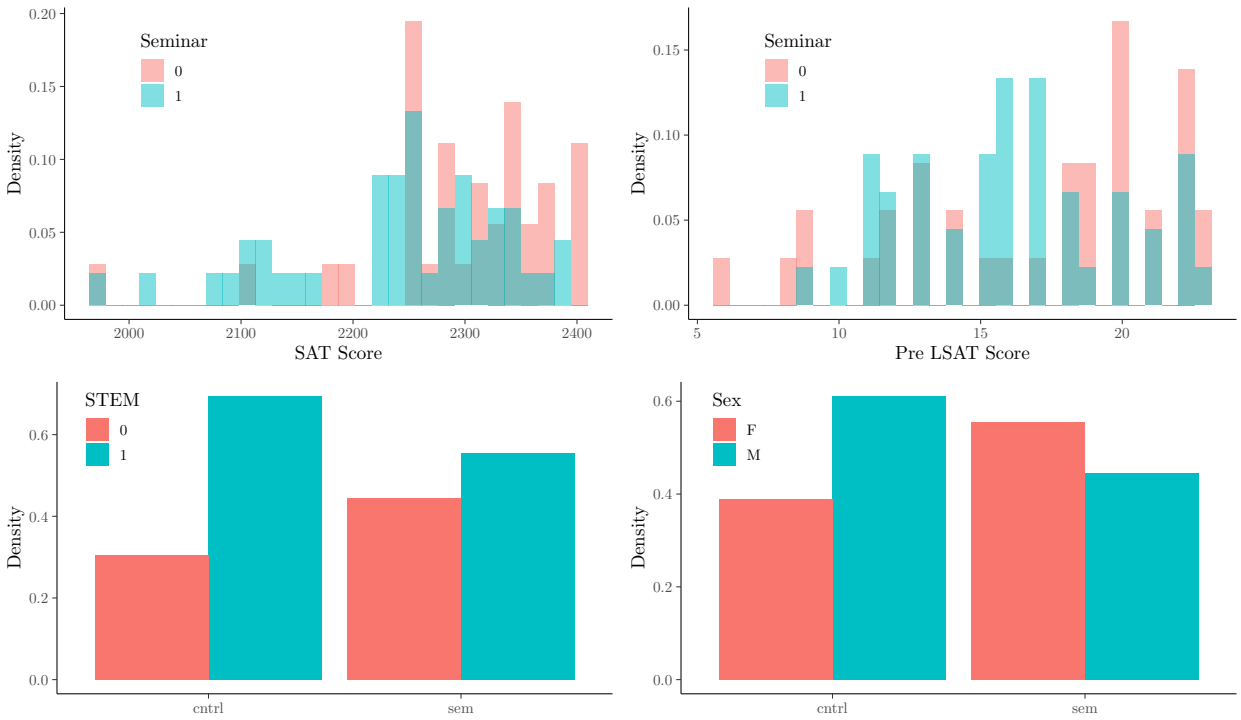

Figure 3: Top four most imbalanced covariates

Table 4: Comparison of pre-test and post-test scores (filtered data)

Seminar	Students	Pre-Score	Post-Score	Change in Score
0	36	16.97	17.75	0.78
1	45	16.07	18.49	2.42

logical reasoning score, pre-seminar LSAT form (A or B), major type (STEM or non-STEM), sex, ESL, and SAT Score. The histogram of propensity scores in Figure 4 shows that these covariates have predictive power for participation in the seminar.

Analysis

As the study was not randomized, we need to ensure that our analysis attempts to balance the two treatment groups to make them more comparable. There are many methodologies that can be used to account for the imbalance in the observed data. Here, we consider a matching analysis, inverse propensity weighted estimators, and a propensity score subclassification analysis and discuss their relative tradeoffs.

Adaptation of Original Analysis

After excluding control units from Fall 2013, treatment units taught by another instructor, and units who received LSAT forms C and D, our sample size is decreased by 50%, from 161 to 81. As a baseline for our analysis, we repeat the t-test and ANCOVA performed in the original paper, but using the smaller dataset.

Among the 81 students in the smaller sample, seminar participants performed 2.4 points better in the post-test than on the pre-test ($t = 6.17$, $p = 1.89\text{e-}07$, $d = 0.7$). The students in the control group only improved by 0.78 points on average ($t = 1.39$, $p = 0.17$, $d = 0.18$). As in the full dataset, the improvement of seminar

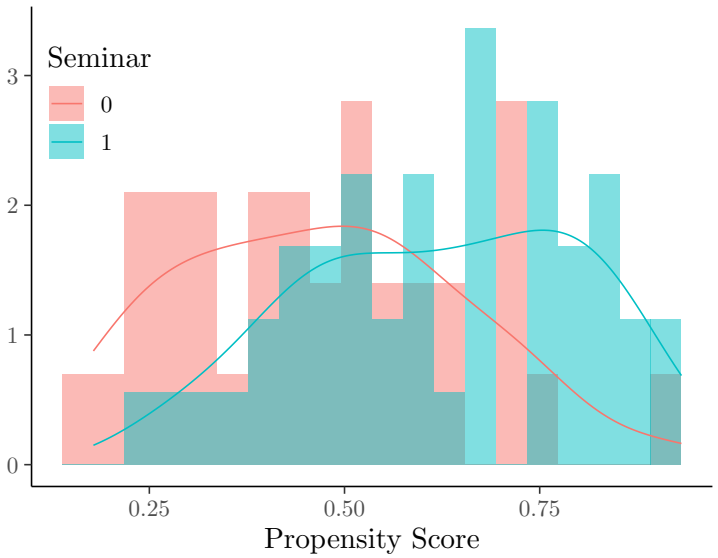

Figure 4: Difference in propensity score distribution between treatment and control

Table 5: Replication of ANCOVA with pre-test score (filtered data)

term	estimate	std.error	statistic	p.value	conf.interval
(Intercept)	7.223	1.229	5.875	0.000	[4.81 , 9.63]
preLR	-0.380	0.068	-5.581	0.000	[-0.51 , -0.25]
seminar	1.301	0.569	2.285	0.025	[0.18 , 2.42]

students was significantly greater than that of control students (t = -4.31, p = 2.91e-05, d = 0.56).

The modified ANCOVA results (presented in Table 5) provide significant, but weaker, evidence that seminar participation had a positive effect on performance improvement, after accounting for the pre-test score.

The modified ANCOVA results provide significant, but weaker, evidence that seminar participation had a positive effect on performance improvement, after accounting for the pre-test score. Similarly, the F-test (F(1, 78) = 5.22, p = 0.03) provides significant but relatively weaker support that seminar participation explains some of the variance in scores. It is expected that the effects will be weaker, given that the sample size is smaller. These results will serve as a baseline for the results obtained after accounting for covariate imbalance.

Matching Analysis

In order to reduce imbalance in observed covariates between treated and control units, we use matching to compare only similar treated and control units. In our case, we have more treated than control units and so we sub-select treated units in order to find the closest matches to our smaller pool of controls. We use the Mahalanobis measure as our distance measure for matching and compute distances using the seven covariates shown in Figure 5. This is a reasonable distance metric to use when we have low dimensional data (recommended to have less than 8 covariates) such as in our case here (Rubin 1980). We use optimal matching as our matching algorithm to match pairs (Rubin 1979).

Standardized differences in our covariates of interest before and after applying matching are plotted in Figure 5. Importantly, we see that the Pre-Test Version is perfectly balanced, which is important since one of the test versions was much easier and we saw this had a substantial effect on our outcome. Additionally, we achieve good balance on the Pre-Test Score and ESL, which are also important variables likely related to

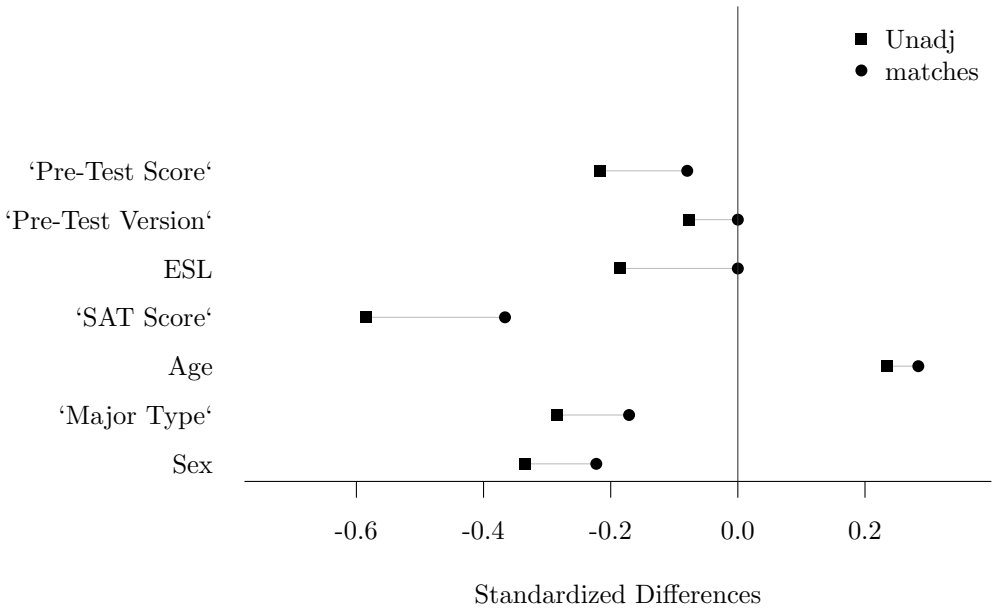

Figure 5: Standardized differences in covariates before and after matching

the outcome. Balance generally improves on the remaining covariates with the exception of age. The most concerning is the remaining imbalance in SAT scores. Achieving better matches appears to be a limitation of small sample size. For example, matching solely to balance SAT scores at most reduces the standardized difference to -0.25.

A common way to account for remaining imbalance in the covariates is to do a regression adjustment using the covariates on the matched data (Stuart 2010). This will help adjust for the remaining imbalances that matching wasn't able to fully smooth out. Results for this regression can be found in Table 6.

We see the main result still holds with an estimated effect of 1.66 points greater improvement in LSAT LR score for students who took the seminar compared to control students ($p = 0.007$). This effect is close to the original effect of 1.5 in the original study and thus in line with the original findings.

A simpler non-parametric analysis approach is to use a paired Fisher randomization test on the matched data to test the strict null hypothesis that there is zero treatment difference for all units. The resulting FRT p-value is 0.004, which is in line with our previous analysis.

Inverse Propensity Weighted Estimator

A second way to account for covariate imbalance is to use inverse propensity weighted estimators. In their comparative analysis of propensity score weighting methods, Lunceford and Davidian describe three versions of inverse propensity weighting (Lunceford and Davidian 2004). The simplest is the Horvitz-Thompson estimator, which weights each unit by it's inverse propensity score, π_i for treatment units and $1 - \pi_i$ for control units. The Hayek estimator also includes estimators for the number of units in each group, $\hat{N}_1$ and $\hat{N}_0$. The third estimator presented in Lunceford and Davidian (2004) achieves a reduction in variance relative to the other two.

The results of the IPW estimators are not as strong as the matching results; the second and third estimators

Table 6: ANCOVA results for matched treated and control units

term	estimate	std.error	statistic	p.value	conf.interval
(Intercept)	-7.019	9.950	-0.705	0.483	[-26.52 , 12.48]
'Pre-Test Score'	-0.406	0.073	-5.529	0.000	[-0.55 , -0.26]
seminar	1.662	0.594	2.800	0.007	[0.5 , 2.83]
'Pre-Test Version'	-1.333	0.579	-2.304	0.024	[-2.47 , -0.2]
'SAT Score'	0.006	0.003	1.688	0.096	[0 , 0.01]
ESL	0.927	0.794	1.167	0.247	[-0.63 , 2.48]
'Major Type'	0.020	0.585	0.034	0.973	[-1.13 , 1.17]
Age	0.087	0.424	0.206	0.837	[-0.74 , 0.92]
Sex	0.511	0.585	0.873	0.385	[-0.64 , 1.66]

Table 7: IPW Estimators

	Point Estimate	Standard Err	t-Statistic	p-Value
Horvitz-Thompson	1.389	0.744	1.867	0.066
Hayek	1.437	0.684	2.103	0.039
IPW3	1.432	0.664	2.157	0.034

are significant at the 5% level, although the Horvitz-Thompson estimator is not. This makes sense given that IPW estimators tend to have high variance, particularly the Horvitz-Thompson estimator; low propensity scores that appear in the denominator are magnified when computing the variance.

Propensity Score Subclassification

Another approach to propensity score estimation of causal treatment effects is stratification by propensity score. Propensity score subclassification has the drawback of being a fairly coarse classification; even within a stratum, there is potential for large variation between units. Theoretically this could be resolved by increasing the number of strata, but we are limited by the fairly small sample size. On the other hand, propensity score subclassification, unlike matching analysis, allows us to use all of our data. This is a non-trivial benefit, given the small sample size. Although Cochran recommends five strata, this results in some strata with as few as two treatment units, which would make the difference in means unreliable (Cochran 1968). Therefore, we use four strata for the propensity score subclassification.

The counts of treatment and control units in each quartile are shown in Table 8. The Neymanian estimator of the stratified difference in means is 1.627, and the estimated variance of the estimator is 0.439. The corresponding 95% confidence interval is [0.3288, 2.9249]. The 95% confidence interval does not contain zero; therefore, Neymanian inference on this stratified design rejects the null hypothesis of no treatment effect for $\alpha = 0.05$.

Table 8: Quartiles of propensity score

Quartile	Treated Units	Control Units	Difference in Means	Variance
1	5	15	-1.40	0.09
2	11	9	2.36	0.06
3	13	7	5.36	0.16
4	16	5	0.25	0.13

As a non-parametric alternative, we conduct a stratified Fisher Randomization Test. The resulting p-value is 0.017, which confirms the conclusion from the Neymanian analysis.

Sensitivity Analysis

So far all of our analysis has attempted to account for imbalance in observed covariates in order to make the treated and control units more comparable. However, another limitation of not conducting a randomized experiment is that the observed effect is susceptible to potential confounding from unobserved variables. While there is no way to correct for this, we can assess how sensitive our result is to potential confounding using a sensitivity analysis. Here we use the senm R function from the sensitivitymult package to conduct an FRT sensitivity analysis on our matched data (Rosenbaum 2007).

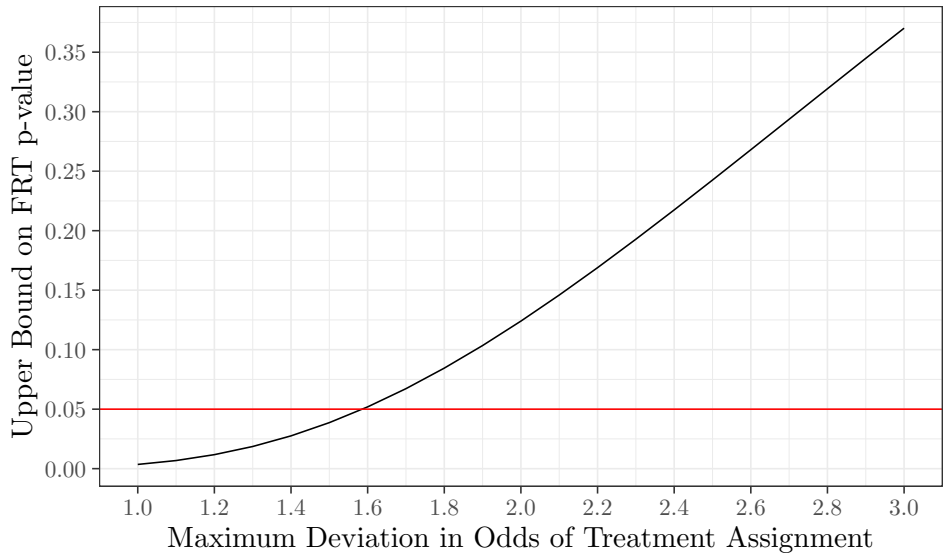

Figure 6: Sensitivity Analysis for Matched Pairs FRT

Figure 7 shows the results of the sensitivity analysis. The interpretation is that, at a significance level of 5%, a difference in likelihood of receiving the treatment of 1.6x would result in a change in our qualitative conclusion (i.e. we reject the null). In other words, these results could be explained by some confounding variable that could have caused treated units to be 1.6 times more likely to receive treatment. This is not an exceptionally high threshold and it is conceivable that such a variable with this effect could exist. For example, maybe students who felt that they especially needed help with logical reasoning were more likely to self select in and it would be reasonable to expect that these students have more to gain from training provided by the seminar.

Discussion

The original study considered here found that training students to visualize arguments via a semester-long seminar resulted in a significant improvment in analytical reasoning as measured by the LSAT LR assessment. The study had a "quasi-experimental" design where treatment assignment was not fully randomized.

In order to validate the causal claim that the visualization training improved students' LSAT LR scores, we conducted several analyses to account for the non-experimental nature of the study. First, in order to ensure that treated and control units were comparable on observed covariates, we compared analyses via matching, inverse propensity weighted estimators, and propensity score subclassification. All of these analyses showed similar results, which in turn were similar to the original study's findings. In particular, all of our estimates of the improvement in LSAT scores due to the seminar ranged from 1.38 to 1.66 points, close to the original estimated effect of 1.5 points. An additional contribution of this paper is a sensitivity analysis to determine

how sensitive the observed results are to potential unobserved confounders. We found that the results were sensitive to a 1.6x difference in treatment assignment probability within matched pairs.

There are several aspects of the current study design that could be adapted to better support causal analysis. Most importantly, a completely randomized design would allow an assessment of the causal effect of the intervention with less worry about unobserved confounders. While randomizing treatment is difficult to do in a university setting, it would be interesting to see how the techniques taught in the class could be repackaged and distributed outside of a formal class setting. In addition, future iterations of this study should investigate the sensitivity of the intervention to differences in instructors and in the assessment form, since both of these seemed to non-trivially impact the outcome. Absent the possibility of a completely randomized design, we recommend collecting a larger sample to improve matching, and collecting control data in each semester of treatment to account for time effects.

References

Cochran, WG. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24. International Biometric Society: 295–313.

Fraccia, Kristin. 2016. "ACT to New Sat to Old Sat Score Conversion Chart." https://magoosh.com/hs/act/act-scores/2016/act-to-new-sat-to-old-sat-score-conversion-chart/.

Lunceford, Jared, and Marie Davidian. 2004. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study." *Statistics in Medicine* 23. Wiley: 2937–60.

Princeton. 2019. "Freshman Seminars." https://odoc.princeton.edu/curriculum/freshman-seminars.

Rosenbaum, Paul R. 2007. "Sensitivity Analysis for M-Estimates, Tests, and Confidence Intervals in Matched Observational Studies." *Biometrics* 63 (2). Wiley Online Library: 456–64.

Rubin, Donald B. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74 (366a). Taylor & Francis: 318–28.

———. 1980. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics*. JSTOR, 293–98.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25 (1). NIH Public Access: 1.