The Effect of Extraversion on College Grades
A Causal Mediation Analysis

Hanchun Jiang, Baichen Tan, Ce Zhang

Dec 09 2022

1 Introduction

Who is more likely to succeed academically? An extravert or an introvert?

The debate on the relationship between personality and academic achievement has been going on for centuries in both education and psychology fields. Previous work has been done by Philip Oreopoulos (2019), who examined the short term impact of personality on the first semester grade using linear regression models. In his research, Oreopoulos used a data set collected at the University of Toronto regarding various aspects of the student's school life and academic performance.

While the original paper aims to find the total effect of extraversion on grades, we take a different route to this question: is the effect of extraversion on grade completely mediated by mediators? Specifically, we want to focus on two potential mediators: study hours (StudyHours) which is measured by the number of hours the student spends for academic purposes per week, and study habits (StudyHabits) which is measured by the frequency the student seeks outside helps from professors, TA, and writing tutors.

Instead of using traditional linear models, we tackled this problem by adopting non-parametric assumptions. Utilizing the same data set as Oreopoulos did, we estimated the Natural Direct Effect and Total Effect of extraversion on grades using a doubly robust estimator. We first identified the Natural Direct Effect of extraversion on grades by blocking different mediator paths. In particular, we calculated 3 different types of direct effects: the direct effect after blocking both mediator paths, the direct effect after blocking the mediator StudyHours, and the direct effect after blocking the mediator StudyHabits. Secondly, we calculated the total effect by estimating the Average Treatment Effect. In both steps, we found negative effects of extraversion on grades. But we are also aware of the fact that our study is limited by model performance and data availability.

2 Traditional Linear Regression Method

The traditional method of estimating causal mediation effect uses the following framework:

1. $Y = \beta_0 + c \times A + e_0$
2. $M = \beta_1 + k \times A + e_1$
3. $Y = \beta_2 + c' \times A + b \times M + e_2$

where Y is the response, A is the treatment, M is the mediator, and e_0, e_1, e_2 are errors. The basic approach follows that we first regress Y over A using formula 1 and obtain the coefficient c. This coefficient measures the total treatment effect of A on Y. We then regress the mediator M over A using formula 2 and obtain the coefficient k. Finally, we regress Y over both A and M and obtain the respective coefficients c' and b. Then the direct effect of A on Y excluding the path through the mediator is c', and the indirect effect of A on Y through the path M is c-c' or kb.

However, while the procedure is simple, the traditional method of linear regression has many limitations that prevent us from acquiring a valid value of the causal effects. Specifically, the path coefficients can be interpreted as causal effects only when:

i. The functional form of each of the models is correctly specified
ii. No confounders between the T-Y relation
iii. No confounders between the T-M relation
iv. No confounders between the M-Y relation
v. No interaction exists between T and M affecting Y, which is often violated by the fact that a treatment may have an impact on the outcome not only through the mediator value but also through changing the mediator-outcome relationship.

However, in our case, since the data is obtained through an observational studies on the first year economics students of the University of Toronto, there exists in reality many confounders between the treatment and the response, as well as between the mediator and the response, which we will examine later in section 5. Given the many limitations above, we decided to estimate the average Natural Direct Effect with a non-parametric model that is suitable for our case.

3 Total Effect and Group Average Natural Direct Effect Identification

3.1 Total Effect

The total effect of the treatment on the outcome Y is the group Average Treatment Effect obtained using the AIPTW method. The formal definition of ATE is written as

$$E_x[E[Y|A=1,X] - E[Y|A=0,X]|A] \tag{1}$$

where Y is the response, A is the treatment and X is the set of confounders.

We can use a $\tau^{A\hat{I}PTW}$ to estimate the ATE

$$\tau^{A\hat{I}PTW} = \frac{1}{n}\sum_i \hat{Q}(1,X_i) - \hat{Q}(0,X_i) + A_i \times \frac{Y-\hat{Q}(A=1,X_i)}{\hat{g}(X_i)} - (1-A_i)\frac{Y_i-\hat{Q}(0,X_i)}{1-\hat{g}(X_i)} \tag{2}$$

Where the g function is propensity score $Pr(A=1|X)$ and the Q function is $E(Y|A,X)$.

In practice, we need to make sure that the propensity score lies within 0 and 1, which we will verify in the following parts. The above method fails when the groups are small. In our data sample, we have 1053 units, which is large enough to employ the method above.

3.2 Natural Direct Effect

One standard way to estimate the Natural Direct Effects in the causal mediation analysis is to adopt cross-world potential outcome to define the group average Natural Direct Effect. The assumptions using this method are less strict than the traditional linear method, which we will discuss shortly. For now, the group average Natural Direct Effect (NDE) is defined as

$$NDE = E[Y_{1M_0}] - E[Y_0] = E_X[E_{M|X,A=0}[E[Y|A=1,M,X]]] - E_X[E[Y|A=0,X]] \quad (3)$$

In this equation, the first term $E_X[E_{M|X,A=0}[E[Y|A=1,M,X]]]$ represents the counterfactual case $E[Y_{1M_0}]$, which means the effect of the treatment A=1 when the effect of the mediators behaves as if the treatment is 0. The second term $E_X[E[Y|A=0,X]]$ represents the effect of the treatment A = 0 while the mediator behaves as $M(A=0)$. In this way, we successfully defined the identification of the Natural Direct Effect.
Expanding the formula, the NDE can be written as

$$\begin{aligned} E[Y_{1M_0}] - E[Y_0] = \sum_X [E[Y|A,M=m,X=x] - E[Y|A=0,M=m,X=x]] \\ Pr(M=m,|A=0,X=x)Pr(X=x) \end{aligned} \quad (4)$$

We notice that in this formula, if we change $Pr(X=x)$ to $Pr(X=x|A=0)$, it does not change the causal question we want to answer. To further explain this point, notice that in the naive estimator

$$\begin{aligned} E[Y_{1M_0}] - E[Y_0] = \sum_X [E[Y|A=1,M=m,X=x]Pr(X=x|A=1) \\ -E[Y|A=0,M=m,X=x]Pr(X=x|A=0)]Pr(M=m,|A=0,X=x) \end{aligned} \quad (5)$$

Conditioned on A, the conditional probability $Pr(X=x|A)$ already expresses the biased information embedded in A (i.e. knowing A already gives us some information about the values of the confounders X). Therefore, we want to cancel out the bias brought by $Pr(X=x|A)$ by using a consistent probability function. One way to achieve this is to use the overall distribution $Pr(X=x)$ to make sure we are integrating over a consistent probability function. However, as long as we use a consistent probability function of X in estimating the causal effect, the causal question doesn't change. Therefore, by changing $Pr(X=x)$ to $Pr(X=x|A=0)$, the causal effect we want to estimate is still valid. Then, we rewrite our new causal estimand as

$$\begin{aligned} \sum_X [E[Y|A,M=m,X=x] - E[Y|A=0,M=m,X=x]]Pr(M=m,|A=0,X=x)Pr(X=x) \\ = E_{M.X}[E[Y|A=1,X,M] - E[Y|A=0,X,M]|A=0] \end{aligned} \quad (6)$$

We observe that now the causal estimand is expressed in a form similar to the Average Treatment Effect on the treated group. Instead of Conditioning on A = 1, we condition on A = 0. This type of estimand is also called Average Treatment Effect on the Untreated (ATU). Notice that this does not mean our estimand on the Natural Direct Effect itself is ATU, but simply that it shares a similar form as ATU. Hence, we use the same method that we use to estimate ATU to estimate our causal

estimand on the Natural Direct Effect by conditioning both on the mediators and the confounders. To put it formally, we use the formula

$$\frac{1}{n}\sum \frac{A_i(1-g(X,M_i))}{(1-Pr(A=1))(g(X,M_i))}(Y-\hat{Q}(1,M_i,X_i)) - \frac{1-A_i}{1-Pr(A=1)}(Y-\hat{Q}(1,M_i,X_i)) \quad (7)$$

Where the g function is propensity score function $Pr(A=1|X,M)$ and the Q function is $E[Y|A,X,M]$. In practice, we need to make sure that the propensity score lies within 0 and 1, which we will verify in the following parts. Again the above method fails when the groups are small, but our data sample is large enough to employ the method above.

3.3 Natural Indirect Effect

After obtaining the Natural Direct Effect (NDE) and the total effect (TE), the Natural Indirect Effect (NIE) of the treatment A through the mediator M is NIE = TE - NDE.
In our case, because we have two mediators, StudyHours and StudyHabits, we estimated three types of Natural Direct Effects:

1.Natural Direct Effect excluding the paths of both mediators

$$E_{M_1,M_2,X}[E[Y|A=1,X,M_1,M_2] - E[Y|A=0,X,M_1,M_2]|A=0] \quad (8)$$

2.Natural Direct Effect excluding the path of StudyHours

$$E_{M_1,X}[E[Y|A=1,X,M_1] - E[Y|A=0,X,M_1]|A=0] \quad (9)$$

3.Natural Direct Effect excluding the path of StudyHabits

$$E_{M_2,X}[E[Y|A=1,X,M_2] - E[Y|A=0,X,M_2]|A=0] \quad (10)$$

After calculating the respective Natural Direct Effects, we then use NIE = TE - NDE to calculate the Natural Indirect Effect, both through StudyHours and through StudyHabits. The non parametric methods we employed here has the advantage of allowing us to use flexible machine learning methods.

4 Identification of the Causal Problem

Our data set consists of first-year students taking the introductory economics class in the fall semester of 2016-2017 at the University of Toronto. We apply the identification method mentioned above to our research question:

1. Treatment A is a binary indicator of the extraversion level of a student (1=extraverted, 0=introverted), which is obtained through a mandatory personality test assigned at the beginning of 2016-2017 fall semester. Specifically, the test assigned each individual a score from -3 to 3. A score 0 means that the person is neither extraverted nor introverted, a positive score means that the person

is extraverted and a negative score means that the person is introverted. We used 0 as a cutoff to make the treatment A a binary variable. Note that while we used the entire data set for model training, in the prediction stage, we excluded the individuals with moderate extraversion scores to satisfy the "no hidden variations of treatment" assumption. This is discussed in detail in section 5.

2. Outcome Y is the student's fall semester average grade over a 100 scale.

3. The first mediator M1 is study hours, which is measured by a survey distributed at the end of the fall semester asking students to self-report the number of hours they spent on study each week.

4. The second mediator M2 is study habits related to one's extraversion. Since extraversion is defined as one's tendency to interact with the outside world, we focus on study habits involving seeking outside help. Here, study habits is defined as an average of the tendency to seek free tutor's help, the tendency to meet with the instructors, and the tendency to seek writing help. All these tendencies were collected in the survey distributed at the end of the fall semester.

5. The confounding variables for A and Y are mother's education and father's education. This information was collected in the survey at the beginning of the fall semester regarding the years of education the parent received.

6. The confounding variables for mediators and Y are stress level, depression level, and motivation (recognition of the importance of academic success, and expectation to apply to graduate schools). The expectation to apply to graduate school was collected as a binary variable at the beginning of the semester (1=applying). All the other variables were collected at the end of the semesters retrospectively regarding the student's subjective experience throughout the semester.

The identification and inclusion of the confounding variables is discussed later in section 5.

5 Assumptions

Compared to the conventional linear regression method, we identified the average Natural Indirect Effect and average Natural Direct Effect under relatively less strict conditions using the method mentioned above. The list of assumptions we made is presented below:

5.1 Consistency Assumptions

No interference
We assume each individual unit to be independent of each other. In other words, the grade outcome of a student does not depend on other students' extraversion. We based this assumption on the fact that the University of Toronto is a large public university and the size of each section is large, which does not facilitate strong interactions between students. Furthermore, the students in our data set share only the introductory economics class, but not necessarily other classes and extracurricular activities, which again dilutes the interference between individuals.

No hidden variations of treatment
We assume that for each unit, there are no different forms or versions of treatment. In particular, we assume that the behaviors of extraverted students don't vary significantly among themselves. In

order to satisfy this assumption, we excluded individuals with a moderate extraversion score. Specifically, we excluded standardized extraversion scores between -0.5 and 0.5 in order to make both the extraverted population and introverted population more representative in terms of personality. we were left with a reasonably large data set after this step (669 units). We further assume that after excluding the moderate extraversion scores, the extraverted group and the introverted group behave uniformly amongst themselves given their relatively similar background in the same university.
In mathematical language, we write these assumptions as:

a. $Y = Y_{am}$ if A = a and M = m for m in the support of the observed M given C, $A = a'$
b. $Y_{aMa'} = Y_{am}$ if $Ma' = m$

5.2 Sequential Ignorability (Conditional independence assumption)

We need the Conditional Independence Assumptions to identify important confounders and ensure that no unobserved confounder has been neglected. We need to satisfy two conditional independence assumptions:

a. Conditional Independence of A: $Y(a', m), M(a) \perp\!\!\!\perp A|X$ for all $a, a' \in 0, 1$ and $m \in M$.

A is the treatment, and in our case, whether a person is extraverted or not was measured by the Big-5 personality test. M is the mediator, which in our case is StudyHours and StudyHabits. X is the set of confounders. In other words, we need to ensure that no unobserved confounders jointly affect A and M, Y.

b. Conditional Independence of M: $Y(a', m), M(a) \perp\!\!\!\perp A = a|X = x$ for all $a, a' \in 0, 1$ and $m \in M$ and $x \in X$.

In other words, conditioned on A and X, no unobserved confounders jointly affect M and Y.
In order to satisfy conditional independence assumption, we need to examine potential confounders in both a and b category so as not to neglect important confounding variables. First, we think about conditional independence (a). To find confounders that affect A, M, and Y, we introduce the following assumption:

We consider the personality scores measured by the Big-5 Personality test to be independent of each other. Importantly, we assume that the personality test assesses each personality as a fixed variable not affected by variables other than prior socio-economic status. This is because, although personalities may vary from childhood to puberty, the personality test measures the students' personality traits at the time of freshman year, when the students' mentalities have developed quite maturely and the personalities are relatively stable. Besides, the research was conducted over the fall semester, which was a relatively short range of time. We assume, on average, no significant shift of personality occurred.

Based on the assumption above, we assume that no variables other than prior socio-economic/family background affect our exposure variable extraversion. Hence, in order not to neglect important confounders that affect A, M and Y, we need to consider confounders that represent each student's socioeconomic/family background. In our study, we use parent's education (separated into father and mother education considering that some students may come from one-parent families) to represent students' socio-economic background. Students whose parents receive a higher level of education

are likely to come from more affluent backgrounds that support them with appropriate educational resources, and are therefore more likely to earn a good grade.

Next, we identified confounders that affect M and Y. To accurately capture all the important confounders, we divided potential confounder candidates into three categories: socio-economic background, psychological factors, and academic level factors. We have discussed socio-economic confounder and will discuss how we found confounders in the second and third categories.

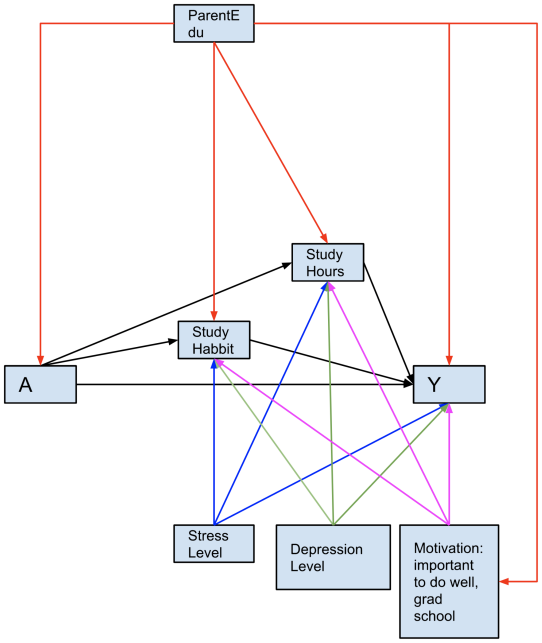
We propose three most important confounders in the psychological factors: stress level, depression level, and motivation score. We assume that for first year college students, the above three variables capture the majority of potential psychological factors that affect both the mediators (StudyHours and StudyHabits) and the response (grades). We also want to point out that there is no causal relationship between extraversion and stress level/depression level/motivation. Extraversion is collected before the other covariates are collected and therefore the other variables cannot be causal to extraversion. We also assume that extraversion has no effect on stress level/depression level, since there are inadequate psychological studies showing a correlation between one's personality type and mental health. Furthermore, motivation is measured as one's expectation of going to graduate school and one's recognition of the importance to do well in college, which is therefore only affected by parents' education but not one's personality.

For the academic factors, we assume it is not a confounder because the levels of course difficulty of the students of our sample size are on average uniform. All students are first-year students taking introductory economics class. For freshmen we expect the course difficulty to be medium and there is not yet severe division among different majors.

By the above reasoning, we identified the important confounders to be adjusted.

6 Causal DAG and Identification

Based on all the assumptions above, we propose the following causal DAG:



7 Model Fitting

As we have discussed earlier, one advantage of the non-parametric identification method is we can use flexible machine learning models to fit the nuisance functions. In the figures below, we compare several different outcome and treatment models for our data set.

For all the models, we used a set of confounders for various causal effects we want to estimate. For estimating the Total Effect, we only need to consider the socioeconomic confounders, which is measured by father and mother education. For estimating the Natural Direct Effect, besides parent's education, we need to consider confounders including depression level, stress level, and motivation that affect both the mediators and the response.

For each outcome model, we calculated its cross validated mean squared error using 5 folds, and for each treatment assignment, we did the same with its cross entropy. The figure below shows the fit diagnostics of each model.

	Outcome	Mediators	Model	Q model MSE	Q model baseline	g model CE	g model baseline
1	Fall_grade	none	RandomForest	152.611823	151.601951	0.696954	0.692866
2	Fall_grade	none	LogisticRegression	149.51816	151.601951	0.690182	0.692866
3	Fall_grade	none	XGBoost	157.875933	151.601951	0.714846	0.692866
4	Fall_grade	both	RandomForest	169.400559	182.096039	0.682537	0.693418
5	Fall_grade	both	LogisticRegression	168.510748	182.096039	0.677284	0.693418
6	Fall_grade	both	XGboost	181.723137	182.096039	0.711953	0.693418
7	Fall_grade	StudyHours	RandomForest	121.740535	130.438943	0.688138	0.693117
8	Fall_grade	StudyHours	LogisticRegression	121.582864	130.438943	0.690446	0.693117
9	Fall_grade	StudyHours	XGBoost	127.014899	130.438943	0.707408	0.693117
10	Fall_grade	StudyHabits	RandomForest	146.637146	149.82793	0.692113	0.692866
11	Fall_grade	StudyHabits	LogisticRegression	160.548133	149.82793	0.693429	0.692866
12	Fall_grade	StudyHabits	XGBoost	155.398148	149.82793	0.708847	0.692866

Among the three models, the logistic regression model fits better than the other models. In general, the models fit well for Q. However, all the models does not fit as well for the g model, though the differences are small. This again demonstrates the value of defining the causal estimand in a non-parametric fashion. This problem may be caused by unobserved confounders, leading to the problem of under-fitting. The poor predictive accuracy of the treatment models causes concerns: the asymptotic results for the double machine learning models rely on consistent estimators for the nuisance function. We are unable to detect immediate remedies to improve the predictive performance of the models given the limited number of covariates. As we continued the causal estimation process, we kept in mind that the results may be biased by the poor estimation of the nuisance functions.

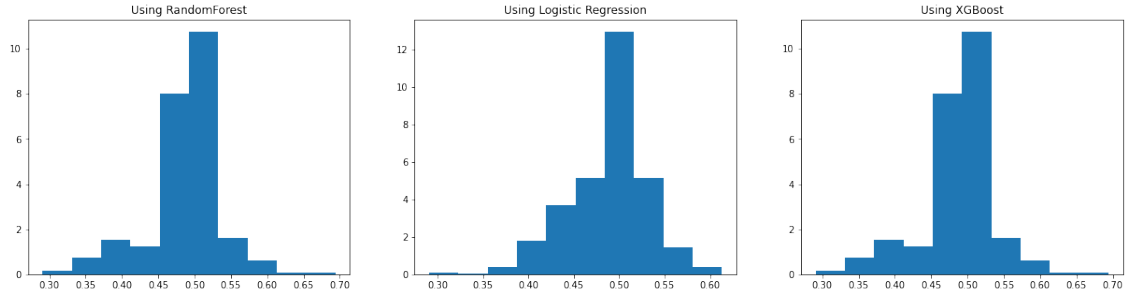
8 Overlap Condition

We need the overlap condition to be satisfied in order to identify the Total Effect and Natural Direct Effect after excluding the mediators, which are estimated by ATE and ATU respectively.

For all the histograms below, we see that the overlap condition is satisfied when using all the three models Random Forest, Logistic Regression, and XGBoost, since the distribution of the propensity scores is not skewed, and the majority of the scores lies around 0.5.

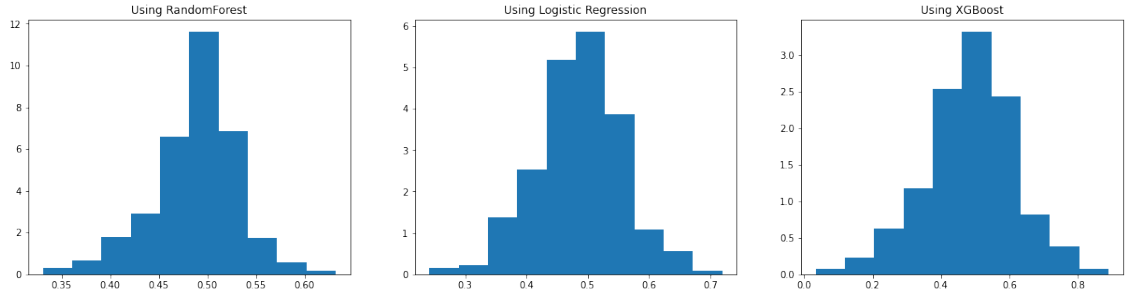
8.1 Total Effect

The histogram below shows the distribution of the propensity scores $Pr(A = 1|X)$ where X is the socio-economic confounder (i.e. Parent's education).



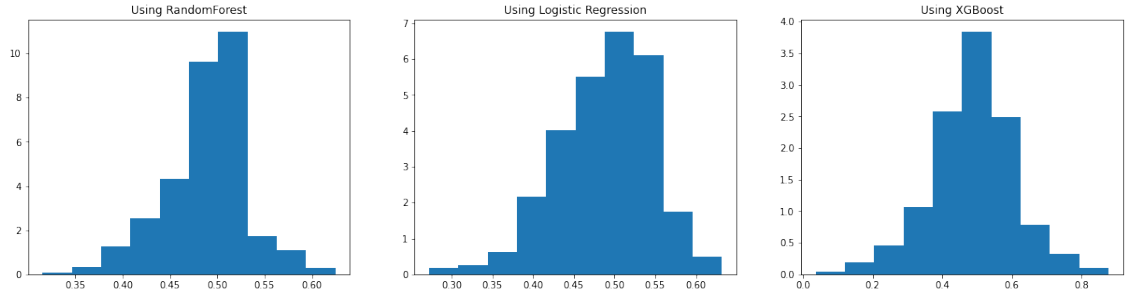
8.2 NDE excluding both StudyHours and StudyHabits

The histogram below shows the distribution of the propensity scores $Pr(A = 1|X, M1, M2)$, where the set of confounders X now includes both the socioeconomic confounders as well as confounders that affect both the mediators and the response Grade, such as stress level, depression level, and motivation. Besides conditioning on the set of confounders, We also need to condition on both mediators when calculating the propensity scores.



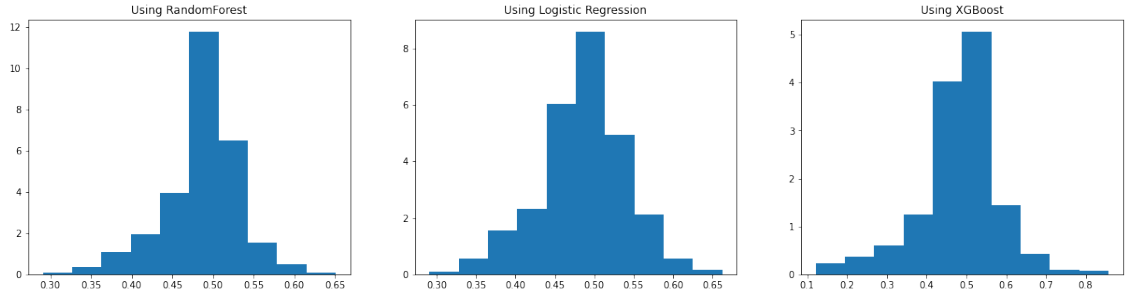
8.3 NDE excluding StudyHours

The histogram below shows the distribution of the propensity scores $Pr(A = 1|X, M1)$, where the set of confounders X now includes both the socioeconomic confounders as well as confounders that affect the mediator StudyHours and the response Grade. Besides conditioning on the set of confounders, We also need to condition on the mediator StudyHours while calculating the propensity scores.



8.4 NDE excluding StudyHabits

The histogram below shows the distribution of the propensity scores $Pr(A = 1|X, M2)$, where the set of confounders X now includes both the socioeconomic confounders as well as confounders that affect both the mediator StudyHabits and the response Grade. Besides conditioning on the set of confounders, We also need to condition on the mediator StudyHabits while calculating the propensity scores.



9 Results

9.1 Total Effect

After plugging our estimated nuisance functions in to the standard double machine learning estimator, we obtained the estimated results of the Average Total Effects of extraversion on the grade. The results are displayed in the table:

	Outcome	Type of causal effect	Model	Estimate	p/m	Standard Error
1	Fall_grade	Total Effect	RandomForest	-2.201365	1.669351	0.85171
2	Fall_grade	Total Effect	Logistic Regression	-2.209192	1.635284	0.834328
3	Fall_grade	Total Effect	XGBoost	-1.862314	1.825616	0.931437

We notice that the estimated values of NDE are all negative and are around -1.86 to -2.21, with confidence intervals not covering 0. Thus regardless of the model we chose, we got the same qualitative conclusions. Since the confidence intervals do not cover 0, we conclude that there exists some negative effect of extraversion on grades. However, the upper bound of the confidence interval is close to 0, which means that the actual total effect is weak.

The standard errors are small for all the models, which relates to the previous section when we checked all the overlap conditions holding true for all three models. The small standard error also means that our sample size is large enough to estimate a causal effect.

9.2 Natural Direct Effect

After plugging our estimated nuisance functions into the standard double machine learning estimator, we obtained the estimated results of the average Natural Direct Effects of extraversion on the grade. The results are displayed in the table below. Notice that Natural Direct Effect is a relative

concept. By Natural Direct Effect we mean the Natural Direct Effect of extraversion on grade without considering the path of study hours, of study habits, or both.

	Outcome	Type of causal effect	Mediator excluded	Model	Estimate	p/m	Standard Error
1	Fall_grade	Natural Direct Effect	Study Hours	RandomForest	-1.837253	1.609899	0.821377
2	Fall_grade	Natural Direct Effect	Study Hours	Logistic Regression	-1.698285	1.598712	0.815669
3	Fall_grade	Natural Direct Effect	Study Hours	XGBoost	-2.571081	2.448955	1.249467
4	Fall_grade	Natural Direct Effect	Study Habits	RandomForest	-2.114952	1.631117	0.832203
5	Fall_grade	Natural Direct Effect	Study Habits	Logistic Regression	-2.057961	1.631844	0.832574
6	Fall_grade	Natural Direct Effect	Study Habits	XGBoost	-2.811477	2.388753	1.218751
7	Fall_grade	Natural Direct Effect	Study Hours and Study Habits	RandomForest	-1.822925	1.602896	0.817804
8	Fall_grade	Natural Direct Effect	Study Hours and Study Habits	Logistic Regression	-1.754583	1.611691	0.822291
9	Fall_grade	Natural Direct Effect	Study Hours and Study Habits	XGBoost	-2.373297	2.57755	1.315077

The estimated NDE values excluding the mediator path StudyHour using the Random Forest, Logistic Regression, and XGBoost are -1.83, -1.69, and -2.57. The estimated NDE values excluding the mediator path StudyHabit using the Random Forest, Logistic Regression, and XGBoost are -2.11, -2.05, and -2.81. And the estimated NDE values excluding both mediators using the three models are -1.82, -1.75, and -2.37 respectively.

We notice that the confidence interval does not cover 0 for all the estimated values of the Natural Direct Effects, regardless of the model we choose.

This means that for the three scenarios below
1. When we excluded the mediator path StudyHour,
2. When we excluded the mediator path StudyHabit,
3. When we excluded both mediator paths
there exists some negative effect of extraversion on grade Y.

However, like the Total Effect, although the confidence intervals do not cover 0, the upper bound of the confidence interval is extremely close to 0, which means that the Natural Direct Effect is very weak. Furthermore, the standard errors of the estimated results are small, which means that our data sample size is large enough to identify a causal effect.

9.3 Natural Indirect Effect

After obtaining the Natural Direct Effect and the Natural Indirect Effect, we used the formula NIE = TE - NDE to get the values of Natural Indirect Effects. The results are presented in the table below.

	Outcome	Type of causal effect	Through which Mediator	Model	Estimate
1	Fall_grade	Natural Indirect Direct Effect	Study Hours	RandomForest	-0.364112
2	Fall_grade	Natural Indirect Direct Effect	Study Hours	Logistic Regression	-0.510907
3	Fall_grade	Natural Indirect Direct Effect	Study Hours	XGBoost	0.708768
4	Fall_grade	Natural Indirect Direct Effect	Study Habits	RandomForest	-0.086413
5	Fall_grade	Natural Indirect Direct Effect	Study Habits	Logistic Regression	-0.151231
6	Fall_grade	Natural Indirect Direct Effect	Study Habits	XGBoost	0.949163

We see that for the Natural Indirect Effect of extraversion on grade through either StudyHours or StudyHabits, the estimated value is positive when using XGBoost, but becomes negative when using Random Forest and logistic Regression.

To sum up, there exists some negative effect of extraversion on grades, and the effect is not completely mediated through the mediators StudyHours and StudyHabits. However, both the Natural Direct Effect and the Total Effect are very weak.

10 Unobserved Confounders

Our approach based on the assumption that the parent's education is a valid variable that measures the socio-economic status of the student. However, there may exist confounders other than socio-economic status that we failed to identify. Furthermore, it is unclear whether parent's education is the sole determinant of socio-economic status. Other factors such as family income, community, and prior educational resources can have an impact on the student's social well being. In the following sections, we discuss methods to validity the soundness of our research regarding potential unobserved confounding variables.

10.1 Placebo Check

Placebo test is used to determine if there are unobserved confounding variables and to test the general soundness of the research design. In order to perform a placebo check, we first identified a "placebo treatment" that is confounded with the outcome in a fashion similar to the true treatment of interest, but which has no causal effect on the outcome. In our placebo check design, we chose the personality trait "agreeableness" as the placebo treatment. Similar to extraversion, agreeableness is another trait in the big-5 personality traits. It is therefore confounded similarly as extraversion in terms of how it is influenced by socioeconomic factors, family background, global trends, etc. Furthermore, we assumed agreeableness to be independent of the outcome, which is the students' grades. We based this assumption on the fact that the students were taking economics classes and were likely to be in STEM majors, which depend more on one's logical thinking and less on one's social ability (i.e. agreeableness) compared to sociology/humanities majors.

Like extraversion, agreeableness is converted into binary forms, and we estimated the total effect of agreeableness on grades using the same estimation methods through the three models. Similar to extraversion, Individuals with moderate agreeableness scores were also excluded in the prediction

stage. The outcome is displayed in the following table:

	Outcome	Type of causal effect	Model	Estimate	p/m	Standard Error
1	Fall_grade	Placebo Effect on Agreealeness	RandomForest	-4.735492	8.943276	4.562896
2	Fall_grade	Placebo Effect on Agreealeness	Logistic Regression	0.473542	1.825678	0.931469
3	Fall_grade	Placebo Effect on Agreealeness	XGBoost	0.189305	1.880532	0.959455

For all the three models, the confidence intervals contain 0 and the results are therefore insignificant, which means agreeableness does not have an effect on grades, just as we predicted. Therefore we concluded that the research design correctly captured the causal relationships between different covariates, taking the majority of confounding variables into consideration.

10.2 Sensitivity Analysis

Since the Placebo analysis already checked that there were no significant unobserved covariates that we failed to take into account, we did not proceed with a sensitivity analysis. It is sufficient to say that there is nearly no unobserved confounding variables after applying the placebo check. Furthermore, sensitivity analysis is used to identify how strong an unobserved confounder should be, if it exists, to undermine the identified causal effect. Since only a extremely weak causal effect was identified, there was not much to undermine and a sensitivity analysis didn't add value to our study.

11 Limitations

While we were able to estimate the Total Effect and Natural Direct Effect of extraversion on grades, we noticed that the estimation method had certain limitations.

11.1 Model Fitting

In our study, while the Q model fitting has small MSE, it is not the case for the g model. The poor performance of the models in estimating g score may be a result of the way the data was collected. Some of the most important variables of the data were collected through surveys, which were highly subjective. For example, in the raw data set, there were units with StudyHours more than 50 hours per week and many units with StudyHours less than 5 hours per week (414 units among 1053 units). The strong oscillation doesn't seem reasonable. Therefore, the data does not perfectly reflect the true effect of the mediator on the treatment. The results we obtained can be biased and are unable to produce the most accurate analysis of the effect.

11.2 Consistency: student interaction

We made the assumption that the interaction between our sample units were small enough to be neglected based on the large population size and class size of the University of Toronto. In reality, there exist, to some degree, interactions between the students both in class and in extra-curriculum activities.

11.3 Binary treatment of extraversion

We proposed there were not much variation within the treatment and therefore set extraversion as a binary variable. In reality, the extent to which an individual is extraverted or introverted differs a lot. Even though we excluded units that were moderately extraverted or introverted when estimating the causal effect, there exist various levels of extraversion and introversion among the rest of the individuals.

11.4 Consistency of personality in the long term

We built our conditional independence assumption by assuming that personality is a fixed variable unlikely to drastically change within the period of our study. However, if we want to examine the causal effect of extraversion on academic performance throughout the four years of undergraduate education, this assumption no longer holds. Personality traits may change as a result of other variables such as academic pressure, social circles, etc.

11.5 Exhaustive mediators

There could be unobserved mediators between extraversion and grades. Extraversion can affect grades in ways other than study habits and study hours. For example, extraversion may have an influence on one's social relationships, which in turn impacts the kind of outside support resources available. Taking unobserved mediators into consideration, the Natural Direct Effect we identified is a "relatively" direct effect independent of study hours and study habits. It is possible that it does not accurately indicate the absolutely unmediated effect of extraversion on grades.

11.6 NIE no confidence intervals

We obtained the Natural Indirect Effect using the following equation: NIE = TE - NDE, instead of directly estimating the Natural Indirect Effect. Our approach prevented us from getting the confidence intervals of the Natural Indirect Effect. Thus we are unable to identify whether the confidence intervals cover zero, which indicates whether the Natural Indirect Effects do exist. Therefore, we could only suggest a positive or negative Natural Indirect Effect, but were unable to verify the soundness of our conclusion.

12 Conclusion

In this project, we set up a non-parametric framework for estimating the Natural Direct Effect and Total Effect of extraversion on grade. By adopting different machine learning models, we find some weak negative Natural Effect and Total Effect of extraversion on grades. However, because the upper bound of the confidence interval we obtained are all close to 0, the existence of actual Natural Direct Effect and Total Effect of extraversion is ambiguous.

Because of the limitations we identified in the previous part, our study is unable to provide a definitive conclusion on the effect of extraversion on the grade. Nevertheless, we believe this approach is informative in estimating the mediating and natural effects of personality traits on academic performance. In future studies, if complemented with more objective data sources and a more thorough examination of confounders, we will be able to draw stronger conclusions using the same approach.

Regardless of the actual effect identified, the results will become a crucial guidance on how education systems should be designed to empower students of different personalities.

13 Code Resources

All codes can be found at the following link

14 Reference

Oreopoulos, Philip, What Limits College Success? A Review and Further Analysis of Holzer and Baum's Making College Work. *Journal of Economic Literature*, 2021.