THE VALUE-ADDED OF STUDENTS' EXPECTATIONS FOR FUTURE JOB CATEGORIES ON THEIR PERFORMANCE

SUPERVISOR:

Prof. Victor Gay

STUDENT:

Zhihan Zhang (Student Number 21613596)

Contents

Toulouse, September 25, 2022

1 INTRODUCTION

Under the topic of student achievement determinants, researchers generally recognized that students' gender, learning time, socioeconomic status, and emotional factors are essential variables. Another factor is still not given the same attention as the others: whether students' aspirations (willingness to pursue higher education or future career choices) impact student performance, under the premise of considering the above factors.

Career aspirations are defined as personal expressed career-related goals or choices under ideal conditions and are identified as a determinant of coursework and educational goals determining an individual's future career pursuits (Rojewski (2005)). In Khattab (2015), the authors argue that the definitions of career aspirations and expectations need to be distinguished. The former leans toward ″what the individual hopes will happen in the future″, while the latter leans toward ″what will happen to the individual believes in the future″. Moreover, in his view, neither students' desires nor expectations are predictors of future educational behavior, but various forms of matching between the two. The authors, therefore, studied the impact of ″mismatched aspirations and expectations″ on the performance of British adolescents. Khattab (2015) does not deny the role of aspirations in improving student achievement. The author believes that ″aspirations can arguably help students improve their achievement″, but if it is accompanied by ″high expectations″, the improvement will be even more considerable.

We use the PISA 2015 and 2018 databases in this thesis, an international assessment project conducted by the Organisation for Economic Co-operation and Development (OECD). The project assesses the literacy of 15-year-old students in reading, mathematics, and science at various educational institutions. Due to data limitations, the variable ″career expectations″ is missing, while the variable ″career aspirations: jobs students expect to have at age 30″ is available. We are interested in two treatments: ″Do students have clear career expectations″ and ″Do students expect high-skilled level careers″. Many high-skilled careers, such as science, engineering, and technicians, are science-related. According to the PISA report, researching student interest in science-related careers is important given the impact of science and technology on our daily lives and the expected growth in science-related employment globally. Science-related careers are often inseparable from science literacy. Therefore, we focus on whether students' future career aspirations impact the current scientific literacy in this thesis.

The analysis consists of the following four parts. We explore the results in various literature on this or similar questions in the first section. The second part is data description. In the third part, we show different identification strategies, OLS, PSM, IV, and their assumptions. Finally, we display the results of each identification strategy. Furthermore, robustness tests and heterogeneity analyses are performed for each method.

2 LITERATURE REVIEW

Khattab (2015) applied a multi-level regression model to study how ”mismatched aspirations and expectations” could influence student achievement. The results show that students with high aspirations and expectations achieve higher grades in school than students with low aspirations and low expectations. Moreover, when a student’s aspiration or expectation is high, the achievement is more than half of that of a student who is low in both. Gutman and Schoon (2012) applied structural equation modeling. The study showed that teenagers with uncertain career aspirations had a higher academic performance at age 16 than those with higher, definite aspirations when students’ background were considered. They also tested various indirect effects. Unlike the above studies, this thesis identifies the impact of student aspirations on student performance more directly.

Some studies suggest that academic achievement has an impact on career aspirations. However, our thesis suggests a causal relationship between the two contrary to those papers. Mau (2003) applied logistic regression to analyze the data of the National Educational Longitudinal Survey of 1988 and show that academic proficiency has a significant positive effect of 0.03 on the persistence of students’ career aspirations in science and engineering (SE) professional careers. Some other researchers found that improvements in students’ science performance significantly affected science-related career aspirations (Park, Khan, and Petrina (2009)). These studies do not take into account that the career aspirations that students have are highly likely to be influenced by social norms and parents’ guidance rather than performance and that different career aspirations produce different levels of performance, so that the career expectations were established even before they show their academic performance. However, these papers inspire us that the effect of performance on career aspirations is also one of the causal relationships that can be considered, and such simultaneity may be a potential endogenous factor in our case.

Since there is not much literature studying the impact of student career aspirations on performance, we discuss different identification strategies of the literature examining other factors affecting student performance. Chetty, Friedman, and Rockoff (2014) construct a quasi-experimental design with teacher mobility at school grade level for identification, and by constructing a Value Added model, they found that each 1 standard deviation (SD) increase in teacher value-added was related to approximately 0.14 SD improvement in math standardized test scores. Fryer (2018) applied ITT and LATE to study the average effect of teacher specialization on student achievement by setting up a randomized field experiment in traditional public elementary schools in Houston. Cortes and Goodman (2014) investigates the effect of Double-Dose Algebra on student performance in a longitudinal dataset of CPS and applies difference-of-differences as an identification strategy.

Unlike many of these papers, we use cross-sectional data, the above identification strategies, such as difference-in-difference, do not work in our case. However, the propensity score method is a feasible identification strategy for cross-sectional data. In Imbens (2014), the author discussed that OLS estimators rely on the same unconfoundedness assumption as propensity score matching. Both mitigate the self-selection bias caused by the observable variables by controlling for variables related to the explained and treatment variables. However, OLS relies more on functional form restrictions and leads to biased estimates in case of functional form misspecification. Drake (1993) proved that the bias caused by the incorrect propensity score model is smaller than that caused by the misspecification of the function form of the linear regression model. Furthermore, the Imbens (2014) also suggests that the OLS estimator is sensitive to the distribution of covariates and is, therefore, not robust to the imbalance in covariates between treatment and control groups, in contrast, the PSM method is robust when outliers exist. We expect that the coefficient and significance obtained by the OLS and PSM are not different, thus verifying the robustness of our regression methods, even if non-obvious nonlinearities may exist in our linear models and imbalances in the distribution of covariates. However, the above two methods cannot solve the general endogeneity problems. Because of the possible endogeneity in the model, this thesis uses the proportion of teachers giving career guidance to students within a school as an instrumental variable for students' career aspirations. The innovation of this thesis is that there is currently less research on the impact of student aspirations on literacy and little literature to find instrumental variables for students' career aspirations. We examine the impact of career aspirations in the absence or presence of endogeneity separately.

3 DATA & DESCRIPTIVE STATISTICS

PISA evaluates the quality of education in participating countries through cross-country comparisons and provides recommendations for improving education and promoting the long-term development of education, and the PISA test covers reading, math, and science for 15-year-old students. Using a stratified random sampling method, PISA first selects a suitable sample of schools with 15-year-old students by the probability proportional to size (PPS) sampling. The selected schools provided a list of 15-year-old students, then 35 to 40 students were selected by random sampling for testing. Students in the schools selected by PISA will fill out a student questionnaire before the test, and the principal and teachers of the school will also fill out the school questionnaire and the teacher questionnaire, respectively, to obtain more information on the student's education process.

However, only 2015 and 2018 contain the instrumental variable we need: the proportion of teacher-to-student career guidance. To keep the specifications of our various methods the same across variables, countries and years, we only select the 2015 and 2018 datasets. Finally, from 29 million 15-year-old students in 72 participating countries and economies worldwide, 519,334 students were selected to participate in the scientific literacy test and filled out the relevant questionnaire in 2015. And in 2018, 612,004 students were selected from 32 million 15-year-olds from 79 countries and economies to participate in the PISA 2018 survey. Among the 17,908 schools in 2015, 12,611 are public schools, 1,392 are private independent and 1,538 are private government-dependent. Two thousand two hundred sixty-seven other schools did not indicate whether they were public or private. Of all the schools, 2,399 are from rural areas, 3,477 are in small towns, and 4,619 are in towns. There are 3,634 schools in medium-sized cities, 2,269 in large cities, and 1,510 schools that do not indicate their school location. In 2018, among the 21,903 schools, 16,808 are public schools, 1,966 are private independent and 1,558 are private government-dependent. The rest of the schools did not give a response. Of all the schools, 3,333 are from rural areas, 3,967 are in small towns, and 5,487 are in towns. There are 4,732 schools in medium-sized cities, 3,021 in large cities, and 1,363 schools that do not indicate their location.

We matched the student data and school data by the school IDs to form new data. We removed individuals missing school, student characteristics, treatments, and the instrumental variables from the sample, forming data with 60,139 observations from 2,457 schools in 13 countries in 2015, 24 students per school, and 71,905 students from 3,532 schools in the same countries in 2018, 20 students per school. Table 1 shows the countries with instrumental variables and the countries that appear in both years that are used in subsequent analyses. Among the control variables in our econometric model, student characteristics include family background, gender, science learning time, grade, ICT resources, psychological factors, and support from parents and teachers. The findings of Arulampalam (2008) show that student absences significantly negatively impact student performance. Teacher absence is also a factor in student achievement. Every 10 days a teacher is absent from school, a student's math performance decreases by 3.3% standard deviation (Miller, Murnane, and Willett (2007)). In addition to the student characteristics, the school characteristics included in the control variables are school size, average class size, shortage of educational material, and teacher- and student-related factors affecting school climate. Table 2 shows student and school characteristic variables, explanations, mean and standard deviation summarized from 60,139 observations in 2015 and 71,905 observations in 2018.

3.1 Outcome Variable

In the PISA 2015, 2018 data, students have ten plausible values (PV) each for their reading, math, and science test scores. To accurately measure students' ability in a particular field, too few questions will not work. Nevertheless, on a global scale, testing multiple subjects with many questions is difficult to operate, so PISA uses matrix sampling to resolve the contradiction between the two. Matrix sampling is a sampling design at the test content level: all test questions covering a subject area are divided into several approximately parallel, small test booklets. In this way, each student participating in the test only accepts one set of the test booklet, which reduces the number of questions each student needs to test and ensures an accurate estimate of the students' ability. The design ensures that the number of candidates in each question book is equal so that the accurate measure of the literacy of students in a particular country in a specific subject. The plausible values are estimated by item response theory (IRT) and the latent regression model and are often used in large-scale assessments, such as PISA, PIAAC, and TIMSS. To estimate student ability, in addition to considering students' responses to the questions, the probability distribution of students' ability values is estimated by combining relevant background variables, and plausible values are randomly selected to present the possible range of students' scores.

Von Davier, Gonzalez, and Mislevy (2009) demonstrates that the mean of plausible values cannot be used as the dependent variable, as it leads to biased estimates. The official PISA technical report suggests the final coefficient estimate should equal the average of the ten coefficients generated from 10 plausible values when using PV as the dependent variable. And the final sampling standard errors estimate obtained using Rubin's formula for multiple imputation (D. B. Rubin (1987)), equals to $\sqrt{\frac{1}{M}\sum_k s_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k\left(b_k-\bar{b}\right)^2}$, where each of the M plausible values as dependent variables and ${s_k}^2$ indicates the variance of coefficient b_k.

3.2 Variable of interest

$JExpectation$ is students' response to "What kind of job do you expect to have when you are about 30 years old?" are labeled from 0000 up to 10000 in the codebook. We classify students' expectations for future jobs in our data according to the International Standard Classification of Occupations (ISCO, Table 3). For further discussion, we will separately estimate the impact of career expectations on performance in terms of two different treatments, $HavePlan$ and $High$.

In Gutman and Schoon (2012), students' uncertain career aspirations play an important role in student performance. Therefore, we classify jobs 0-9 in $JExpectation$ as "having a clear plan" and 10 as "unclear plan" and construct a treatment variable $HavePlan$. In 2015, 3,704 students with $HavePlan = 0$ served as the control group, while 56,435 students with $HavePlan = 1$ served as the treatment group. The weighted average literacy of the treatment group is 489.8003, and that of the control group is 484.7855. In 2018, the weighted average science literacy of 65,764 treated observations was 496.1526, and that of 6,141 controls was 500.6107. Table 4 shows the balancing test for $HavePlan$ and $High$, respectively, given p-values with or without clustering at the school level. In both 2015 and 2018, parents' emotional support, gender, ICT resources, class size and student-related factors affecting school climate are significantly different between the control and treatment groups at the 5% statistical significance level. Overall, the distribution of covariates between the treatment and control groups is quite different.We also distinguish Skill levels 3 and 4 from $JExpectation$, i.e., $High$ as the treatment variable. In 2015, the control group consisted of 12,533 students with an average test score of 467.288. There are 47,606 students in the treatment group, with a mean science performance of 495.8578. In 2018, the weighted average science literacy of 55,064 treated observations was 501.8626, and that of 16,841 controls was 483.3059. Table 4 shows that there is little difference in teacher-related factors affecting school climate in 2015, and school size and shortage of educational material in 2018. All other variables show a strong imbalance between the two groups.

3.3 Weights

The PISA technical report states that weighting is necessary for any calculation using PISA data as the participation rates of different types of schools are different, and the response rates of students with different characteristics are different. We use the final weight variable (W_FSTUWT) in the regressions.

4 IDENTIFICATION STRATEGY

In our standard OLS Model, y_{ij} is the plausible value of the science literacy score of the individual i nested in country j, and $JExpectation_{ij}$ is a particular expected job of individual i considered as a treatment. $x_{1ij}, \ldots, x_{pij}$ are the predictors, including student and school characteristics:

$$y_{ij} = \gamma_1 JExpectation_{ij} + \sum_{k=1}^{p} \beta_k x_{kij} + \varepsilon_{ij}$$

In the following sections, we are interested in Average Treatment Effect on Treated (ATT), and will compare the ATT obtained by various methods.

4.1 Standard OLS regression & Fixed-Effects Model

We apply standard OLS and fixed-effects models in our first stage. Since the instrument used in the subsequent IV method varies at the school level thus school fixed effects are unavailable. To better compare the OLS, PSM, and IV methods for the same specification, we control for unobserved country-specific heterogeneity. The control variables of all OLS and FE regression models in this thesis are all variables in Table 2 except for instrumental variables $guide1$ and $guide2$. In addition, we add the interaction term between treatment and school-level variables to measure the mechanism of treatment effect. We use clustering robust standard error, where the error terms of individuals within a country are allowed to be correlated, while those of individuals across countries are not. Considering that career aspirations were formed before students fostered scientific literacy and that we include a wealth of variables in the model, we can reasonably assume that the model is free of endogeneity (zero conditional mean condition) to a certain extent. If the linearity assumption is satisfied, plus no perfect collinearity and very rare outliers, we get an unbiased estimate of the effect of treatment on scientific performance. Student weights are using in regressions.

4.2 Propensity Score Method

REASONS FOR APPLYING PROPENSITY SCORE MATCHING

✓ 1. Alleviate the reliance on the functional form specification to avoid biased estimates.

✓ 2. The PSM method assigns a weight to each individual according to the number of times they are matched, reducing the effect of outliers on estimates.

✓ * PSM method has advantages over the regression method in the presence of non-linearity and different distributions of covariates between the controls and treated.

With $HavePlan$ or $High$ (Jexpectation) as the treatment, the initial background of the treatment group and the control group are not the same. Therefore, the idea is to find samples in the control group with similar characteristics to the individuals in the treatment group and then compare them. When two people have similar family backgrounds, have the same gender, and have similar learning times and other control variables. Therefore, the difference in performance can only be caused by treatment to verify the causal effect of treatment and performance.To estimate the change in the outcome variable y of the individual A after treated $(D_i = 1)$, the ideal way is to find the individual A' of this person in parallel time and space, and let A' not receive the treatment $(D_i = 0)$, after the same period, get the result variable and make the difference to get the ATT. In reality, we are more likely to obtain the average treatment effect of all samples (ATE), as shown in the following formula:

$$ATE = E\left(Y_i(1) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 0\right)$$

$$= \underbrace{E\left(Y_i(1) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 1\right)}_{\text{ATT}} + \underbrace{E\left(Y_i(0) \mid D_i = 1\right) - E\left(Y_i(0) \mid D_i = 0\right)}_{\text{Selection bias}}$$

When the self-selection bias exists, whether being involved in treatment is an endogenous behavior. We are going to apply Propensity Score Matching and Weighting to alleviate the self-selection bias caused by observable covariates and estimate the ATT.

We use the same covariates as OLS and fixed-effects models in propensity score methods (propensity score matching and propensity score weighting). PSM relies on the same unconfoundedness (exogeneity) assumptions (Imbens (2014)) as ordinary least squares, which means that treatment is independent of any unobservable variables that affect the science performance, thus its effect on the outcome and treatment received can be consistently estimated.

THE ASSUMPTIONS IN PSM APPROACH

✓ 1. $\{(Y(1), Y(0)) \perp D\} \mid X$ (unconfoundedness).

- * Potential outcomes $Y(0)$ and $Y(1)$ are independent of treatment D conditioning on student and school characteristics . (Stronger assumption)

✓ 2. $0 < P(T = 1 \mid X) < 1$ (overlap).

- * Each individual has a positive probability of receiving a treatment.

4.2.1 Approach 1: Naive Propensity Score Matching

We first apply naive PSM to all observations, and all individuals, regardless of country origin, may be matched with individuals in their own country or other countries.

Propensity scores were measured using logistic regression with student weights:

$$logit\left(e_{ij}\right) = \alpha_0 + X_{ij}\beta$$

We apply propensity scores matching in all samples. The set of control units which matched to $tj \in I_T$ is defined as follows, where I_C and I_T represent the set of control units and treated units respectively, tj represents treated units in cluster j and cj' indicates control units in cluster j'. Since the matching is between all observations, we do not restrict $j = j'$, ie matching individuals are not forced to belong to the same school (Arpino and Cannas (2016)).

$$A_{tj} = \left\{ cj' \in I_C : \hat{e}_{cj'} = \min_{cj' \in I_C} \left| \hat{e}_{tj} - \hat{e}_{cj'} \right| < 0.25\hat{\sigma}_e \right\}$$

We apply 1:1 nearest neighbor matching with replacement, according to the principle of the closest propensity score value, each treatment group observation is matched to 1 control group sample. D. Rubin and Rosenbaum (1985) suggested a caliper with a standard deviation of 0.25 propensity score provides satisfactory results. All the matched observations form the following new dataset:

$$M = \{tj : A_{tj} \neq \emptyset\} \cup \left\{ \bigcup_{tj} A_{tj} \right\}$$

And the ATT estimator is as follows:

$$\widehat{ATT} = \frac{1}{\text{card}(M)} \left\{ \sum_{tj \in I_T \cap M} \left(Y_{tj} - \sum_{cj' \in I_C} Y_{cj'} w\left(tj, cj'\right) \right) \right\}$$

where $w\left(tj, cj'\right)$ is the estimated weight when the treated units tj are assigned to the control units cj'. We expect ATT by naive PSM to be close to the OLS estimator.

4.2.2 Approach 2: Within-Cluster Propensity Score Matching

In Approach 1, we did not account for cluster-level confounding factors, so that the ATT estimator could be biased. We use the within-cluster PSM method in Arpino and Cannas (2016) as approach 2 to account for the country-level confounders, and the matching is applied within each country. The matched dataset is different from that in Approach 1, as we set $j = j'$ in the within-cluster approach. Using the same propensity score model as naive PSM results in less balanced individual-level variables, even less than naive PSM (Arpino and Cannas (2016)). To obtain a more accurate match, our propensity scores are calculated within countries, and a caliper is obtained for each country based on the propensity score.

The ATT estimated is at the country/regions level, that is, within a particular country, the ATT of every student is the same. Finally, we average the ATT of all students regardless of the country to get within-cluster PSM method-based ATT. The advantage of within-cluster PSM compared to Naive PSM is that unobserved school confounders are considered. However, the procedure of computing ATT indicates that the standard errors in such clustered data are not available. Therefore, the within-cluster PSM method does not show whether ATT is significant. We expect ATT by within-country PSM to be close to the FE estimator.

4.2.3 Approach 3: Inverse Probability Weighting

The Inverse probability of treatment weighting (IPTW) method increases the weight of individuals who are less likely to receive treatment on the outcome variable by using the inverse probability of receiving treatment (i.e., propensity score) to reduce the effects of observed confounders. The covariates used for propensity score modeling here are the same as in the linear models and PSM for the corresponding treatment. We compute IPTW-ATT weights using estimated propensity scores:

$$w_{att} = \mathbb{1}(Treated = 1) + \frac{e(1 - \mathbb{1}(Treated = 1))}{1 - e}$$

Then weighted outcomes of treated and controls are calculated using w_{att}, and the difference is the IPW-based ATT. However, IPW does not perform well when the propensity score is too close to zero or one because this violates the positivity assumption. We apply the IPW methods of naive and within-cluster, respectively, and the weighted propensity score model is built for each county or region in the within-cluster IPW method.

4.3 Instrumental Variable

The OLS and PSM methods above rely on the strong assumption of unconfoundedness. We consider the following two sources of endogeneity such that the zero conditional mean (exogeneity) assumption is violated, resulting in biased estimates:

TWO SOURCES OF ENDOGENEITY:

✓ Endogenous source 1: Omitted variable.

- $\mathrm{E}\left(\varepsilon_{ij} \mid \boldsymbol{JExpectation}_{ij}, x_{kij}\right) \neq 0$. OLS is biased when the treatment variable and other controls are correlated to unobserved omitted variables in the error term.

✓ Endogenous Source 2: Simultaneity.

- $$\begin{cases} y_{ij} = \gamma_1 JExpectation_{ij} + \sum_{k=1}^{p} \alpha_k x_{kij} + u_{ij} \\ JExpectation_{ij} = \gamma_2 y_{ij} + \sum_{k=1}^{p} \beta_k x_{kij} + v_{ij} \end{cases}$$

  $$\Rightarrow JExpectation_{ij} = \frac{\beta_k x_{kij} + \gamma_2 \alpha_k x_{kij}}{1-\gamma_1\gamma_2} + \frac{1}{1-\gamma_1\gamma_2} v_{ij} + \frac{\gamma_2}{1-\gamma_1\gamma_2} u_{ij},$$

  which indicating $\mathrm{E}\left(JExpectation_{ij} u_{ij}\right) \neq 0$ and OLS estimator is biased.

To solve this problem, the idea is to find a filter (instrumental variable) to filter out the part related to the error term in $JExpectation_{ij}$, leaving only the orthogonal part. We find an instrumental variable, the proportion of "Student career guidance and counseling" within a school, for the treatment variable $HavePlan$ and $High$. The instrumental variable is inapplicable if we use school fixed effects.

The identification assumptions according to Angrist and Pischke (2009) are: i) Independence: $\left[\{\mathrm{Y}_i(d, z); \forall d, z\}, \mathrm{D}_i(1), \mathrm{D}_i(0)\right] \mid \mathrm{Z}_i$, where $Y_i(d; z)$ is the potential outcome of an individual with treatment status d and instrumental value z. This assumption implies that the instrumental variables are independent of both potential treatment status and potential outcomes, i.e. randomness. We believe that students are not easily aware of teachers' career guidance in advance and do not choose a level of guidance on their own initiative. ii) Exclusion Restriction: $\mathrm{Y}_i(d, 0) = \mathrm{Y}_i(d, 1) \equiv \mathrm{Y}_{di}$ for $d = 0, 1$. This means that the proportion of teachers' guidance to students at a school cannot affect a person's performance in ways other than career aspirations. iii) Relevance: $E\left[\mathrm{D}_i(1) - \mathrm{D}_i(0)\right] \neq 0$, indicating the heterogeneous causal effect of the instrument on treatment not 0. iv) Monotonicity: $\mathrm{D}_i(1) - \mathrm{D}_i(0) \geq 0, \forall i$, an increase in instrument level will not decrease treatment level. We will apply the two-stage least squares model (2SLS) and the fixed-effects two-stage least squares model (FE-2SLS).

We use ten plausible values as outcome variables for each identification strategy. Take the mean of the ten estimated coefficients or ATTs, and use Rubin's formula to calculate the final sampling standard errors as the final result.

5 EMPIRICAL RESULTS

5.1 Standard OLS Regression & Fixed-Effects Model

5.1.1 Standard OLS & Fixed-Effects Model - HavePlan

We calculated the variance inflation factor (VIF), the VIF of all variables is less than 2, whether it is $HavePlan$ or $High$ as the treatment variable or whether it is in data 2015 or 2018, indicating that there is no multicollinearity in our model. Table 5 shows that, in 2015, the coefficient of $HavePlan$ in standard OLS model in Column (2) is 8.388 (1.73%=8.388/484.7855). Controlling for country fixed-effects, Column (3) shows that $HavePlan = 1$ would result in students' science literacy being 5.882 (1.21%=5.882/484.7855) higher than without clear career expectations. Column (4), (5) show that the effect of $HavePlan$ is insignificant with adding the interaction term. Furthermore, from the significance of the interaction terms between treatment and school characteristics, we can conclude that the effect of having a clear plan does not depend on any one observable school characteristic. In 2018, We found that $HavePlan$ has no significant effect on student performance. In Column (10), the coefficient of the interaction between treatment and teacher-related factors affecting school climate is significantly negative, which indicates that the better the teacher-related school climate, the lower the positive impact of students' aspirations on grades. We conclude that at the mean of teacher-related factors affecting school climate (0.259), the impact of $HavePlan$ on performance is insignificantly 4.636 (=6.073+0.259*(-5.549)).

5.1.2 Standard OLS & Fixed-Effects Model - High

Consider the treatment variable is not $HavePlan$ but whether the desired occupation belongs to skill levels 3 and 4 (High). In Table 6, Column (2) and (3) of 2015 show that the effect of $High$ in the standard OLS model is 17.702 (3.79%=17.702/467.288), and adding fixed effects makes the effect of $High$ on student performance stronger, the coefficient of $High$ in the fixed-effects model is 23.728 (5.08%=23.728/467.288). In Column (5), the effect of interaction between treatment and school size is significantly positive, indicating a partial effect of 33.61 (=28.453+0.005*1032.609) at the mean of school size. A larger school size could improve the positive effect of $HavePlan$. In 2018, the fixed-effects model in Column (8) showed the impact of $High$ is significantly 15.884 (3.29%=15.884/483.3059). We can conclude from Column (10) at the mean of the Shortage of educational material (-0.222), the treatment effect is 25.292 (=26.179+3.994*(-0.222)) - the more scarce educational resources, the stronger the positive effect of $High$ on performance. See the further robustness check for OLS and FE in appendix.

Compare to Hien et al. (2019), where the occupational aspirations improved the average score of mathematics, physics, chemistry and biology by 0.337 (1.69%=0.337/19.95). This proportion is close to the OLS result in 2015 of our $HavePlan$ and slightly larger than that in FE. Contrary to the conclusions of Gutman and Schoon (2012), we found that a certain career expectation has a positive effect on student performance. The significant impact of $High$ aspirations is in line with Khattab (2015) that having only one of high aspirations or expectations positively affects school performance.

5.2 Propensity Score Matching & Inverse Probability Weighting

5.2.1 Propensity Score Matching & Inverse Probability Weighting - HavePlan

In the naive PSM method, all students with $HavePlan = 1$ are in the treatment group, and those with $HavePlan = 0$ are in the control group. The mean of the propensity score in 2015 is 0.933 with a standard deviation of 0.024, and the mean in 2018 is 0.877 with a standard deviation of 0.040. It can be seen that the students' propensity scores in 2015 are highly concentrated and tend to have a strong tendency to $HavePlan$, while propensity scores in 2018 were relatively discrete. Table 7 shows the outcome of controls is 474.993, and the ATT of $HavePlan$ by 1:1 matching with replacement in 2015 is significantly 8.201 (1.73%=8.201/474.993), quite close to the OLS estimator. And the ATT by 1:1 naive matching with replacement in 2018 is insignificant at 10%. Both results are very close to the standard OLS estimator in the corresponding year. It can be seen that 22 samples in 2015 and 45 samples in 2018 in the treatment group were outside the common support because their propensity scores are higher than the maximum or less than the minimum of the controls' propensity score. Note that the number of observations with weight is always slightly smaller than the number of observations on support. The gap between the propensity score of some treated and the matched control is beyond the caliper range. IPTW estimator shows similar results to PSM, 10.211 (2.16%=10.211/472.847) in 2015 and insignificant effect in 2018.

The balance test shows that, after matching, the absolute value of %bias for all variables is less than 15 except for school size. As shown in Figure 1(a) and 2(a), %bias after 1:1 matching with replacement for most of the covariates are less than %bias before matching in 2015, but the balance in 2018 does not seem to be improved. What's more, 11 of the 14 covariates in 2015 used for matching and 10 of 13 in 2018 reject the null hypothesis (H_0) of the t-test: there is no systematic bias in the values of the covariates between the two groups, statistically significant at the 1% level. After matching, the pseudo R square in the regression results is smaller, in 2015 from 0.015 to 0.008 and in 2018 from 0.011 to 0.008. The reduced pseudo R square means that the imbalance of the distribution of covariates is reduced, however, there is still much explanatory power for the changes of the explained variables in the Logit regression. To sum up, the naive PSM matching quality is not good. The results of 1:1 without replacement and 1:2 with replacement macthing and %bias across covariates are also shown in Table 7 and Figure 1(c), 1(e), 2(c) and 2(e), among them, no replacement matching has the worst balance after matching, and too many samples off support greatly reduce the external validity.

Frequency-weighted regressions are used for matched samples according to the different matching weights of individuals to obtain model-based ATT and cluster adjusting standard errors (standard errors in Table 7 assuming homoscedasticity). The coefficients we get from OLS with no covariates are in perfect agreement with those of ATT by naive PSM and the clustering adjusted standard errors, see Table 8. The coefficients of OLS and FE are always far from each other, which means that there are still country-level confounders in the samples after naive PSM.

Then we apply within-country PSM. The propensity score model built for the within-country PSM takes into account the heterogeneity of countries, and the propensity score of $HavePlan$ is no longer as high as that of the naive PSM. The ATT of $HavePlan$ is not significant in 2015 nor in 2018, see Table 9, and the results are similar to the previous FE model for all samples. All the %bias are less than 4 in Figure 1(b), 2(b). Seven of 14 in 2015 and only 2 of 14 in 2018 covariates reject the hypothesis of no systematic bias at 1% level. The pseudo R^2 decreased from 0.015 to 0.001 in 2015, and from 0.011 to 0.000 in 2018. All of the above show that within-country PSM performs much better than naive PSM. Figure 1(d) of 1:1 no replacement matching still perform the worst with large number of samples are off support, and as shown in Column (2) in Table 9. The within-cluster IPW estimator for the ATT is 4.2896 (0.09%=4.2896/473.0101) in 2015. Note that Chinese Taipei in 2015 is omitted, because all students in both countries have $HavePlan$ of 1 in the corresponding year.

Using the within-cluster matched sample for regression analysis, the results in Table 10, OLS and the fixed-effects models without covariates show the same coefficients and provide model-based standard errors. The model-based ATT is comparable to the ATT by within-country PSM method and is very close to the coefficient of the fixed-effects model for all samples. All OLS and FE results in the table are similar because we control for country-level confounders.

5.2.2 Propensity Score Matching & Inverse Probability Weighting - High

Students with $High = 1$ are in the treatment group, and those with $High = 0$ are in the control group, regardless their schools. The mean of the propensity score of $High$ in 2015 is 0.776 with standard deviation of 0.109, and the mean in 2018 is 0.733 with standard deviation of 0.105. For 1:1 naive matching with replacement, the ATT of $High$ after matching is 15.4119 (3.26%=15.4119/472.06192) in 2015 and 18.9352 (3.97%=18.9352/477.38486) in 2018, significant at the 1% level in Table 11, and there is no obvious difference with the results of matching and IPW method.

After 1:1 naive matching with replacement, the absolute value of %bias for all variables is less than 8 in 2015 and 15 in 2018, as shown in Figure 3(a) and 4(a). However, 11 of 14 covariates in 2015 and 11 of 13 variables in 2018 reject the hypothesis of no systematic bias between the two groups. The pseudo R square in the regression results dropped from 0.054 to 0.006 in 2015 and 0.040 to 0.013 in 2018 after matching, which means that all the covariates between the two groups are still different.

The ATT by 1:1 within-country matching with replacement of $High$ is 18.705 (3.96%=18.705/472.55237) in 2015 and 17.649 (3.67%=17.649/481.30222) in 2018 in Table 13, not very different from that by fixed-effects models before matching. The IPW based ATT is 18.21861 (3.85%=18.21861/473.13196) in 2015 and 16.84323 (3.49%=16.84323/481.95532) in 2018. All the %bias are less than 4, and the null hypothesis is only rejected for 5 of 14 covariates in 2015 and 2 of 13 covariates in 2018 at 1% level. The pseudo R square in the regression results dropped from 0.054 to 0.001 in 2015 and 0.040 to 0.000 in 2018. Within-country PSM displays its excellent matching quality. Frequency-weight regressions are applied for matched samples. The results are shown in Table 12 and 14, and We reach the same conclusion as previous regression analysis on matched samples.

We found that among the naive methods for each treatments, the largest gap is the coefficient of $High$ between OLS for all samples and 1:1 with replacement Naive PSM in 2018, which is 71% (=(18.935-11.049)/11.049). For within-country methods for $HavePlan$ and $High$, the largest gap is only the coefficient of $High$ between FE for all samples and 1:1 with replacement within-country PSM in 2015, which is -21% (=(18.705-23.728)/23.728). We prove the robustness of fixed-effects model by within-country PSM method. See the further robustness check for naive PSM and within-cluster PSM in appendix.

5.3 Instrumental Variable Method: Two-Stage Least Squares & Fixed-effects Two-Stage Least Squares

5.3.1 Exclusion Restriction

Exclusion restriction denotes that the instrumental variable does not directly affect the dependent variable. Any effect of the instrument on the outcome must be through the treatment variables. The instrumental variable detects movements in the endogenous variables that are uncorrelated with the error terms and then use them to estimate the causal effects. The proportion of giving career guidance to students is a school-level instrumental variable and is very likely to affect student performance only through the treatment variables $HavePlan$ and $High$ and is thus exogenous.

Innate or fixed characteristics, such as gender, ESCS, school size, classroom size, ICT resource, grade, teacher support in science classes, and parent emotional support, are almost impossible to be affected by the proportion of career guidance to students affected. However, we think of two potential violations of the exclusivity constraint. First, career guidance may not only affect performance through students' $HavePlan$ or $High$ but may also be through achieving motivation and learning time. However, considering the career expectations of students in our data are more inclined to be an aspiration, not a well-thought-out plan. In addition, students are generally 15-year-old junior high school students who lack career understanding, life experience, and professional experience. Therefore, students may not be significantly affected by career guidance to study habits, and thus affect their performance. Second, career guidance is an intervention from teachers, and guidance and assistance from teachers may improve student-related school climates, such as reducing absenteeism, which leads to the potential to improve student performance. However, career guidance may not be enough to improve the atmosphere of middle school students.

The exclusion restriction is not testable because the true error term is not observable. We may not be able to completely rule out that the proportion of career guidance affects performance in other ways, but we believe that the role of other ways may not be significant.

5.3.2 First-Stage Diagnostics: Relevance of the instruments

In Table 15, the output from the first stage diagnostics for under-identification (Anderson canon. corr. LM statistic) indicates that the proportion of ”Student career guidance and counseling included in my professional development activities during the last 12 months”(guide2) is a relevant instrument for

$HavePlan$ in IV-2SLS and 2SLS-FE models in 2015. And the proportion of "Student career guidance and counseling included in my teacher education or training program or other professional qualification" (guide1) shows relevance to the treatment in IV-2SLS and 2SLS-FE models in 2018. Neither the two guidances show their relevance to $High$. We will use $guide2$ as the instrumental variable for the two treatments in 2015 and $guide1$ as the instrumental variable for the two treatments in 2018.

The weak identification test is used to test the strength of instrumental variables. Kleibergen-Paap rk Wald F statistic is robust to heterogeneity, serial correlation, and clustering. The critical value under 10% bias is 16.38. And 8.96, 6.66 and 5.53 are for 15%, 20%, 25% maximal IV size, respectively. For 2SLS models in Columns (1) and (3) in Table 16, F statistics are much larger than the critical value, indicating that $guide2$ in 2015 and $guide1$ in 2018 are strong instrumental variables for 2SLS models. For the FE-2SLS models, only $guide1$ for $HavePlan$ in 2018 is not too weak (close to the 25% maximal IV size) .

5.3.3 Second Stage

Without considering country fixed-effects, the effect of $HavePlan$ is significantly -406.616 (-83.88%=-406.616/484.7855) in 2015 and insignificant in 2018, see Table 16. The 2SLS-FE model in 2018 shows a significantly negative effect of -441.638 (-91.10%=-441.638/484.7855) from $HavePlan$. Our conclusions contrast to previous OLS and PSM methods but in line with Gutman and Schoon (2012)'s finding, teenagers with ambiguous career aspiration have higher academic performance. The estimated coefficients of 2SLS and FE-2SLS are very large. They are dozens of times the magnitude of OLS estimators and ATTs by PSM.

Our instrumental variable for $High$ performs worse than for $HavePlan$, as can be seen from Table 17, the proportion of career guidance is a relevant instrument for $High$ in 2SLS models in 2018, see Column (3), however, in 2SLS-FE models the guidance is a very weak instrumental variable. When there is little correlation between the instrumental and the explanatory variables, the part of the explanatory variables determined by the instrumental variables is tiny, and there may be a large bias in the results.

In our case, the local average treatment effects are much larger than the coefficients of OLS and PSM in this way. The coefficients obtained by 2SLS and 2SLS-FE are tens or hundreds of times larger than the coefficients of OLS, fixed-effects models and PSM methods in the previous sections. In the case of binary treatment and instrument, LATE can be considered the Intention-to-treat (ITT) estimator divided by the proportion of compliers in the sample. Therefore, when the instrumental variables are not strong enough, the small scale of the compliers will magnify the estimates of the coefficients. In our case, the reason for this upward bias is similar, that the proportion of teachers in schools to student career guidance only affects students who are sensitive to the guidance. For students who will anyway have a plan, (always-takers), or no plan (never-takers), no matter how high the proportion of guidance, it cannot affect their choice. In this way, instrumental variables only affect compliers, and the coefficient is enlarged because the proportion of the compliers is small.

See the further robustness check for 2SLS and FE-2SLS in appendix.

6 Heterogeneity Analysis

We select a western country Germany (DEU), and an eastern region Hong Kong (HKG) to interact with treatment, respectively. In Table 20, we find that in the fixed-effects model in Column (6) in 2015, the significant positive effect of students' $High$ on performance is reduced in Hong Kong. Fixed-effects models in 2015 showed that the effects of both treatments on student achievement were not associated with Germany. In 2018, the fixed effect model showed that Hong Kong could improve the value-added of students' $HavePlan$ to performance. And Germany increases the influence of $High$. In 2SLS models, Hong Kong can increase the impact of $HavePlan$ and $High$ in 2015 and $HavePlan$ in 2018. The interaction terms of Germany with $HavePlan$ and $High$ are both significantly favorable in 2SLS models in 2015. Among all 2SLS-FE regressions, only $guide1$ for $HavePlan$ in 2018 in Column (4) is a relatively non-weak instrumental variable, Hong Kong and Germany show an improved effect on $HavePlan$.

7 Conclusion

In this thesis, we first assumed the absence of endogenous sources. We applied OLS and FE models to study the effects of two treatments on student performance and whether students have clear career expectations and high-skilled career expectations. However, the ordinary least squares model is sensitive to the specification of functional forms and the imbalance of the distribution of covariates, we thus apply naive propensity score matching and within-country propensity score matching, which do not rely on linearity assumptions, balance the covariates in treated and control groups, to verify the accuracy of standard OLS and FE models. Naive PSM and Within-Country PSM are compared to the results of OLS and FE, respectively.

In the PSM section, we briefly compare several matching strategies, 1:1 replacement, 1:1 no replacement, and 1:2 replacement, where 1:1 replacement may have slightly higher variance but less bias than the other two methods. The 1:1 matching method without replacement leads to lower matching quality and a smaller sample size. In matching without replacement, many observations do not meet the common support assumption, and a large number of samples are excluded, making the sample less representative and affecting the external validity of the results. Using within-cluster PSM can capture cluster-level confounding, but the disadvantage is that we cannot get the standard error of ATT obtained by within-cluster PSM without the help of other methods. We use frequency-weighted regressions with no covariates on the matched samples in the paper to obtain the same coefficients as PSM-based ATT, as well as cluster-adjusted standard errors.

We find that all the results of within-country PSM are in line with the coefficients in FE models, while some of the OLS results show a gap between the ATT by naive PSM. We thus conclude fixed-effect models to be robust, assuming no endogeneity in the model. According to the regression results of the fixed effect model, the impact of clear career expectations on performance in 2015 was 5.882 (1.21%=5.882/484.7855), and the impact in 2018 was not significant. The impact of high-skilled career expectations on performance was 23.728 (5.08%=23.728/467.28) in 2015 and 15.884 (3.29%=15.884/483.3059) in 2018, which means those clear career aspirations or high-skill career aspirations are favorable for improving students' personal qualities. By comparing different years, we find that the coefficient of $High$ becomes smaller from 2015 to 2018, while $HavePlan$ even becomes insignificant. The impact of student career aspirations decreases over the two years.

We then assume that the models have two sources of endogeneity from omitted variables and simultaneity. We used the school-level instrumental variable, the proportion of teacher-to-student career guidance. In both 2015 and 2018, career guidance was a strong instrumental variable for $HavePlan$ in 2SLS models, while in 2SLS-FE models, it was only a non-weak instrumental variable with close to the critical value of 25% maximal IV size In the instrumental variable method, we find that the coefficients of 2SLS and 2SLS-FE are generally dozens or hundreds of times larger than those of OLS and FE. The 2SLS-FE model in 2018 shows that clear career expectations have a significantly negative effect of -441.638 (-91.10%=-441.638/484.785). This coefficients is much larger than those of the original FE and show a negative effect, contrary to our previous overall conclusion. The main reason is that the instrumental variable is not strong enough. In fact, only the compliers are affected by the instrumental variables. Considering that LATE is the result of dividing the average effect of experimental assignment on outcomes by the proportion of compliers, and when the ratio of compliers is smaller, the coefficients will be scaled inversely proportionally. Therefore, we need stronger instrumental variables to identify the causal effects of career aspirations.

References

Rubin, Donald and Paul Rosenbaum (Feb. 1985). "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score". In: *The American Statistician* 39. DOI: 10.1080/00031305.1985.10479383.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, p. 258.

Drake, Christiana (1993). "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect". In: *Biometrics* 49 (4), pp. 1231–1236.

Mau, Wei-Cheng (Mar. 2003). "Factors That Influence Persistence in Science and Engineering Career Aspirations". In: *The Career Development Quarterly* 51.

Rojewski, Jay W (2005). "Occupational aspirations: Constructs, meanings, and application". In: *Career development and counseling: Putting theory and research to work*, pp. 131–154.

Miller, Raegen, Richard Murnane, and John Willett (Jan. 2007). "Do Teacher Absences Impact Student Achievement? Longitudinal Evidence From One Urban School District". In: *National Bureau of Economic Research, Inc, NBER Working Papers* 30. DOI: 10.3102/0162373708318019.

Arulampalam Wiji, Naylor (2008). "Am I missing something? The effects of absence from class on student performance". In: *IZA Discussion Papers*.

Angrist, Joshua and Jörn-Steffen Pischke (Jan. 2009). "Mostly Harmless Econometrics: An Empiricist's Companion". In: ISBN: 9780691120348 (hardcover : alk. paper).

Park, Hyeran, Samia Khan, and Stephen Petrina (2009). "ICT in Science Education: A quasi-experimental study of achievement, attitudes toward science, and career aspirations of Korean middle school students". In: *International Journal of Science Education* 31, pp. 1012–993.

Von Davier, Matthias, E Gonzalez, and RJ Mislevy (Jan. 2009). "What are plausible values and why are they useful". In: *IERI monograph series* 2, pp. 9–36.

Gutman, Leslie and Ingrid Schoon (June 2012). "Correlates and consequences of uncertainty in career aspirations: Gender differences among adolescents in England". In: *Journal of Vocational Behavior* 80, pp. 608–618. DOI: 10.1016/j.jvb.2012.02.002.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (Sept. 2014). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates". In: *American Economic Review* 104(9), pp. 2593–2632.

Cortes, Kalena E. and Joshua S. Goodman (May 2014). "Ability-Tracking, Instructional Time, and Better Pedagogy: The Effect of Double-Dose Algebra on Student Achievement". In: *American Economic Review* 104(5), pp. 400–405.

Imbens, Guido (Mar. 2014). *Matching Methods in Practice: Three Examples*. Working Paper 19959. National Bureau of Economic Research. DOI: 10.3386/w19959. URL: http://www.nber.org/papers/w19959.

Khattab, Nabil (2015). "Students' aspirations, expectations and school achievement: what really matters?" In: *British Educational Research Journal* 41(5), pp. 731–748.

Arpino, Bruno and Massimo Cannas (2016). "Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score". In: *Statistics in Medicine* 35(12), pp. 2074–2091.

Fryer Roland G., Jr. (Mar. 2018). "The "Pupil" Factory: Specialization and the Production of Human Capital in Schools". In: *American Economic Review* 108(3), pp. 616–56. DOI: 10.1257/aer.20161495. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20161495.

Hien, Hien et al. (Sept. 2019). "Reading Habits, Socioeconomic Conditions, Occupational Aspiration and Academic Achievement in Vietnamese Junior High School Students". In: *Sustainability* 11, p. 5113.

8 Appendix

Country or Region Code	Country or Region Name	# of observations in 2015	# of observations in 2018	Selected
ALB	Albania	-	2,116	
QAZ	Baku (Azerbaijan)	-	1,091	
AUS	Australia	8,532	-	
BRA	Brazil	7,693	4,384	X
CHL	Chile	4,818	3,301	X
TAP	Chinese Taipei	5,459	5,890	X
COL	Colombia	7,593	-	
CZE	Czech Republic	5,471	-	
DOM	Dominican Republic	2,382	974	X
DEU	Germany	2,576	1,644	X
HKG	Hong Kong	3,124	3,522	X
ITA	Italy	6,194	-	
KOR	Korea	4,398	6,119	X
MAC	Macao	3,703	3,542	X
MYS	Malaysia	-	5,501	
MAR	Morocco	-	718	
PAN	Panama	-	1,085	
PER	Peru	5,317	2,106	X
PRT	Portugal	2,960	2,963	X
ESP	Spain	4,631	21,345	X
ARE	United Arab Emirates	8,704	12,556	X
GBR	United Kingdom	-	1,353	
USA	United States	4,374	3,559	X
QCH	B-S-J-G (China)	8,389	-	
	Total	96,318	83,769	
	Total # of observations selected	60,139	71,905	

Table 1: Number of Observations in Different Countries

		2015		2018	
Student Variables	Explanation	Mean	Sd	Mean	Sd
Female	1:Female; 0:Male	0.500	0.500	0.507	0.500
ESCS	Index of economic, social and cultural status	-0.667	1.262	-0.062	1.082
SMINS	Science Learning time (minutes per week)	227.555	154.758	226.625	148.073
EMOSUPS	Parents emotional support	-0.063	0.997	0.067	0.983
ANXTEST	Personality: Test Anxiety	0.206	0.962	-	-
MOTIVAT	Achieving motivation	0.176	0.969	-	-
COMPETE	Students' competitiveness achievement motive	-	-	0.181	0.996
TEACHSUP	Teacher support in a science classes	0.173	0.934	0.192	0.946
ICTRES	ICT Resources	-0.726	1.270	-0.130	1.137
GRADE	Grade compared to modal grade in country	-0.234	0.760	-0.060	0.672
School Variables	Explanation	Mean	Sd	Mean	Sd
SCHSIZE	School Size	1032.609	1097.047	1245.939	946.0996
CLSIZE	Class Size	31.73	10.237	28.605	7.359
EDUSHORT	Shortage of educational material	0.221	1.165	-0.222	1.029
STUBEHA	Student-related factors affecting school climate	-0.021	1.200	0.282	1.001
TEACHBEHA	Teacher-related factors affecting school climate	0.029	1.151	0.259	0.931
guide1	The proportion of "Student career guidance and counseling included in the teacher education/ training program/other professional qualification"	0.337	0.216	0.291	0.196
guide2	The proportion of "Student career guidance and counseling included in my professional development activities during the last 12 months"	0.252	0.171	0.235	0.151

Table 2: Student & School characteristics - Descriptive Statistics

Broad skill level	Code	Occupation	Label
Skill levels 3 and 4 (High)	1	Manager	[1000 2000)
	2	Professionals	[2000 3000)
	3	Technicians and Associate Profession	[3000 4000)
Skill level 2 (Medium)	4	Clerical Support Workers	[4000 5000)
	5	Services and Sales Worker	[5000 6000)
	6	Skilled Agricultural, Forestry and Fishery Work	[6000 7000)
	7	Craft and Related Trades Worker	[7000 8000)
	8	Plant and Machine Operators and Assemblers	[8000 9000)
Skill level 1 (low)	9	Elementary Occupations	[9000 9700)
Armed Forces	0	Armed Forces Occupation	[0000 1000)
Unemployed or Vague	10	Not elsewhere classified	[9700 9997)
-	11	invalid	[9997 10000)

Table 3: International Standard Classification of Occupations

	2015				2018			
	HavePlan	Don't HavePlan	P-value	P-value with clustering	HavePlan	Don't HavePlan	P-value	P-value with clustering
ESCS	-0.156	-0.086	0.061	0.083	-0.070	-0.007	0.077	0.115
SMINS	228.140	228.931	0.852	0.865	225.854	231.728	0.189	0.206
ANXTEST	0.258	0.144	0.001	0.002	-	-	-	-
MOTIVAT	0.427	0.218	0.000	0.000	-	-	-	-
COMPETE	-	-	-	-	0.187	0.141	0.154	0.137
TEACHSUP	0.258	0.207	0.115	0.142	0.203	0.120	0.007	0.006
EMOSUPS	0.082	-0.026	0.002	0.003	0.075	0.013	0.055	0.075
female	0.520	0.455	0.000	0.000	0.523	0.399	0.000	0.000
GRADE	-0.057	-0.030	0.173	0.175	-0.063	-0.040	0.199	0.257
ICTRES	-0.211	-0.128	0.028	0.042	-0.146	-0.024	0.002	0.006
CLSIZE	29.051	28.043	0.000	0.000	28.690	28.037	0.004	0.009
STUBEHA	0.044	0.155	0.002	0.007	0.274	0.331	0.026	0.058
TEACHBEHA	0.080	0.112	0.394	0.435	0.255	0.286	0.225	0.371
SCHSIZE	1165.947	1199.350	0.267	0.296	1231.746	1339.837	0.001	0.002
EDUSHORT	-0.134	-0.157	0.412	0.466	-0.210	-0.304	0.004	0.021
guide1	0.340	0.298	0.000	0.000	0.295	0.264	0.000	0.000
guide2	0.254	0.221	0.000	0.000	0.237	0.223	0.000	0.001
Sample Size	56435	3704	.	.	65764	6141	.	.
	2015				2018			
	High	Not High	P-value	P-value with clustering	High	Not High	P-value	P-value with clustering
ESCS	-0.089	-0.364	0.000	0.000	-0.008	-0.203	0.000	0.000
SMINS	232.900	211.941	0.000	0.000	230.129	217.441	0.000	0.000
ANXTEST	0.272	0.176	0.000	0.000	-	-	-	-
MOTIVAT	0.491	0.142	0.000	0.000	-	-	-	-
COMPETE	-	-	-	-	0.198	0.137	0.008	0.009
TEACHSUP	0.272	0.195	0.000	0.001	0.216	0.130	0.000	0.000
EMOSUPS	0.112	-0.054	0.000	0.000	0.097	-0.012	0.000	0.000
female	0.551	0.392	0.000	0.000	0.557	0.375	0.000	0.000
GRADE	-0.038	-0.112	0.000	0.000	-0.043	-0.102	0.000	0.000
ICTRES	-0.169	-0.327	0.000	0.000	-0.106	-0.193	0.001	0.004
CLSIZE	29.196	28.232	0.000	0.000	28.915	27.792	0.000	0.000
STUBEHA	0.024	0.147	0.000	0.000	0.253	0.357	0.000	0.000
TEACHBEHA	0.076	0.106	0.155	0.266	0.242	0.301	0.001	0.030
SCHSIZE	1177.882	1135.246	0.011	0.039	1252.468	1228.828	0.288	0.375
EDUSHORT	-0.152	-0.078	0.000	0.004	-0.231	-0.198	0.135	0.252
guide1	0.336	0.339	0.446	0.567	0.288	0.298	0.007	0.038
guide2	0.252	0.249	0.280	0.432	0.234	0.238	0.274	0.397
Sample Size	47606	12533	.	.	55064	16841	.	.

Table 4: Balancing Test for $HavePlan$ & $High$

Table 5: OLS Regression Model: Treatment - HavePlan

	2015					2018				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	OLS	OLS	FE	OLS	FE	OLS	OLS	FE	OLS	FE
	science	science	science	science	science	science	science	science	science	science
HavePlan	5.015	8.388**	5.882*	1.782	8.024	-4.458	-0.651	0.510	9.102	6.073
	(3.731)	(3.267)	(3.153)	(12.191)	(11.824)	(3.784)	(3.176)	(3.184)	(12.175)	(12.263)
c.HavePlan#c.CLSIZE				-0.021	-0.250				-0.412	-0.262
				(0.375)	(0.362)				(0.398)	(0.400)
c.HavePlan#c.STUBEHA				-1.796	0.198				3.657	2.902
				(3.072)	(2.952)				(3.267)	(3.209)
c.HavePlan#c.TEACHBEHA				1.789	2.190				-6.674**	-5.549*
				(3.259)	(3.148)				(3.232)	(3.221)
c.HavePlan#c.SCHSIZE				0.006	0.004				0.002	0.002
				(0.004)	(0.004)				(0.003)	(0.003)
c.HavePlan#c.EDUSHORT				2.839	3.435				2.536	2.293
				(3.425)	(3.322)				(2.631)	(2.642)
_cons	484.785***	476.087***	467.760***	482.243***	465.657***	500.611***	496.176***	472.713***	487.440***	467.686***
	(4.201)	(6.635)	(6.686)	(12.357)	(12.429)	(4.358)	(6.891)	(7.113)	(12.827)	(13.107)
With Covariates	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
With Interaction Terms	No	No	No	Yes	Yes	No	No	No	Yes	Yes
N	60,139	60,139	60,139	60,139	60,139	71,905	71,905	71,905	71,905	71,905
R^2	0.000	0.272	0.356	0.272	0.356	0.000	0.244	0.289	0.244	0.289

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

	2015					2018				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	OLS	OLS	FE	OLS	FE	OLS	OLS	FE	OLS	FE
	science	science	science	science	science	science	science	science	science	science
High	28.570***	17.702***	23.728***	29.807***	28.453***	18.557***	11.049***	15.884***	25.929***	26.179***
	(2.254)	(1.979)	(1.893)	(7.952)	(7.850)	(2.640)	(2.179)	(2.196)	(7.909)	(8.002)
c.High#c.CLSIZE				-0.713***	-0.403				-0.635**	-0.440
				(0.262)	(0.261)				(0.270)	(0.271)
c.High#c.STUBEHA				3.555*	2.413				2.983	1.301
				(1.974)	(1.940)				(2.114)	(2.087)
c.High#c.TEACHBEHA				1.123	0.017				-4.276*	-3.480
				(2.034)	(2.016)				(2.326)	(2.361)
c.High#c.SCHSIZE				0.006**	0.005**				0.003	0.003
				(0.003)	(0.003)				(0.002)	(0.002)
c.High#c.EDUSHORT				-2.558	-2.184				3.467*	3.994**
				(1.964)	(1.870)				(1.880)	(1.913)
_cons	467.288***	473.798***	458.184***	464.154***	454.383***	483.306***	490.290***	464.294***	479.100***	456.555***
	(2.485)	(6.101)	(6.478)	(8.010)	(8.752)	(3.100)	(6.639)	(7.019)	(8.947)	(9.494)
With Covariates	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
With Interaction Terms	No	No	No	Yes	Yes	No	No	No	Yes	Yes
N	60,139	60,139	60,139	60,139	60,139	71,905	71,905	71,905	71,905	71,905
R^2	0.014	0.277	0.365	0.278	0.365	0.007	0.246	0.294	0.247	0.294

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: OLS Regression Model: Treatment - High

TABLE: Naive Propensity Score Matching & Inverse Probability Weighting Method

	Naive Propensity Score Matching					
	2015			2018		
	(1) One to One With Replacement	(2) One to One No Replacement	(3) One to Two With Replacement	(4) One to One With Replacement	(5) One to One No Replacement	(6) One to Two With Replacement
Outcome of Treated[1]	483.19412	489.85507	483.19412	490.63751	492.69272	490.63751
Outcome of Controls[2]	474.99329	478.65750	474.78491	490.54071	492.02542	490.53506
ATE by PSM	8.02116	8.43960	8.30612	-0.06847	-0.37823	-0.02150
ATT by PSM	**8.20085*****	**11.19759*****	**8.40922*****	**0.09681**	**0.66730**	**0.10246**
Standard Errors of ATT	2.36608	2.41535	2.15581	1.83094	1.98133	1.69713
# of obs On Common Support	60,117	7,408	60,117	71,860	12,282	71,860
Treated	56,413	3,704	56,413	65,719	6,141	65,719
Controls	3,704	3,704	3,704	6,141	6,141	6,141
# of obs Off Common Support	22	52,731	22	45	59,623	45
Treated	22	52,731	22	45	59,623	45
Controls	0	0	0	0	0	0
# of obs with weight	60,060	7,408	60,117	71,697	12,282	71,856
Treated	56,413	3,704	56,413	65,719	6,141	65,719
Controls	3,647	3,704	3,704	5,978	6,141	6,137
Sum of Weights	112,826	7,408	112,826	131,438	12,282	131,438
	Naive Propensity Score Weighting					
	2015			2018		
Weighted Outcome among Treated	483.16974			490.63519		
Weighted Outcome among Controls	472.84735			490.30420		
ATT by IPTW	**10.32238*****			**0.33097**		
Standard Errors of ATT by Reg	2.25433			1.59177		
N	60,139			71,905		

[1] Mean of scientific performance for supported and treated individuals.

[2] Mean of scientific performance of control group individuals matched with supported and treated individuals.

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Table 7: Naive Propensity Score Matching & Inverse Probability Weighting Method - HavePlan

(Weighted) OLS Regressions: All Samples vs. Naive Matched Sample Without considering unobserved country-level confounders - HavePlan														
2015														
	All Samples		One to One With Replacement Matched Sample				One to One No Replacement Matched Sample				One to Two With Replacement Matched Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE
HavePlan	8.388**	5.882*	8.201***	6.989***	5.898***	3.832**	11.198***	6.818***	9.023***	4.977**	8.409***	6.862***	6.264***	3.505**
	(3.267)	(3.153)	(2.664)	(2.200)	(2.267)	(1.788)	(2.656)	(2.340)	(2.417)	(1.963)	(2.455)	(2.058)	(2.122)	(1.700)
_cons	476.1***	467.8***	475.0***	475.6***	475.4***	462.8***	478.7***	480.8***	470.9***	470.0***	474.8***	475.6***	475.5***	462.5***
	(6.635)	(6.686)	(2.987)	(2.403)	(7.242)	(5.575)	(2.469)	(2.022)	(7.464)	(5.963)	(2.796)	(2.270)	(6.941)	(5.420)
N	60,139	60,139	112,826	112,826	112,826	112,826	7,408	7,408	7,408	7,408	225,652	225,652	225,652	225,652
R^2	0.272	0.356	0.002	0.261	0.218	0.465	0.003	0.226	0.192	0.440	0.002	0.261	0.217	0.462
2018														
	All Samples		One to One With Replacement Matched Sample				One to One No Replacement Matched Sample				One to Two With Replacement Matched Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE
HavePlan	-0.651	0.510	0.097	3.224*	-4.815***	0.321	0.667	3.712*	-4.507**	0.791	0.102	3.238*	-4.933***	0.005
	(3.176)	(3.184)	(2.017)	(1.824)	(1.848)	(1.605)	(2.193)	(2.041)	(2.059)	(1.837)	(1.886)	(1.702)	(1.740)	(1.513)
_cons	496.2***	472.7***	490.5***	489.0***	504.0***	481.6***	492.0***	490.5***	500.2***	484.4***	490.5***	489.0***	505.0***	482.4***
	(6.891)	(7.113)	(2.247)	(1.999)	(5.189)	(4.493)	(2.039)	(1.831)	(5.526)	(5.094)	(2.134)	(1.892)	(4.976)	(4.242)
N	71,905	71,905	131,438	131,438	131,438	131,438	12,282	12,282	12,282	12,282	262,876	262,876	262,876	262,876
R^2	0.244	0.289	0.000	0.136	0.138	0.301	0.000	0.121	0.122	0.284	0.000	0.138	0.135	0.299
Covariates	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Weights	Sampling Weight		Frequency Weight for OLS, FE on matched samples, and Sampling Weight for Propensity Score Model before Matching											

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Table 8: Regressions: All Samples vs. Naive Matched Sample - HavePlan

TABLE: Within-Country Propensity Score Matching & Inverse Probability Weighting Method						
	Within-Country Propensity Score Matching					
	2015			2018		
	(1) One to One With Replacement	(2) One to One No Replacement	(3) One to Two With Replacement	(4) One to One With Replacement	(5) One to One No Replacement	(6) One to Two With Replacement
Outcome of Treated[1]	477.18112	470.74280	477.18112	491.14001	487.88980	491.14001
Outcome of Controls[2]	473.21713	469.21854	473.90906	489.08502	487.90402	489.98016
ATE by PSM	3.98687	2.60054	3.29515	2.00643	0.80800	1.21838
ATT by PSM	**3.96398**	**1.52424**	**3.27205**	**2.05502**	**-0.01423**	**1.15987**
Standard Errors of ATT	-	-	-	-	-	-
# of obs On Common Support	54,323	7,398	54,323	71,596	12,257	71,596
Treated	50,619	3,694	50,619	65,455	6,116	65,455
Controls	3,704	3,704	3,704	6,141	6,141	6,141
# of obs Off Common Support	357	47,282	357	309	59,648	309
Treated	357	47,282	357	309	59,648	309
Controls	0	0	0	0	0	0
# of obs with weight	54,224	7,388	54,308	71,313	12,232	71,557
Treated	50,619	3,694	50,619	65,455	6,116	65,455
Controls	3,605	3,694	3,689	5,858	6,116	6,102
Sum of Weights	101,238	7,388	101,238	130,910	12,232	130,910
	Within-Country Propensity Score Weighting					
	2015			2018		
Weighted Outcome among Treated	477.17062			491.07513		
Weighted Outcome among Controls	473.01010			490.36679		
ATT by IPTW	**4.28961****			**0.66235**		
Standard Errors of ATT by Reg	1.87455			1.37212		
N	54,680			71,905		

[1] Mean of scientific performance for supported and treated individuals.

[2] Mean of scientific performance of control group individuals matched with supported and treated individuals.

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Table 9: Within-Cluster Propensity Score Matching & Inverse Probability Weighting Method - Have-Plan

(Weighted) OLS Regressions: All Samples vs. Within-Country Matched Sample Considering unobserved country-level confounders - HavePlan														
							2015							
	All Samples		One to One With Replacement Matched Sample				One to One No Replacement Matched Sample				One to Two With Replacement Matched Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE
HavePlan	8.388**	5.882*	**4.048**	**4.048***	4.334*	4.344**	1.335	1.335	2.674	3.961**	3.344	3.344	3.891*	3.821**
	(3.267)	(3.153)	(2.748)	(2.367)	(2.454)	(2.031)	(2.575)	(2.209)	(2.289)	(1.866)	(2.570)	(2.249)	(2.277)	(1.912)
_cons	476.1***	467.8***	472.5***	472.5***	475.7***	454.9***	478.9***	478.9***	477.4***	463.6***	473.2***	473.2***	477.6***	457.0***
	(6.635)	(6.686)	(3.258)	(2.633)	(7.931)	(6.432)	(2.471)	(2.007)	(7.471)	(5.747)	(3.103)	(2.540)	(7.797)	(6.171)
N	60,139	60,139	101,238	101,238	101,238	101,238	7,388	7,388	7,388	7,388	202,476	202,476	202,476	202,476
R^2	0.272	0.356	0.000	0.251	0.214	0.456	0.000	0.257	0.223	0.464	0.000	0.248	0.215	0.455
							2018							
	All Samples		One to One With Replacement Matched Sample				One to One No Replacement Matched Sample				One to Two With Replacement Matched Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE
HavePlan	-0.651	0.510	**2.060**	**2.060**	1.768	1.678	-0.682	-0.682	-0.111	0.497	1.192	1.192	0.578	0.287
	(3.176)	(3.184)	(2.138)	(1.971)	(2.071)	(1.837)	(2.006)	(1.894)	(1.895)	(1.750)	(1.993)	(1.851)	(1.919)	(1.711)
_cons	496.2***	472.7***	488.6***	488.6***	503.7***	482.3***	492.2***	492.2***	498.3***	482.8***	489.5***	489.5***	504.4***	482.5***
	(6.891)	(7.113)	(2.452)	(2.142)	(5.893)	(4.842)	(2.022)	(1.849)	(5.550)	(5.217)	(2.337)	(2.046)	(5.619)	(4.413)
N	71,905	71,905	130,910	130,910	130,910	130,910	12,232	12,232	12,232	12,232	261,820	261,820	261,820	261,820
R^2	0.244	0.289	0.000	0.146	0.119	0.303	0.000	0.129	0.137	0.287	0.000	0.144	0.123	0.305
Covariates	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Weights	Sampling Weight		Frequency Weight for OLS, FE on matched samples, and Sampling Weight for Propensity Score Model before Matching											

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Table 10: Regressions: All Samples vs. Within-Cluster Matched Sample - HavePlan

TABLE: Naive Propensity Score Matching & Inverse Probability Weighting Method

	Naive Propensity Score Matching					
	2015			2018		
	(1)	(2)	(3)	(4)	(5)	(6)
	One to One	One to One	One to Two	One to One	One to One	One to Two
	With Replacement	No Replacement	With Replacement	With Replacement	No Replacement	With Replacement
Outcome of Treated[1]	487.47382	477.58447	487.47382	496.32004	490.35147	496.32004
Outcome of Controls[2]	472.06192	465.61743	473.03186	477.38486	472.57153	478.25125
ATE by PSM	14.83459	12.32424	14.22602	19.25867	18.62944	18.65992
ATT by PSM	**15.41190*****	**11.96702*****	**14.44197*****	**18.93520*****	**17.77994*****	**18.06879*****
Standard Errors of ATT	1.44101	1.33000	1.29030	1.26887	1.15017	1.12406
# of obs On Common Support	60,138	25,020	60,138	71,899	33,667	71,899
Treated	47,605	12,487	47,605	55,058	16,826	55,058
Controls	12,533	12,533	12,533	16,841	16,841	16,841
# of obs Off Common Support	1	35,119	1	6	38,238	6
Treated	1	35,119	1	6	38,238	6
Controls	0	0	0	0	0	0
# of obs with weight	58,015	24,974	59,623	68,984	33,652	71,236
Treated	47,605	12,487	47,605	55,058	16,826	55,058
Controls	10,410	12,487	12,018	13,926	16,826	16,178
Sum of Weights	95,210	24,974	95,210	110,116	33,652	110,116
	Naive Propensity Score Weighting					
		2015			2018	
Weighted Outcome among Treated		487.47443			496.32138	
Weighted Outcome among Controls		474.12262			480.21814	
ATT by IPTW		**13.35180*****			**16.10323*****	
Standard Errors of ATT by Reg		1.45754			1.19942	
N		60,139			71,905	

[1] Mean of scientific performance for supported and treated individuals.

[2] Mean of scientific performance of control group individuals matched with supported and treated individuals.

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Table 11: Naive Propensity Score Matching & Inverse Probability Weighting Method - High

Table 12: Regressions: All Samples vs. Naive Matched Sample - High

(Weighted) OLS Regressions: All Samples vs. Naive Matched Sample Without considering unobserved country-level confounders - High														
2015														
	All Samples		One to One With Replacement Matched Sample				One to One No Replacement Matched Sample				One to Two With Replacement Matched Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE
High	17.702***	23.728***	**15.412*****	28.038***	11.115***	21.638**	11.967***	24.661***	11.885***	21.851***	14.442***	27.453***	10.182***	21.049***
	(1.979)	(1.893)	(1.681)	(1.513)	(1.513)	(1.242)	(1.564)	(1.345)	(1.374)	(1.188)	(1.539)	(1.386)	(1.367)	(1.138)
_cons	473.8***	458.2***	472.1***	465.8***	475.8***	454.7***	465.6***	459.3***	456.2***	455.6***	473.0***	466.5***	476.5***	456.7***
	(6.101)	(6.478)	(2.133)	(1.712)	(6.435)	(4.532)	(1.816)	(1.375)	(6.174)	(4.475)	(2.038)	(1.606)	(6.295)	(4.431)
N	60,139	60,139	95,210	95,210	95,210	95,210	24,974	24,974	24,974	24,974	190,420	190,420	190,420	190,420
R^2	0.277	0.365	0.006	0.263	0.189	0.456	0.004	0.247	0.209	0.424	0.005	0.261	0.189	0.436
2018														
	All Samples		One to One With Replacement Matched Sample				One to One No Replacement Matched Sample				One to Two With Replacement Matched Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE
High	11.049***	15.884***	**18.935*****	27.698***	12.159***	20.521***	17.780***	25.299***	14.613***	22.052***	18.069***	27.168***	11.582***	20.225***
	(2.179)	(2.196)	(1.512)	(1.337)	(1.402)	(1.157)	(1.377)	(1.178)	(1.271)	(1.038)	(1.362)	(1.190)	(1.264)	(1.032)
_cons	490.3***	464.3***	477.4***	473.0***	493.3***	470.33***	472.6***	468.8***	485.0***	467.9***	478.3***	473.7***	493.2***	469.7***
	(6.639)	(7.019)	(1.737)	(1.490)	(4.418)	(3.635)	(1.453)	(1.241)	(4.481)	(3.743)	(1.643)	(1.392)	(4.298)	(3.490)
N	71,905	71,905	110,116	110,116	110,116	110,116	33,652	33,652	33,652	33,652	220,232	220,232	220,232	220,232
R^2	0.246	0.294	0.010	0.149	0.128	0.302	0.008	0.141	0.128	0.296	0.009	0.149	0.128	0.302
Covariates	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Weights	Sampling Weight		Frequency Weight for OLS, FE on matched samples, and Sampling Weight for Propensity Score Model before Matching											

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

TABLE:Within-Country Propensity Score Matching & Inverse Probability Weighting Method						
	Naive Propensity Score Matching					
	2015			2018		
	(1) One to One With Replacement	(2) One to One No Replacement	(3) One to Two With Replacement	(4) One to One With Replacement	(5) One to One No Replacement	(6) One to Two With Replacement
Outcome of Treated[1]	491.25702	466.46210	491.25702	498.95117	479.81860	498.95117
Outcome of Controls[2]	472.55237	454.35678	472.67889	481.30222	465.63928	480.89636
ATE by PSM	18.65671	15.65305	18.58409	18.16547	17.56408	18.59498
ATT by PSM	**18.70464**	**12.10533**	**18.57813**	**17.64897**	**14.17933**	**18.05482**
Standard Errors of ATT	-	-	-	-	-	-
# of obs On Common Support	60,017	24,573	60,017	71,791	33,191	71,791
Treated	47,484	12,040	47,484	54,950	16,350	54,950
Controls	12,533	12,533	12,533	16,841	16,841	16,841
# of obs Off Common Support	122	35,566	122	114	38,714	114
Treated	122	35,566	122	114	38,714	114
Controls	0	0	0	0	0	0
# of obs with weight	57,350	24,080	59,105	67,816	32,700	70,449
Treated	47,484	12,040	47,484	54,950	16,350	54,950
Controls	9,866	12,040	11,621	12,866	16,350	15,499
Sum of Weights	94,968	24,080	94,968	109,900	32,700	109,900
	Naive Propensity Score Weighting					
	2015			2018		
Weighted Outcome among Treated	491.34027			499.00183		
Weighted Outcome among Controls	473.13196			481.95532		
ATT by IPTW	**18.21861*****			**16.84323*****		
Standard Errors of ATT by Reg	1.29498			0.97739		
N	60,139			71,905		

[1] Mean of scientific performance for supported and treated individuals.

[2] Mean of scientific performance of control group individuals matched with supported and treated individuals.

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Table 13: Within-Cluster Propensity Score Matching & Inverse Probability Weighting Method - High

Table 14: Regressions: All Samples vs. Within-Cluster Matched Sample - High

(Weighted) OLS Regressions: All Samples vs. Naive Matched Sample
Considering unobserved country-level confounders - High

	All Samples		One to One With Replacement Matched Sample				One to One No Replacement Matched Sample				One to Two With Replacement Matched Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE
High	17.702***	23.728***	**18.639*****	18.639***	18.704***	18.755***	12.945***	12.945***	17.038***	19.086***	18.458***	18.458***	18.544***	18.543***
	(1.979)	(1.893)	(1.776)	(1.566)	(1.701)	(1.399)	(1.448)	(1.289)	(1.327)	(1.160)	(1.611)	(1.417)	(1.504)	(1.218)
_cons	473.8***	458.2***	468.7***	468.7***	468.4***	458.2***	466.7***	466.7***	468.6***	464.2***	468.9***	468.9***	465.4***	456.5***
	(6.101)	(6.478)	(2.363)	(1.847)	(7.037)	(5.261)	(1.846)	(1.366)	(6.243)	(4.491)	(2.237)	(1.718)	(6.637)	(4.824)
N	60,139	60,139	94,968	94,968	94,968	94,968	24,080	24,080	24,080	24,080	189,936	189,936	189,936	189,936
R^2	0.277	0.365	0.008	0.282	0.189	0.446	0.004	0.248	0.196	0.411	0.008	0.281	0.193	0.448
							2018							
	All Samples		One to One With Replacement Matched Sample				One to One No Replacement Matched Sample				One to Two With Replacement Matched Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE	OLS	FE
High	11.049***	15.884***	**17.530*****	17.530***	16.899***	16.955***	14.908***	14.908***	18.896***	19.955***	17.999***	17.999***	17.243***	17.360***
	(2.179)	(2.196)	(1.485)	(1.380)	(1.376)	(1.208)	(1.167)	(1.110)	(1.077)	(1.003)	(1.320)	(1220)	(1.232)	(1.077)
_cons	490.3***	464.3***	478.7***	478.7***	489.0***	466.7***	473.8***	473.8***	488.6***	475.5***	478.3***	478.3***	490.0***	467.4***
	(6.639)	(7.019)	(1.903)	(1.594)	(5.194)	(4.213)	(1.456)	(1.256)	(4.174)	(3.570)	(1.780)	(1.453)	(4.870)	(3.857)
N	71,905	71,905	109,900	109,900	109,900	109,900	32,700	32,700	32,700	32,700	219,800	219,800	219,800	219,800
R^2	0.246	0.294	0.008	0.162	0.123	0.313	0.006	0.142	0.131	0.286	0.009	0.164	0.123	0.313
Covariates	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Weights	Sampling Weight		Frequency Weight for OLS, FE on matched samples, and Sampling Weight for Propensity Score Model before Matching											

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Underidentification test (Kleibergen-Paap rk LM statistic) H_0: the instrument is not correlated with the endogenous regressor				
	2015			
Treatment	HavePlan		High	
Instrument	Guide1	**Guide2**	Guide1	Guide2
IV-2SLS	22.006***	**19.442*****	0.590	2.426
2SLS-FE	2.055	**3.445***	0.073	1.044
	2018			
Treatment	HavePlan		High	
Instrument	**Guide1**	Guide2	Guide1	Guide2
IV-2SLS	**14.899*****	3.633*	3.538*	1.679
2SLS-FE	**5.194****	0.465	0.399	0.199

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 15: Underidentification test for Instrumental Variables

	2015		2018	
	(1)	(2)	(3)	(4)
	IV-2SLS	IV-2SLS-FE	IV-2SLS	IV-2SLS-FE
	science	science	science	science
HavePlan	-406.616**	-165.179	-66.166	-441.638*
	(171.468)	(328.816)	(106.364)	(236.116)
_cons	842.640***		549.443***	
	(152.117)		(87.843)	
With Covariates	Yes	Yes	Yes	Yes
With Interaction Terms	No	No	No	No
N	60,139	60,139	71,905	71,905
adj.R^2	-0.858	-0.025	0.181	-2.388
Underidentification test	19.442***	3.445*	14.899***	5.194**
Cragg-Donald Wald F statistic	87.234	16.239	114.317	47.278
Kleibergen-Paap rk Wald F statistic	19.392	3.481	15.847	5.512
Sargan statistic	0.000 (equation exactly identified)			

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 16: IV-2SLS Model: Treatment - HavePlan

	2015		2018	
	(1)	(2)	(3)	(4)
	IV-2SLS	IV-2SLS-FE	IV-2SLS	IV-2SLS-FE
	science	science	science	science
High	-692.867	-204.657	105.708	-1328.601
	(490.225)	(450.462)	(185.365)	(2084.098)
_cons	863.047***		444.400***	
	(270.750)		(89.621)	
With Covariates	Yes	Yes	Yes	Yes
With Interaction Terms	No	No	No	No
N	60,139	60,139	71,905	71,905
adj.R^2	-8.052	-0.857	0.013	-39.771
Underidentification test	2.426	1.044	3.538*	0.399
Cragg-Donald Wald F statistic	11.933	4.267	26.429	3.119
Kleibergen-Paap rk Wald F statistic	2.394	1.039	3.540	0.403
Sargan statistic	0.000 (equation exactly identified)			

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 17: IV-2SLS Model: Treatment - High

Robustness Check: Adding Science-related Variables, School Type and Parents' Occupation								
	2015							
	HavePlan				High			
	OLS	FE	2SLS	2SLS-FE	OLS	FE	2SLS	2SLS-FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	7.459*	4.862	-347.752**	-160.396	15.217***	20.399***	-888.699	-189.498
	(3.828)	(3.751)	(137.945)	(209.067)	(2.182)	(2.111)	(763.873)	(267.145)
_cons	436.178***	444.675***	727.742***	-	433.413***	436.613***	961.328**	-
	(26.586)	(24.558)	(127.478)	-	(26.626)	(24.838)	(474.132)	-
N	41,399	41,399	41,399	41,399	41,399	41,399	41,399	41,399
R^2	0.308	0.386	-0.495	0.060	0.311	0.392	-13.083	-0.591
	2018							
	HavePlan				High			
	OLS	FE	2SLS	2SLS-FE	OLS	FE	2SLS	2SLS-FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	1.491	3.320	-68.626	-419.441*	13.348***	18.536***	175.581	-1725.953
	(3.816)	(3.788)	(98.378)	(232.851)	(2.673)	(2.655)	(312.106)	(3859.713)
_cons	449.594***	442.389***	514.637***	-	446.993***	439.573***	398.586***	-
	(26.783)	(28.851)	(98.312)	-	(25.409)	(26.952)	(96.753)	-
N	52,103	52,103	52,103	52,103	52,103	52,103	52,103	52,103
R^2	0.241	0.288	0.173	-2.120	0.245	0.294	-0.401	-66.233

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Table 18: Robustness Check for Linear Models

Robustness Check for PSM: First-order and Second-order Covariates				
	2015			
	HavePlan		High	
	Naive PSM (1)	Within-Country PSM (2)	Naive PSM (1)	Within-Country PSM (2)
ATT	10.038***	2.095	14.735***	19.165
	(2.332)	-	(1.450)	-
Outcome of Controls	473.137	474.977	472.739	472.023
	-	-	-	-
On Support	60,098	53,858	60,108	59,945
	2018			
	HavePlan		High	
	Naive PSM (1)	Within-Country PSM (2)	Naive PSM (1)	Within-Country PSM (2)
ATT	2.101	1.843	18.654***	15.412
	(1.882)	-	(1.326)	-
Outcome of Controls	488.533	489.172	477.655	483.466
	-	-	-	-
On Support	71,868	71,139	71,894	71,770

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

The 'psestimate' command in Stata is used to select the first- and second-order forms of covariates that achieve the best fit by comparing the maximum likelihood values of different models.

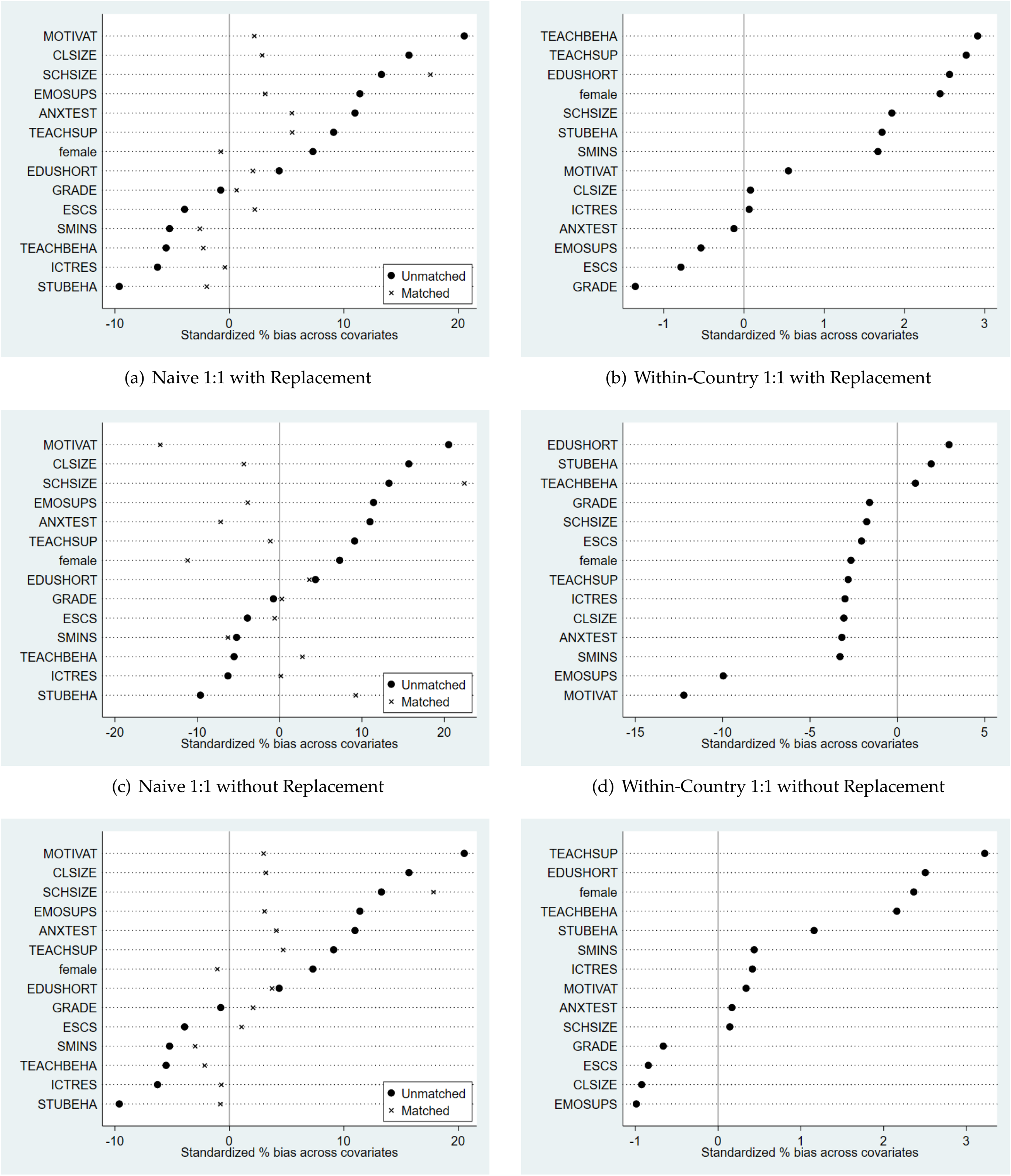
Table 19: Sensitivity Analysis: 1:1 Propensity Score Matching

Heterogeneity Analysis: Interaction of Treatment and Countries								
	2015							
	HavePlan				High			
	OLS	FE	2SLS	2SLS-FE	OLS	FE	2SLS	2SLS-FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	6.906**	6.400*	-291.161**	-209.537	16.101***	23.396***	-339.864*	-205.588
	(3.290)	(3.416)	(128.683)	(427.750)	(2.042)	(2.029)	(182.600)	(452.040)
c.Treatment#c.HKG	57.743***	-7.177	78.289***	207.789	57.214***	-11.722***	138.194***	207.704
	(4.249)	(5.824)	(9.638)	(426.041)	(4.276)	(4.007)	(41.100)	(433.245)
c.Treatment#c.DEU	17.878***	-5.448	49.529***	209.750	25.375***	5.412	123.852**	223.041
	(4.394)	(6.533)	(14.090)	(426.227)	(4.323)	(4.369)	(50.266)	(429.631)
_cons	474.932***	467.592***	733.733***	-	472.806***	458.348***	660.293***	-
	(6.700)	(6.713)	(112.599)	-	(6.125)	(6.494)	(97.566)	-
N	60,139	60,139	60,139	60,139	60,139	60,139	60,139	60,139
R^2	0.275	0.356	-0.304	-0.137	0.311	0.365	-1.779	-0.775
Underidentification Test for IV	-	-	28.223***	2.195	-	-	6.916***	1.047
Cragg-Donald Wald F statistic	-	-	140.615	11.139	-	-	39.584	4.596
Kleibergen-Paap rk Wald F statistic	-	-	27.399	2.214	-	-	6.678	1.041
	2018							
	HavePlan				High			
	OLS	FE	2SLS	2SLS-FE	OLS	FE	2SLS	2SLS-FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-1.546	0.488	-25.377	-440.697*	9.848***	14.896***	136.710	-1099.89
	(3.220)	(3.306)	(78.568)	(234.901)	(2.229)	(2.292)	(866.487)	(1.44e+03)
c.Treatment#c.HKG	45.756***	18.262***	48.936***	454.442*	45.211***	3.844	7.811	1.07e+03
	(4.004)	(6.196)	(10.730)	(232.355)	(4.105)	(4.222)	(255.906)	(1.38e+03)
c.Treatment#c.DEU	13.067**	-1.802	16.647	435.149**	23.941***	17.714***	-12.847	1.07e+03
	(5.256)	(9.158)	(12.378)	(233.022)	(5.165)	(5.848)	(251.933)	(1.36e+03)
_cons	494.924***	472.692***	513.864***	-	489.099***	464.648***	430.056	-
	(6.959)	(7.112)	(63.824)	-	(6.686)	(7.017)	(402.249)	-
N	71,905	71,905	71,905	71,905	71,905	71,905	71,905	71,905
R^2	0.245	0.289	0.232	-2.260	0.248	0.294	-0.907	-25.71569
Underidentification Test for IV	-	-	25.962***	5.238**	-	-	0.219	0.579
Cragg-Donald Wald F statistic	-	-	214.835	49.771	-	-	1.789	4.774
Kleibergen-Paap rk Wald F statistic	-	-	27.455	5.558	-	-	0.220	0.585

Standard errors in parentheses

* $p < 0.10$,** $p < 0.05$,*** $p < 0.01$

Table 20: Heterogeneity Analysis: Regression with Interactions between Treatments and Certain Countries



(a) Naive 1:1 with Replacement

(b) Within-Country 1:1 with Replacement

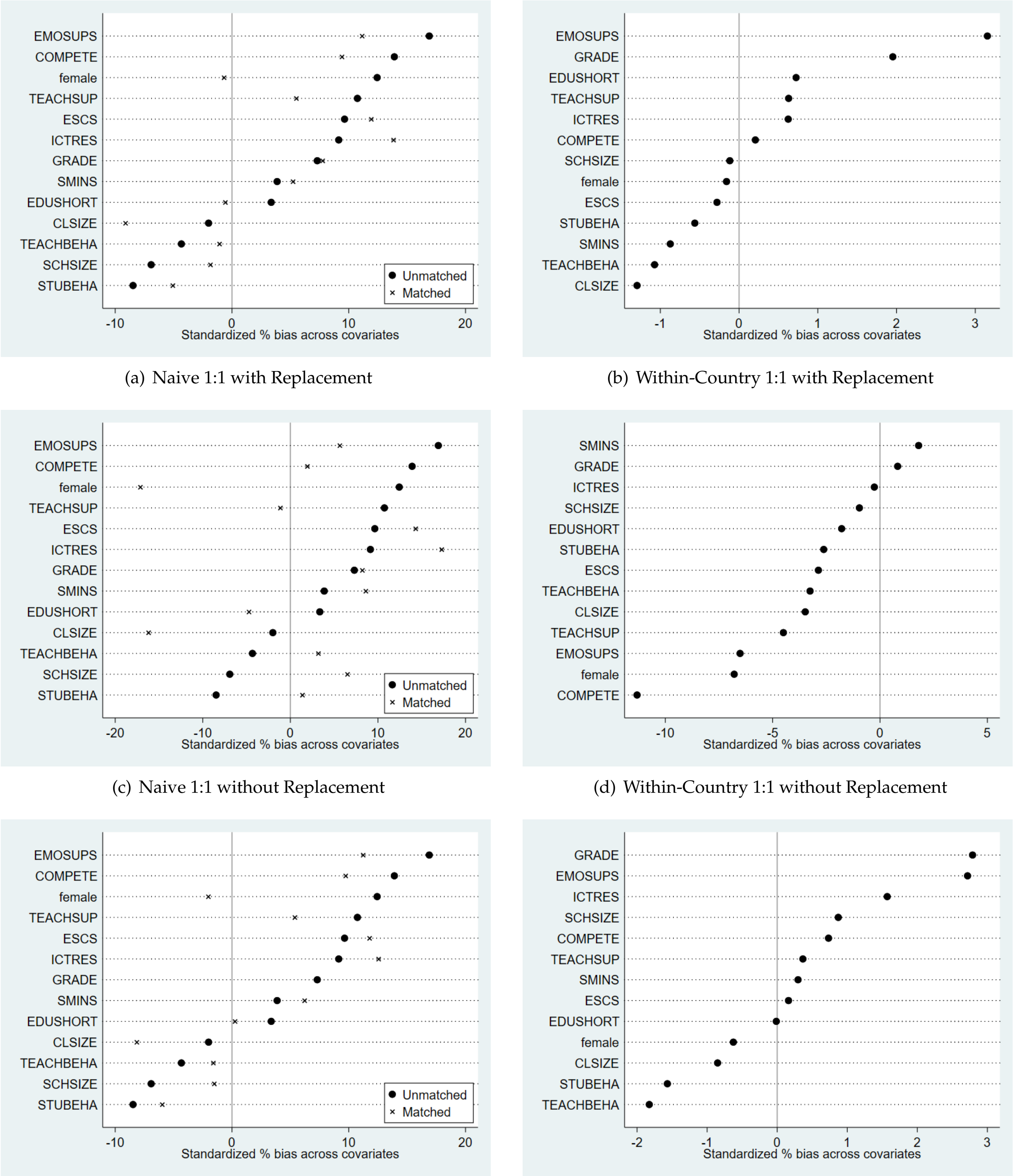
(c) Naive 1:1 without Replacement

(d) Within-Country 1:1 without Replacement

(e) Naive 1:2 with Replacement

(f) Naive 1:2 with Replacement

Figure 1: NAIVE PSM VS. WITHIN-COUNTRY PSM - STANDARDIZED % BIAS ACROSS COVARIATES - HAVEPLAN 2015



(a) Naive 1:1 with Replacement

(b) Within-Country 1:1 with Replacement

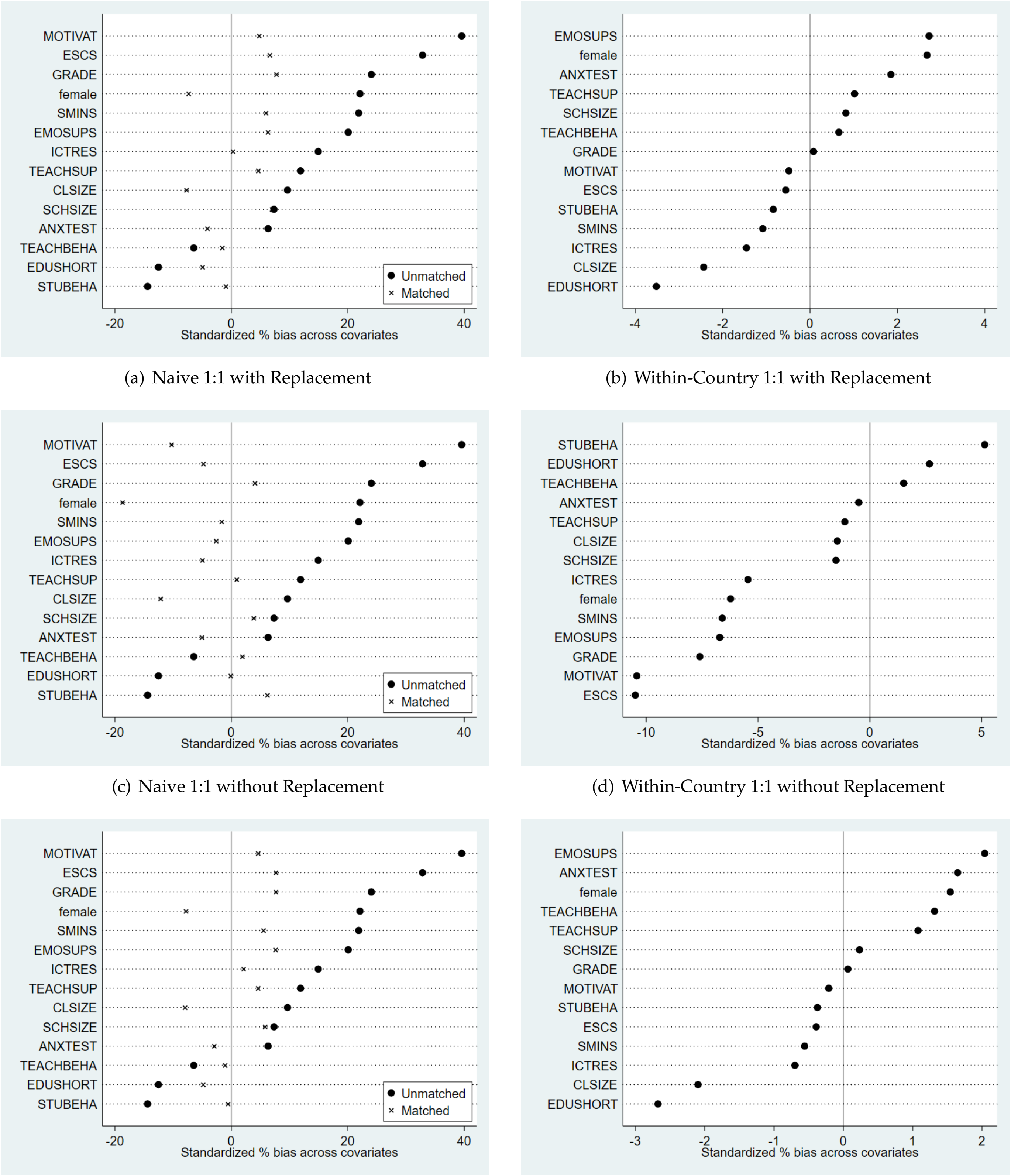
(c) Naive 1:1 without Replacement

(d) Within-Country 1:1 without Replacement

(e) Naive 1:2 with Replacement

(f) Within-Country 1:2 with Replacement

Figure 2: NAIVE PSM VS. WITHIN-COUNTRY PSM - STANDARDIZED % BIAS ACROSS COVARIATES - HAVEPLAN 2018



(a) Naive 1:1 with Replacement

(b) Within-Country 1:1 with Replacement

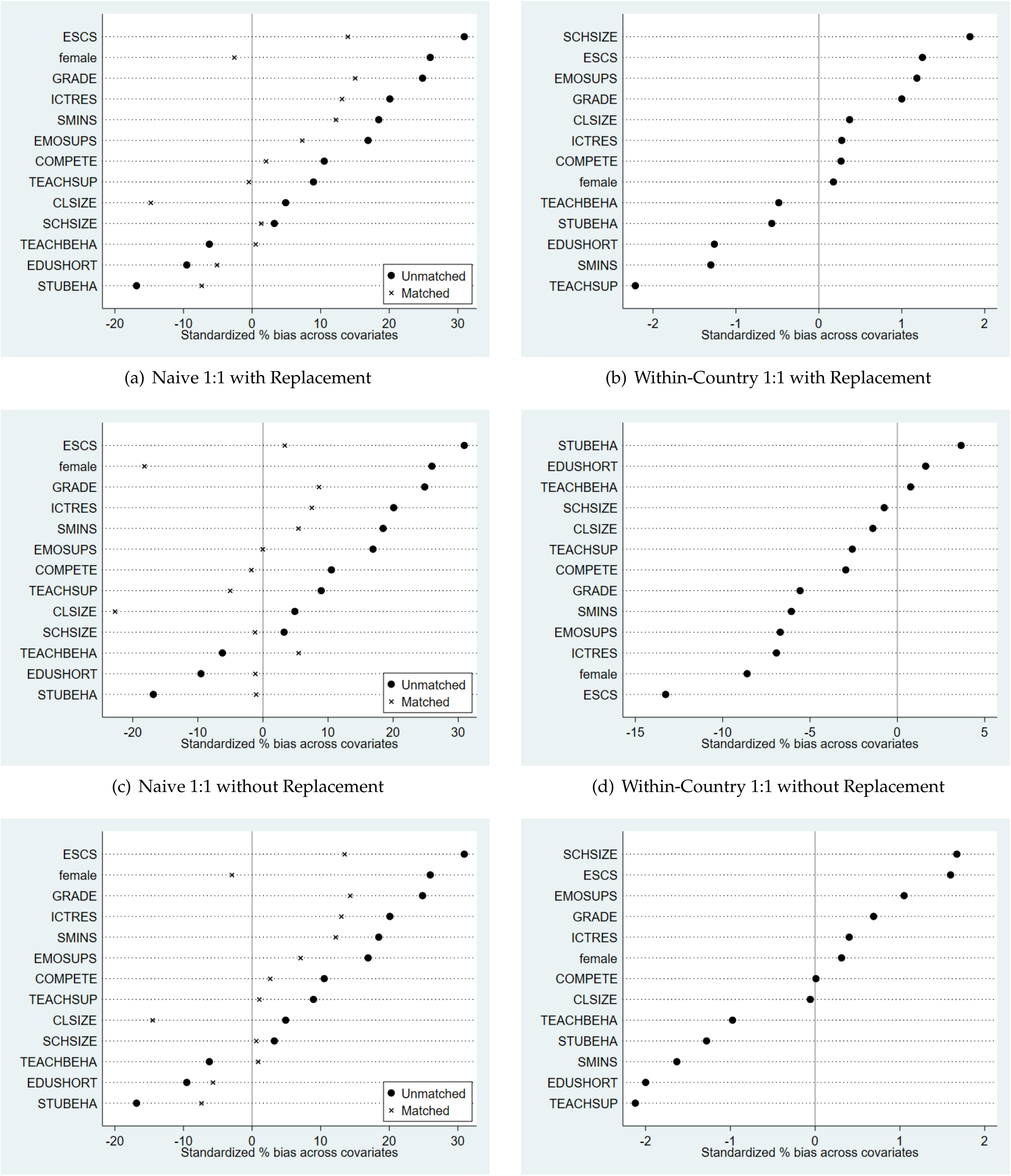
(c) Naive 1:1 without Replacement

(d) Within-Country 1:1 without Replacement

(e) Naive 1:2 with Replacement

(f) Within-Country 1:2 with Replacement

Figure 3: Naive PSM vs. Within-Country PSM - Standardized % Bias across Covariates - High 2015



(a) Naive 1:1 with Replacement

(b) Within-Country 1:1 with Replacement

(c) Naive 1:1 without Replacement

(d) Within-Country 1:1 without Replacement

(e) Naive 1:2 with Replacement

(f) Within-Country 1:2 with Replacement

Figure 4: NAIVE PSM VS. WITHIN-COUNTRY PSM - STANDARDIZED % BIAS ACROSS COVARIATES - HIGH 2018