THE FAILURE OF THE BALANCED CONDITION IN THE NATURAL EXPERIMENT DESIGN

Chen Wang
Research School of Finance, Actuarial Studies and Applied Statistics,
The Australian National University

One important task in economics and finance studies is establishing causal inference. Given the obstacles of obtaining reliable instrumental variables, methodologies such as ordinary least squares (OLS) that fall under the umbrella of natural experiments by taking advantage of exogenous events are becoming prominent, without questioning on the balanced condition hypothesis. One empirical problem with these methodologies is that the treatment assignment is not random, which is characterized by non-balanced covariates across the treatment and control groups. This problem is often not obvious to researchers, and they may infer causality when in fact none may exist. By employing the examples from influential journals, we show that the causal inferences from the natural experiment studies may change after we deal with the imbalanced condition problem. We argue that the data quality will affect the estimates of the treatment effect: a better-balanced dataset will require fewer matching activities, in doing so, the estimation results after dealing with the imbalanced condition problem become less volatile compared with those from low quality dataset. Notwithstanding the popularity of nature experiment technique, according to our knowledge, such results are not available in the previous literature.

Keywords: Entropy Balancing, Natural Experiment, Causal Inference, Propensity Score Matching

JEL Codes: B41, C13, C18, C80

"It may be that Campbell and Stanley (1963) should feel guilty for having contributed to giving quasi-experimental designs a good name. There are program evaluations in which the authors say proudly, 'We used a quasi-experimental design.' If responsible, Campbell and Stanley should do penance, because in most social settings, there are many equally or more plausible rival hypotheses."

Campbell and Boruch (1975, 202)

1. Introduction

One important task in economic and financial studies is establishing causal inference. Although causal inference is a simple concept that is a process by which we can claim causality (i.e., one event can cause another), the establishment of such a relationship is never easy. Admittedly, the interpretation of parameters require caution as variation in an independent variable whose effects on the dependent variable is of interest can be determined by unobserved covariates that jointly influence the dependent variable. The unobserved factors can involve pre-existing conditions that differ between treatment and control groups in evaluation of treatment or program studies. Such unobserved factors imply that the estimates obtained from such studies are open to hundreds or even thousands of alternative explanations.

Given the challenges of obtaining reliable instrumental variables to answer the causality of interest, researchers have increasingly exploited natural experiments. In a natural experiment, the assignment of treatment to subjects is serendipitous and random (Freedman, Pisani, and Purves, (1997)). Such alleged natural experiments are generally variations in rules controlling behaviour that are supposed to satisfy random criteria.

In a simplified natural experiment, subjects are randomly divided into treatment groups (observations are exposed to the treatment condition) and control groups (no treatment condition). This random assignment prevents subjects from being selected into particular groups and consequently ensures that treatment and control groups are as similar as possible with respect to measured and un-measured characteristics. Owing to the fact that the only difference between the treatment and control groups is the assignment to the treatment condition, which occurs randomly (which is occasionally called exogeneity), the comparison of outcomes between the two groups provides causal inferences on treatment (Dunning, (2008)).

As the mechanism of treatment assignment is ambiguous and not controlled by the researcher, most natural experiment studies in economics, finance, and other social science disciplines are criticized based on the problem that the assumption of random assignment to the subjects is not credible (Diamond and Robinson, (2010); Angrist and Krueger, (1991); Dunning, (2008, 2010); Gregory, McNulty and Krasno, (2009); Sekhon, (2009); Sekhon and Titiunik, (2012); Dunning, (2012); Rozenzweig and Wolpin, (2000)). Unfortunately, as several commonly used methods (e.g., OLS) are not designed to focus on balance checking for covariates in studies with binary treatments, numerous applications have failed to determine the potential problem of distinct characteristics between treatment and control groups, which is the central dilemma—that is, the randomness assumption (or the balanced condition assumption) in a natural experiment does not hold. Such an imbalanced condition issue perhaps explains why "as if" "causality relationship" can be obtained by such a problematical "natural experiment" in which indeed such causal inference may be different from the original study since the groups compared are fundamentally heterogeneous in terms of their characteristics.

The challenge in finding balanced and comparable observations from observational dataset possibly explains why many natural experiment studies do not check and report the imbalance problems and run regressions in the first place. Imbens (2010) argue that researchers are discouraged to explore certain economic and finance questions since randomization is difficult or even impossible to achieve. Atanasov and Black (2016) review 36 shock-based studies on causal inference in the corporate finance discipline and find that only three of them analyse the risk of non-random assignment problem. Even the table of statistics compares the means between treatment and control groups, while the problems caused by higher moments imbalanced condition or joint imbalanced condition (covariates are not balanced jointly between treatment and control groups) are seldom discussed.

Proceeding without balance checking is a mistake as the criterion of randomness in the nature experiment is not satisfied. Thus, the researcher is left to worry that the "causality" established from the regression method (without a balance-checking process) may produce imprecise estimations, since it may actually stem from the heterogeneous characteristics in the samples compared.

The goal of this paper is to investigate the balanced condition required by the natural experiment study. If the estimation process, such as ordinary least squares (OLS), falls under the umbrella of natural experiments by taking advantage of "exogenous events", the balanced condition assumption is easily ignored by the researchers because of the natural of the randomness in such experiment design.

We classify this issue by asking the following two questions: (1) In a natural experiment study, are the proposed treatment and control groups similar in terms of characteristics? (2) If not, will the causal inference change after we deal with the imbalanced condition problem? To achieve this, we provide empirical examples that appeared in the top Accounting, Economic and Finance journals in which the answer to the first question is "no." For the second question, we compare the estimation results before and after we deal with the imbalanced condition problem and present evidence that the data quality will affect the estimates of the treatment effect: a better-balanced dataset will require fewer matching activities, in doing so, the estimation results after dealing with the imbalanced condition problem become less volatile compared with those from low quality dataset.

Our first investigation focuses on the imbalance problem in the nature experiment studies that scholars may ignore in academic journals. We review 372 empirical natural experiment studies in Economics, Finance, and Accounting journals, only 152 (40.86%), 65 (17.52%) and 31 (8.36%) papers check the first moment, higher moments, and joint imbalanced problems respectively.

In order to figure out what can be wrong if the balanced condition is not satisfied as required by natural experiment, we produce Monte Carlo simulations, where covariates in treatment and control groups are distinct and the randomness assumption does not hold, as the major problem we argue that scholars may ignore in practice. We randomly draw two covariates for the control group from a uniform distribution (0,10) as well as two covariates for the treatment group from a uniform distribution (3,13)[1]. This design provides the overlapping observations for a [3,10]×[3,10] square as a random dataset, and the covariates in the treatment group are fundamentally higher than those in the control group. For estimation purpose, we try to explore ways to combine a direct matching method, Entropy balancing (EB) (Hainmueller (2012)) with the natural experiment method (combined method) to estimate the differential effects between the imbalanced original dataset and the balanced dataset after the combined method step. The results present that the estimations of the original imbalanced dataset implemented by Ordinary Least Square (OLS), a commonly used method in natural experiment study, suffers from the model dependence problem since with different model assumptions (with respect to different number of the regressors included in the model) the estimation results become much more volatile compared with the results after we deal with the imbalanced condition problem via the combined method step.

In addition, we investigate if an indirect matching method-Propensity Score

[1] Including more covariates will not change the conclusion and will make the conclusion stronger.

Matching (PSM), combined with the OLS (OLS is commonly used in the natural experiment study) is appropriate in our case to deal with the imbalanced condition problem in natural experiment design. PSM is widely used in observational studies to deal with the problem of self-selection (non-randomness problem). According to Ho et al. (2007), Sekhon (2009), PSM methods are commonly used for estimation of binary treatment effects based on the assumption of the selection of observables. One of the main criticisms of the PSM method is that, in practice, it can be challenging to ensure that the distributions of the propensity scores of the control and treatment groups are balanced (see for example Rosenbaum and Rubin, (1983, 1985); Dehejia and Wahba, (2002) for tests to check for balance). We investigate the balanced condition required by the PSM and extend the analysis in our studies to show that PSM may not produce both local balanced condition (if each individual covariate is the same between compared groups) and global balanced condition (if all the covariates are balanced jointly between treatment and control groups) if the analysed data quality is low (the natural experiment is not random).We provide the concluding remarks that estimation performance after PSM matching (combined with OLS) may be even worse than the OLS (used in original study) without matching step if the balanced condition required by the PSM does not hold, since the Mean Squared Error (MSE) for PSM method is much higher than that of the OLS method.

We show in the text with an intuitive example on when the PSM combined method may perform worse than the original OLS study without matching step: PSM combined method initially does quite well regarding its estimation performance when the balanced condition is met in random dataset since its estimated parameter (1.981) is the same as OLS (1.981). However, as the balanced condition required by PSM deteriorates, the PSM combined method’s performance suffers compared with OLS without matching step, and the potential MSE with PSM combined method goes up dramatically compared with that from OLS.

In our reviewed 372 natural experiment studies, conditional on the Propensity Score Matching (PSM) method included in the papers, only 10.53% of the papers check the balanced condition required by the PSM method, in doing so, the potential estimation problem may be a concern as we illustrate in the Monte-Carlo section and the empirical study section.

We further examine different natural experiment studies claiming causal inference in influential economic, financial and accounting journals, and compare the estimation performance between their original results (without matching step) and the results after dealing with the imbalanced condition problem with our combined method (direct matching combined with regression).

Our first influential example is to use the Sarbanes–Oxley Act (SOX) as a natural experiment. For instance, concerning the corporate board structure is an endogenous variable, an experiment design proposed by Chhaochharia and Grinstein (2009) employs SOX as a natural experiment to study the cause inference of board independence on CEO compensation. Such a "natural experiment" is widely used in the literature:[2] For example, see Duchin, Matsusaka, and Ozbas (2010); Chen (2014); Chen, Cheng, and Wang (2015); and Armstrong, Core, and Guay (2014), among others.

Chhaochharia and Grinstein (2009) explore the fact that SOX requires NYSE and NASDAQ to propose the regulation that all of the listed firms should have more than 50% independent board of directors. Despite its strength of design, that is, firms which did not satisfy this requirement were serendipitously forced to increase their board independence level (thus, the variation in board structure of the affected firms appears exogenous), this experiment lacks balance checking in terms of the characteristics between the groups

[2] Atanasov and Black (2016) review 36 shock-based natural experiments on causal inference in corporate finance literature, 17 of which are driven by SOX regulation.

compared in the study. Indeed, a part of the predicament in this design is that the assignment of a treatment variable to the subjects is not random as the evidence shows that the characteristics are significantly different among the groups compared. To overcome this challenge, we deal with the imbalanced problem and illustrate that the treatment effect is insignificant and 25% lower than the report in the original study where the imbalanced condition problem exits. To evaluate the statistical properties of these methods, we implement bootstrap simulation tests on this experiment design and show that our proposed direct matching combined method leads to lower bias, and a lower MSE in comparison to the OLS method (used in the original natural experiment study) and the commonly used matching scheme (i.e., PSM method).

Our second influential example is provided by Angrist et al. (2002) who investigate the causal effect of using school vouchers (distributed randomly by lotteries) on the number of hours that the winners of the lotteries work every week. This experiment is also used in Huber, Laffers, and Mellace (2017), Uribe et al. (2006), Angrist, Bettinger, and Kremer (2006), and Bettinger, Kremer, and Saavedra (2010), among others. Angrist et al. (2002) conclude that the winners work less than their peers who lose in the lotteries. Although the randomness assumption is still a concern since Rubin’s B ratio is 30.4, which indicates that the samples compared are not sufficiently and jointly balanced with respect to the background characteristics, the quality of the dataset is much better compared with previous case since most of the covariates are quite similar between compared groups. Hence, the estimation process requires fewer matching activities, in doing so, the estimation results after dealing with the imbalanced condition problem become less volatile

We evaluate the performance of those methods based on the precision of coefficients. Although such a comparison does not include all the statistical properties, we focus on the bias and MSE, and the results show that our proposed direct matching combined method dominates OLS (without matching step) and PSM (combing with the OLS method). Nevertheless, because the data quality is much better than the previous two examples, the estimation results after dealing with the imbalanced condition problem become less volatile, in doing so, the data quality matters with respect to the estimation performance which is consistent with out Monte-Carlo study.

As most of the natural experiment studies in Accounting, Economics and Finance disciplines have the model dependent problem, at least partly, this means that the estimation result depends on the assumption of the model specification (i.e., the assumption on what types of independent variables are included in the model) (Ho et al. 2007). We make use of the "National Job Training Partnership Act (JTPA)," an influential randomized experiment to evaluate the performance of a methodology (i.e., Heckman et at. (1997) among others) to show that if our estimation results will change much with our combined method. The JTPA is devolved by the US Department of Labor to evaluate the effect of the JTPA on program participants' earnings. The JTPA randomly choses participants according to their background characteristics who were qualified to apply to the program in 16 service delivery areas in the United States of America, such that part of the qualified candidates is selected into the treatment group randomly, whereas the others are automatically left to be in the control group.

In contrast to Heckman et at. (1997)'s study who focus on the self-selection bias issue and calculate the probability that if any observation will receive the JTPA program, we combine a direct matching method with OLS method and figure out that our combined method has better performance on reducing the model dependency problem for the estimation of treatment effect (the treatment effect is 1051 with less independent variables and is 1049 with more independent variables), in contrast to OLS without dealing with imbalanced condition step (the treatment effect is 1117 with less independent variables, and is 970 with more independent variables) as well as a combined of PSM and OLS method (the treatment effect is 1197 with less independent variables, and is 1059 with more independent variables)

Concerns about if our combined method in natural experiment study is suitable to be applied in Accounting, Economics and Finance disciplines are consistent with similar methodology studies (i.e., Kullback (1968), Imbens, Spady, and Johnson (1998), Diamond and Sekhon, (2006), Abadie and Imbens (2007), Ho et al. (2007), and Qin, Zhang, and Leung (2009)). We evaluate our combined method with a job training program, the National Supported Work Demonstration (NSW), which was first introduced by LaLonde (1986) and then developed by Dehejia and Wahba (2002). The idea of this experiment is to obtain a benchmark for treatment effect on increased earnings from the NSW experiment. Further, the control group in the randomized experiment is replaced by individuals (obtained from a non-experimental Current Population Survey) who are not involved in the training program and determine which method using the non-experimental data can recover the benchmark of treatment effect of job training program on salary (obtained from NSW).

We expand the NSW analysis and compare the performance of the OLS without balancing checking step with our proposed method (combing direct matching scheme with OLS) and the PSM method (combined with OLS).

In this design, the estimates by OLS without matching step (1672.426) and PSM combing with OLS method (1669.661) are significantly lower than the threshold (1794) set by Dehejia and Wahba (2002) in comparison to the direct matching combing with OLS estimation (1776.685).

In this study, we focus on the imbalanced condition required by the natural experiment design. We make several contributions. First, we investigate the imbalanced problems appeared in the top Economics, Finance and Accounting journals which are easily ignored by the scholars in the natural experiment studies. For our investigation on imbalanced condition we consider four benchmarks: the imbalanced condition in first moment, the imbalanced condition in higher moments, the joint imbalanced condition and the imbalanced propensity scores conditional on the PSM is employed. Secondly, we try to explore ways to combine the direct matching method with the natural experiment method (combined method) to estimate the treatment effects. The typical estimation strategy in direct matching scheme is to calculate the mean differences between the treatment and control groups as the average treatment effect, without focusing on the true economic question where the economic structural models can provide the useful information of interest. Heckman and Urzua (2009) argue that even the perfect randomization experiment cannot answer the question of economic interest because of the lack of the economic mechanism analysis executed by structure models. Together these direct matching results allow us to analyse the economic question of interest with the structure models. A third contrition is to investigate whether the indirect matching scheme, the Propensity Score Matching (PSM) method can be used to combine with the OLS in the natural experiment. We provide evidence that estimation performance after PSM matching (combined with OLS) may be worse than the OLS (used in original study) without matching step, if the balanced propensity scores condition required by PSM is not satisfied. Our fourth contribution is to investigate if the data quality matters with respect to the estimation performance in natural experiment design. We identify scenarios in both Monte-Carlo simulation tests and empirical studies and illustrate that a better-balanced dataset will require fewer matching activities, in doing so, the estimation results after dealing with the imbalanced condition problem become less volatile compared with those from low quality dataset suffering from serious imbalanced condition problem. Fifth, we provide some specific recommendations for future natural experiment design, an intensively used strategy in influential Accounting, Economics and Finance journals. The major recommendation of our study is that researchers in empirical natural experiment designs should implement balance checking process between compared groups, and as a process of routine, combine the direct matching step with the natural experiment to improve the estimation with respect to the causal inference if the balanced condition is not met.

Even though our recommendation of the combined method already provides a faster and easier procedure to understand and implement in Accounting, Economics and Finance disciplines, this technique maybe limited by a comparative case study design (although this drawback is shared with similar matching methods (see Deming and Stephan (1940), Qin and Lawless (1994), Zaslavsky (1988), Schennach (2007)). First, in empirical implementation, some obscurity exists on how the observations compared are chosen. Scholars generally choose the units compared based on their subjective judgment of affinity between the affected and non-affected observations. However, such subjective choice may not produce a reliable standard. Second, the empirical designs in the comparative studies generally use the data from a sample consisting of disaggregated observations but provide statistical inference on aggregated data for the population. Hence, the real statistical inference for the population can be ignored completely if the aggregate dataset is not accessible. Even if the information on population (aggregate data) is used, uncertainty still exits on whether the control group can provide counterfactual outcome for the affected observation, which is short of reliable intervention. Such uncertainty cannot be represented by standard errors in the normal statistical inferential scheme. Third, if the control group is rather different to the treatment group, we may not find positive weights to satisfy the balanced condition, given few observations but strict constraints. Although this problem also exists in other methods, we should be careful with few overlapping problems. Fourth, there can be special cases in which the data will cause the calculated weights to be large for a group of observations. This result is caused by a limited overlap between treatment and control groups. Hence, the study heavily relies on very few observations in the control group that are rather similar to observations in the treatment group. Although it is a common problem in existing matching techniques that well matched observations in a control group are used more, researchers can also use a refined dataset or employ commonly used diagnostics to check whether the results will still hold excluding those cases.[3]

As such, we focus on the combing a direct matching scheme with natural experiment design in Accounting, Economics and Finance disciplines because it reflects robust property within misspecification environment. There are more matching schemes (i.e., Stefano, Gary, and Giuseppe (2012), Kullback (1968), Hirano, Imbens, and Ridder (2003), Hellerstein and Imbens (1999)) which also include data preprocessing for parametric or non-parametric estimation, which may provide similar results as recommended here for claiming causal inference. Above all, as such alternative techniques can preprocess through different matching algorithms, in doing so, such methodologies may not be considered as competitors.

The remainder of the paper is organized as follows. In the next section, we introduce the Monte Caro Simulation, review the balance checking problem in the literature and explain the method issues. Sections 3 to 5 examine the empirical examples that appeared in influential journals. Section 6 concludes our findings.

2 Balanced condition problem, Monte-Caro Simulation and the method issue

2.1 Balance checking problem in the literature

In order to evaluate the balance checking problem in literature, we searched the empirical natural experiment articles from the Web of Science database (from 2000 to

[3] This is not the problem of this paper as we test all the weighting schemes in each subsection and no extreme weights assigned to control groups can be found. We recommend researchers to draw the distributions of the new assigned weights to check whether some special observation receive big weight.

the first half year of 2018) using the following terms or variations: terms focusing on “Natural” or “Quasi” or “Quasi Natural” near experiment; “exogenous” near “shock” or “event”. We manually read each article that satisfied above settings and exclude the papers focusing on Regression Discontinuity Design (RDD) approach, Instrument Variable (IV) approach, the time series models’ approach, the review papers, the discussion or comments papers on previous studies, and the pure theory papers without empirical studies so that the sample of articles satisfy the method settings we argue above. This process produced 372 articles listed in table 1.

We then manually check the balanced condition required by natural experiment for each paper respect to balanced condition regarding on the first and higher moments of covariates, the global balanced condition (if the covariates are jointly balanced between treatment and control groups), whether or not the paper include PSM, a commonly used method to control for self-selection bias, if PSM method is employed, whether or not the balanced condition required by PSM is checked.

Table 2 illustrates the balanced condition check results for our sample. It provides the information on the relative ratio of the balance checked papers and the non-checked articles. The majority papers in our sample (at least 82.48%) did not report the higher moments balanced condition, global balanced condition (91.64%) and balanced condition of PSM (89.47%) if this method is employed, in doing so, the estimation performance produced by these methods may be a concern.

2.2 Monte Carlo Simulation study

To illustrate the argument regarding what can go wrong if the covariates are not balanced in the compared group in the natural experiment’s design, our investigation focuses on the Monte Carlo simulation where the covariates in the treatment and control groups are distinct so that the randomness assumption does not hold.

For intuition, we randomly draw two covariates for the control group (with 10,000 observations) from a uniform distribution (0,10) and two covariates for the treatment group (with 10,000 observations) from a uniform distribution (3,13); including more covariates does not change the implication and will make the argument even stronger. This design provides the overlapping observations as a [3,10]×[3,10] square as a random dataset, and the observations outside of this range provide imbalance to the dataset. In doing so, the covariates in the treatment group are fundamentally higher than those in control group, which is consistent with our assumption that the randomness condition does not hold in the "natural experiment". We then produce the dependent variable from the equation $z_i = y_{1i} + y_{2i} + \varepsilon_i$, where ε_i follows a standard normal distribution with mean of 0 and a standard deviation of 1. We repeat the simulation as suggested above for 100 times.

To evaluate the estimation performance produced by the direct matching (EB) scheme combined with OLS and the indirect matching (PSM) scheme combined with OLS, we use the mean difference from the [3,10]×[3,10] random experiment as a benchmark for the treatment effect, and we test which method can recover the benchmark treatment effect better if different model structure assumptions are used.

We first check the balanced condition for the unbalanced dataset and the result is reported in Table 3. The table reports the imbalance problem between the treatment and control groups on the mean differences, the quantiles difference, and the $L1$ distance for joint covariates imbalance that is calculated by coarsening the data (Iacus, King and Porro (2009)).

The diagram illustrates that almost all the measures of the covariates in the treatment group are 3 units higher than those of the control group, which is consistent with our simulation setting that the compared groups are imbalanced.

To evaluate the performance with the three methods, we have obtained the benchmark treatment effect from the mean difference from the [3,10]×[3,10] random experiment (0.044). The estimation results are reported in Table 4.

For each method, the table reports two model assumptions. One method only includes the treatment indicator as the independent variable (coded as unadjusted), and the other method includes both the treatment indicator and the two covariates as regressors (coded as adjusted).[4]

The results present that both the OLS (without matching step) and PSM (combined with OLS) suffer from the model dependence problem since with different model assumptions (with respect to the number of regressors included in the model) the estimation results change a lot. This is contrast to the direct matching method (EB combined with OLS), which provides very stable estimation results (0.036 for fewer regressors and 0.035 for more regressors). The diagram also implies that the direct matching (combined with OLS) achieves the lowest estimation bias (approximately 0.08 for both model assumptions) within all three methods.

2.3 Compare the indirect matching PSM (combined with OLS) with OLS (without matching step) in Monte Carlo Simulation study

Possibly, the indirect matching scheme-PSM is the most popular strategy and commonly used in causal inference study (Pearl, 2010). The method has been used and referenced for around 116,000 research papers[5]. We reanalyse the examples by using

[4] We only report kernel matching of PSM in the table, while other matching scheme provides quite similar results and thus are not reported.

[5] We count from Google Scholar on Aug 12, 2018.

PSM method, an indirect matching scheme, because it is widely used in observational studies to deal with the problem of self-selection (non-randomness problem). According to Ho et al. (2007), Sekhon (2009), and Hainmueller (2012), PSM methods are commonly used for estimation of binary treatment effects based on the assumption of the selection of observables. One of the main criticisms of the PSM method is that, in practice, it can be challenging to ensure that the distributions of the propensity scores of the control and treatment groups are balanced (see for example Rosenbaum and Rubin, (1983, 1985); Dehejia and Wahba, (2002) for tests to check for balance). We investigate the balanced condition required by the PSM and extend the analysis in our studies to show that PSM may not produce both local balanced condition (if each individual covariate is the same between compared groups) and global balanced condition (if all the covariates are balanced jointly between treatment and control groups) if the analysed data quality is low (the natural experiment is not random). We show here in the Monte Carlo Simulation study that PSM combined with OLS (called PSM combined method thereafter) may perform worse than OLS (without matching step) if the dataset suffers from serious imbalance problem in non-random dataset. In fact, it is difficult for PSM to achieve balanced condition if the dataset is too far from random, in doing so, PSM may not produce well balanced dataset to be analysed.

We trace this problem in the Monte-Caro Simulation study. We create one random dataset, with two covariates for both control group and treatment group (both groups with 10,000 observations) from a uniform distribution (0,10). We also create one non-random dataset, with two covariates for the control group (with 10,000 observations) from a uniform distribution (0,10) and two covariates for the treatment group (with 10,000 observations) from a uniform distribution (8,18). All the other settings are following the same procedure as in the previous section.

We report the frequency distributions of the propensity scores for the two datasets in figure 1. The top panel of the figure shows that PSM has achieved perfect balanced condition in random dataset, while the difficulty for the non-random dataset (lower panel) is that most propensity scores in treatment group are much higher than that of control groups, in doing so, the balanced condition required by PSM is difficult to be met.

We show here the estimations from PSM combined method will be different from that estimated by OLS (without balancing checking step). The results, in the Panel A of table 5 show that PSM combined method initially does quite well regarding its estimation performance when the balanced condition is met in random dataset since its estimated parameter (1.981) is the same as OLS (1.981). However, as the balanced condition required by PSM deteriorates (as shown in bottom Panel of Figure 1), the PSM estimation performance (Panel B of table 5) suffers compared with OLS, and the potential MSE with PSM goes up dramatically compared with a small MSE of OLS (0.001).

In making the decision to adopt PSM in natural experiment design the scholar should keep in mind that the imbalanced propensity scores between treatment and control group should be a concern, and thus estimations without balanced assumption checking process will generally produce improper results.

2.4 The direct matching scheme-Entropy Balancing Matching Methodology

In order to improve the balanced condition between treatment and control groups required by natural experiment, we employ the direct matching scheme -Entropy Balancing (EB) method (Hainmueller (2012)). In the natural experiment design study, we consider a sample including n_1 observations in the treatment group and n_2

observations in the control group. Each observation i is subject to a treatment $D_i \in \{0,1\}$, where $D_i = 1$ means that observation i is in the treatment group, and $D_i = 0$ means that unit i is coded in the control group. X represents a set of covariates. X_i^k is the i'th observation for covariate *k*.

To obtain the weights w_i for each observation in the control group such that the treatment and control groups are balanced with respect to the covariates, we minimize the entropy equation as follows:[6]

$$\mathrm{H}(w^i) = \sum_{\{i|D\ =\ 0\}} w^i \, log \frac{w_i}{q_i} \tag{1}$$

which is an entropy distance function from Kullback (1959), where

$$q_i = \frac{1}{n_2} \tag{2}$$

and "n_2" is the number of units in the control group. The loss function can also be chosen from Cressie Read divergence family (Read and Cressie (1988)). The reason why Kullback function is preferred is as its property is more stable, given the misspecification environment ((Imbens, Spady, and Johnson (1998)).

The entropy distance function in equation (1) is minimized subject to the following three constraints that ensure balance:

$$\sum_{\{i|D\ =\ 0\}} w_i \, c_r^k(X_i^k) = m_r^k, r \in 1,2, \ldots \mathrm{R} \tag{3}$$

$$\sum_{\{i|D\ =\ 0\}} w_i = 1 \tag{4}$$

$$w_i \geq 0 \ \forall_i \ \text{ such that D=0} \tag{5}$$

where the balance constraints $C_r^k(X_i^k) = m_r^k$ are imposed through a reweighting of the

[6] As the entropy equation is a strict convex function, the local optimizing solution is also the global optimizing solution, which can guarantee a unique result on weights. Also, according to Kullback (1959), as the EB function is a non-negative function, it decreases if w_i is closer to q_i. For example, if w_i equals q_i, there is no loss of original information on weights and the function can achieve its minimum value 0.

r moments for X_i^k of the control group.[7] In the natural experiment studies, we include all the three moments (mean, variance, and skewness) as the constraints in the analysis if there is a solution on weights. Here, m_r^k is the moment's "*r*" for covariate k in the treatment group. C_r^k is the moment's function for covariate *k* in the control group.

2.5 Combing the direct matching scheme-EB with the natural experiment design

The typical estimation strategy in direct matching scheme is to calculate the mean differences between the treatment and control groups as the average treatment effect, without worrying about the true economic question where the economic structural models can provide the useful information of interest. Heckman and Urzua (2009) argue that even the perfect randomization experiment cannot answer the question of economic interest because of the lack of the economic mechanism analysis executed by structure models. In order to explore the benefits of both direct matching scheme and the economic structure models used in the natural experiment design, we try to explore ways to combine the direct matching method (EB) with the natural experiment method (a structure model is usually included within the natural experiment design) and to provide evidence that if this combined approach (called direct matching combined method thereafter) is appropriate in the non-randomness "natural experiment" study.

More specifically, we firstly produce the solution obtained from the equation (1) to equation (5),[8] then the resulting weights is compatible with OLS, a commonly used approach in the natural experiment design.

[7] We can include all three moments' (mean, variance, and skewness) constraints, such that the mean, variance, and skewness in the control group are equal to those in the treatment group.

[8] There are cases where the solution on weights cannot be obtained. The reason is that if there are few overlaps between treatment and control groups with limited observations, EB cannot provide a solution on weights. For example, the ages in the treatment group ranging from 5 to 10, whereas the ages in the control group ranging from 11 to 20.

Figure 2 reports an example of density function for a covariate with and without the weighting operation process. The left column reports the raw density function without a weighting operation. The right column reports the covariate density function after a weighting operation. The red curve represents the density function of the covariate in the control group. The blue curve represents the density function of the covariate in the treatment group. The figure shows that after the weighting operation, the covariates are rather similar between treatment and control groups. In doing so, the randomness assumption required by natural experiment is improved.

Note that the above calculation process can be easily implemented in statistical software. For example, the prefix command "*svy*" in Stata Software can be used for setting weights in the reweighted sample to calculate the mean difference of the outcomes of interest, or a weighted least square regression can easily be accomplished if applicable.

3. The effects of Board Independence on CEO compensation

Boards of directors are a key aspect of corporate governance design because they are charged with monitoring executives' activities to ensure that managers act in the owners' interests (Jensen and Meckling, 1976). In this dispersed ownership environment, the board of directors, with its key monitoring function, focuses on reducing conflicts between a diversified ownership and strong executives. Therefore, independent boards of directors by definition[9] have incentives to monitor managers (Fama and Jensen, 1983).

[9] In this paper, we follow the RiskMetrics definition of an *independent director* as a director who has no material connection to the company other than a board seat. See more details at

http://wrdsweb.wharton.upenn.edu/wrds/support/Data/_001Manuals%20and%20Overviews/_115RiskMetrics/RiskMetrics%20Directors%20Definitions.cfm.

Considering the board's potential role in corporate governance design, board independence levels and their influence on corporate governance and CEO compensation are worthy of evaluation. Convincing empirical studies that examine the board independence effect have been rare because both theoretical and empirical studies argue that certain board structures and board independence levels are endogenous (Adams, Hermalin, and Weisbach (2010)). The literature is developing to address this problem. Recent empirical studies (Chhaochharia and Grinstein (2009); Duchin, Matsusaka, and Ozbas (2010); Chen (2014); Chen, Cheng, and Wang (2015); Armstrong, Core, and Guay (2014); Guthrie, Sokolowsky, and Wan (2012))[10] employ the "natural experiment" induced by the SEC regulation to study the effects of board independence. Atanasov and Black (2016) review 36 shock-based natural experiments on causal inference in corporate finance literature, 17 of which are driven by SOX regulation.

Although these studies (i.e., Chhaochharia and Grinstein (2009), Guthrie, Sokolowsky, and Wan (2012), Duchin, Matsusaka, and Ozbas (2010)) analyse the effects of board independence on CEO compensation or the effect of audit independence on firm performance, few studies discuss whether the dataset follows randomness assumption required by natural experiment. Based on the post-SOX exchange regulations that force firms to increase their board independence level if they

[10] There is a debate over the effects of board independence. For example, Chhaochharia and Grinstein (2009) find that increased board independence will decrease CEO compensation, but Guthrie, Sokolowsky, and Wan (2012) argue that this decrease is caused by outliers and that in reality, compensation committee independence will increase CEO pay. But none of the two studies provide an efficient method to address the randomness assumption of natural experiment problem rather than relying on researchers' own knowledge to one or more special events (i.e., the compensation for apple's CEO Steve Jobs) which provides us incentive to use EB method to detect the problem more efficient.

do not satisfy the requirements,[11] we provide new evidence on whether board independence affects the CEO's compensation.

Despite the strength of this design, that is, that firms which did not satisfy this requirement were serendipitously forced to increase their board independence level (thus, the variation in board structure of affected firms appears "exogenous"), this "natural experiment" lacks balance checking on the characteristics between compared groups in the study.[12] Indeed, part of the predicament in this design is that the assignment of a treatment variable to subjects is not random, since we show that the firm characteristics are significantly different among compared groups.

Possibly, the challenge in producing balanced and comparable observations explains why most literatures using the SOX as a "natural experiment" do not check and report imbalance problems and run regressions in the first place. Even the table of statistics compares means between firms which are forced to increase board independence and those are not required to, the imbalance problem caused by variance, skewness and joint imbalanced condition are seldom discussed. Proceeding without balance checking will lead to the concern that the criterion of randomness in the "SOX" experiment is not satisfied. Thus, the researcher is left to worry that the "causality" of

[11] In response to the significant accounting and corporate scandals that occurred between 2000 and 2002 and the SOX, the Securities Exchange Commission (SEC) approved the new exchange listing proposals from NASDAQ and the New York Stock Exchange (NYSE). The major provisions of those proposals that are relevant to this paper are the following: (a) all publicly listed firms should have a majority of independent board directors; (b) compensation committees should be composed of independent directors; (c) nominating committees should be composed of independent directors; (d) audit committees should be composed of independent directors; and (e) in addition to its regular sessions, the board should hold sessions without management. More details can be found on the SEC website in Release No. 34-48745 (http://www.sec.gov/rules/sro/34-48745.htm), in Chhaochharia and Grinstein (2009) and in Guo and Masulis (2015).

[12] In a natural experiment, the assignment of treatment to subjects happens serendipitously and randomly (Freedman, Pisani, and Purves 1997, 4-8). Because the mechanism of treatment assignment is ambiguous and not controlled by the researcher, most natural experiment studies in economic and other social science disciplines are criticized based on the problem that the assumption of random assignment to subjects is not credible (Diamond and Robinson (2010); Angrist and Krueger (1991); Dunning (2008, 2010); Gregory,McNulty and Krasno (2009); Sekhon (2009); Sekhon and Titiunik (2012); Dunning, (2012); Rozenzweig and Wolpin, (2000)).

board independence on CEO compensation established from the regression method (without a balance-checking process) may not exist and actually stem from heterogeneous characteristics in compared samples.

To improve causal inference caused by the problem of unbalanced covariates between treatment and control groups, we combine the direct matching scheme with the NYSE and NASDAQ experiment as discussed above.

3.1 Data and Sample Construction

To implement the study, we define the firm is in the control group if this firm has satisfied with the SEC independence requirement before the SEC issued the new regulations by the end of 2002.[13] Otherwise the firm is in treatment group. For example, if a firm did not satisfy the SEC independence requirements in 2002, it had to increase the independence level to satisfy the new regulations from 2003 and will be coded in the treatment group.

Figure 3 reports the changes in the total board independence ratio from 2003 to 2005 of sample firms we collect from RiskMetrics. The board information including independence classification is obtained from Institutional Shareholder Services (formerly RiskMetrics). The compliance ratio shows the percentage of the sample firms which have more than 50% independent directors in the board. We find that the number of firms that began to adopt the new policy decreased gradually from 2003 to 2005.The trends reflected in the figure show that the firms began to comply with the new regulations in 2003 and most of the firms satisfied the new regulation by 2004, which is

[13] According to the new listing rules, firms had to comply with the new independence requirements from 2003 to 2005. Therefore, the observations in 2002 are the latest clean sample that is not affected by the new regulations.

consistent with both the SEC regulation on NASD and the NYSE Rules relating to corporate governance.[14]

We obtained the director-level data from Institutional Shareholder Services (formerly called RiskMetrics) for 2002 to 2005.[15] The data contained director information on classification (e.g., independent information). The information on CEO tenure, and the Standard Industrial Classification (SIC) four-digit code was obtained from ExecuComp. The data on firm accounting information were collected from Compustat, and the information on stock retains was obtained from the Center for Research in Security Prices (CRSP).

To ensure that the observations in the control and treatment groups are suitable for reflecting the SEC regulations, we implement the following requirements. First, the observation needs the classification information so that the independence status could be observed. Second, the observation needs the exchange information, and the firm has to be listed on the NYSE or NASDAQ to be consistent with the proposals required by the two exchange markets. Third, firms from 2002 to 2005 could not have violated the regulations if they had already previously complied with those regulations.[16] Fourth, all of the firms have to have complied with the regulations by the end of 2005 so that the observations would be consistent with the SEC regulations and could be applied to this study. Finally, the definition of board independence in RiskMetrics needs to be

[14] See Securities and Exchange Commission release No.34-48745 from http://www.sec.gov/rules/sro/34-48745.htm.

[15] This is because firms should have complied with the new requirements by the end of 2005.

[16] Some exceptional firms in 2002 had already complied with the independence requirements, but in later years, for example, 2005, they violated the regulations and did not comply with the new regulations. The potential interpretation may be that they were exempt from the SEC regulations or they did not care about the violation of SEC regulations because it was very rare for firms to be removed from the exchange even if they violated this SEC requirement. These cases are very rare and are not relevant to this study. More details on the reasons that firms do not follow the requirements can be found in Guthrie, Sokolowsky, and Wan (2012) and Dah, Frye, and Hurst (2014).

consistent with the definition in the NYSE and NASDAQ regulations. To adjust the definition, we follow Chhaochharia and Grinstein's (2009) specification that board directors who were former employees but who had left the company more than 3 years previously were independent directors.[17]

3.2 Design

We follow Chhaochharia and Grinstein (2009) who use a NYSE- and NASDAQ-proposed regulation to examine the effect of board independence. We explore the fact that the NYSE and NASDAQ exchanges proposed a regulation in 2002 that all listed firms should have boards of directors that were greater than 50% independent. Consequently, firms that did not satisfy this requirement were serendipitously forced to increase their board independence level (thus, the variation in board structure of affected firms appears exogenous), and they were required to satisfy the new regulation before the year 2005 (inclusive). This experiment compares firms that were forced to increase their board independence level with firms that had already satisfied the regulation in 2002 and thus remain unchanged in board structure. This experiment design is an effort to overcome the challenge of studying board structures. Such examinations constitute a challenge because board structures are endogenous, a statement based on both theoretical grounds and empirical evidence in the literature (Hermalin and Weisbach (1998, 2003)).

We follow the designs presented in Chhaochharia and Grinstein (2009); the regression model is recorded as follows:

[17] The definition of independence in RiskMetrics can be obtained from http://wrds-web.wharton.upenn.edu/wrds/support/Data/_001Manuals%20and%20Overviews/_115RiskMetrics/RiskMetrics%20Directors%20Definitions.cfm. The definition of independence according to SEC regulations can be obtained from http://www.sec.gov/rules/sro/34-48745.htm. The final sample is consistent with previous literature, for more details on data construction, see Chhaochharia and Grinstein (2009).

$$\text{Log}(Compensation_{it}) = \alpha_0 + \alpha_1 \times Dummy(NoncompliantBoard02)_i \times Dummy(03-05)_t + \text{Controls}_{it} + Firm\_Effects_i + Industry\_year\_Effects_{it} + \varepsilon_{it} \quad (6)$$

where the dependent variable is defined as the natural log of CEO compensation[18], Dummy (Noncompliant Board02) is coded as 1 if the firm did not comply with the regulation in the year 2002, and as 0 otherwise. Dummy (03-05) is set to 1 if the observation is in the period 2003–2005, and 0 otherwise. Dummy (00-02) equals 1 if the observation is in the period 2000–2002, and 0 otherwise. The controls include (1) Sales×Dummy (00-02), (2) Sales×Dummy (03-05), (3) ROA×Dummy (00-02), (4) ROA×Dummy (03-05), (5) Stock Returns×Dummy (00-02), (6) Stock Returns×Dummy (03-05), (7) CEO Tenure,[19] (8) Firm fixed effect, and (9) Industry-year fixed effect, which is defined as Fama–French 48 industry factor times year dummy variable.

The sample includes 807 firms. The data are collected according to the instructions of Chhaochharia and Grinstein (2009). The explanations for these can be obtained from the Supplementary Appendix from the Journal of Finance website.

We assume that there is an increase in the board independence level after the regulation is truly exogenous and that such rules randomly affect firms listed on the NYSE or NASDAQ exchanges. We consider that the covariates in the groups compared should be similar if the natural experiment satisfies the randomness assumption. We report the comparison of the covariates for the two groups in Table 6.

[18] Since the treatment effect estimated by the difference in difference approach relies on the OLS regression of Equation (6), all the assumption required by OLS is still required by the difference in difference approach here. Hence, the estimation problem (raised by independence assumption we discussed in chapter 2) encountered by OLS is also applicable to the difference in difference approach here.

[19] The definitions of the covariates are as follows: *Sales* is the natural log of the company sales. *ROA* is the natural log of one plus the net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1). *Stock Returns* is the natural log of the annual gross stock return (dividend reinvested), measured in year (t−1). *Tenure* is the number of years the CEO served in the firm.

Based on the table 6, we consider the following question: is the comparison of two groups guaranteed to be a valid natural experiment in terms of the randomness hypothesis? It can be easily inferred that the answer is no. The table shows that most firm characteristics (i.e., Sales, ROA, and CEO tenure) differ significantly between the groups compared.[20] For example, the mean of the variable "ROA×Dummy (00-02)" for the affected group is 0.031, which is 0.007 greater than that in the non-affected group at the 1% significance level.

As the treatment and control groups are different in terms of their characteristics, the distinct pre-conditions in each group will probably lead to the randomness assumption required by the natural experiment unconvincing. Thus, the observed estimates in terms of the differences in CEO compensation between the two groups cannot simply be attributable to the fact that the affected firms must increase the board independence level. Rather, the distinct pre-conditions do not make the two groups comparable; by doing so, the "causal inference" based on the simple OLS may cause bias estimation and imprecise standard errors.

This problem naturally leads to the following question: how can the unbalanced covariates between the two groups be addressed to improve causal inference? As the discussion in the methodology section provides clarity, we can combine the direct matching method with the natural experiment design in the original study

Following the calculation process shown in the methodology chapter, we produce the weights for each observation in the control group, such that the covariates can be sufficiently balanced between the treatment and control groups with respect to covariates' distributions. The density for the weights' results is reported in Figure 4. The

[20] Table 1 reports the overall covariate comparison for all the firms for the period 2000–2005. In an unreported test, we also find that the characteristics are distinct between the treatment and control groups in the year 2002 when the regulation was imposed.

figure shows that much of the weights are modified from the original weights (the original weights should be the same for all the observations). The figure implies that the EB method has to change the original weights significantly, so that the balanced condition can be satisfied, conditional on the information loss (entropy should be minimized). Although much of the original information is modified by the method, there are no extreme weights produced. In doing so, our analysis will not be sensitive to the observations with big weights assigned to them.[21]

The new summary statistics after direct matching calculation are shown in Table 7. The result shows that the covariates are balanced as the first, second, and third moments are equivalent between the groups compared. For example, the mean, variance, and skewness of CEO tenure are 2.261, 0.764, and -0.084 respectively, for both the groups, which indicate that the assignment to treatment is orthogonal to the firm's background characteristics (with respect to "CEO tenure").

3.3 Main Results for the effect of board independence on firm's CEO compensation

To show the effects on CEO compensation, we initially use the naturel experiment with OLS method (without a data-preprocessing step) as in Chhaochharia and Grinstein (2009). The result is reported in the left column of Table 8. The table suggests that an increased board independence level can "cause" a decrease in CEO compensation (the coefficient is -0.171 at the 10% significance level).[22] The evidence suggests that an

[21] We use the program "ebalance" in stata to generate the weights. The program scaled the calculated weights for the control group, such that the sum of the total weights in the control group equals the sum of the weights in the treatment group (the number of observations in the treatment group). In this sense, no individual observation has been assigned too much weight.

[22] We use the same data and program from the Journal of Finance Supplementary Appendix website provided by Chhaochharia and Grinstein (2009). The data construction method in that the Appendix is marginally modified in comparison to the one used in the original paper; hence, the results are

increased board independence level appears to “cause” a decrease in CEO compensation, which is a major argument in Chhaochharia and Grinstein (2009).

To compare the estimated results of our proposed framework with those in Chhaochharia and Grinstein (2009)’s study, our second experiment is based on the direct matching combined method as explained in the method chapter. The weights are combined with the regression method used in equation (6). The result is reported in the right column of Table 8. The estimates in the diagram display that the effect of board independence on CEO total compensation is much lower compared with the original study (the coefficient is only -0.129 and not significant at the 10% significance level). The estimated results are significantly different between the two methods, except for the situation in which the results are all negative.

Overall, our results suggest that our combined method improves the balanced condition required by the natural experiment study, and our causal inference between board independence and CEO compensation is quite different from that in the original study.

3.4 Simulation studies

As the effects of board independence on CEO compensation estimated by the direct matching combined method and OLS method (used in the original natural experiment study) are rather different, this section implements bootstrap simulation tests to evaluate the two methods’ performance with respect to a variety of benchmark standards.[23] More

marginally different for the modified dataset. Nevertheless, our replicated results are qualitatively similar to those in the original study in Chhaochharia and Grinstein (2009).

[23] For studies implementing simulation to evaluate the properties and performance of estimating methods, see Abadie and Imbens (2007), Diamond and Sekhon (2013), Rubin (2006), Zhao (2004), and Hainmueller (2012).

specifically, we follow Ho et al. (2007), Colin and Trivedi (2009) and examine the sensitivity of the results estimated by both methods.

According to Colin and Trivedi (2009), bootstrapping allows researchers to make statistical inferences by using a resampling technique from a data sample. The statistics evaluated in this section include bias and MSE.

Consider the following example: the mean of the variable θ (the "causal effect" of board independence on CEO compensation in this paper) is generally challenging to obtain by using one dataset as there is only one estimator that can be obtained from the data. Moreover, a reliable distribution for the mean of the estimator (we generally assume that it follows a normal distribution) cannot be obtained with one dataset. To solve this problem, we can draw many (i.e., 1000) random samples with replacement (with the same number of observations) from the data, and we can obtain 1000 different estimates of θ for each method within each bootstrap simulation. Let $\theta_1, \theta_2, \dots \theta_n$denote each estimate for the 1000 random samples for each method. The mean of θ can be calculated as follows:

$$\bar{\theta} = \frac{1}{n}\sum_{b=1}^{n} \theta_b \tag{7}$$

Here, n=1000; further, we can calculate the bootstrap estimate of the variance for θ

$$Var_{Bootstrap}(\theta) = \frac{1}{n-1}\sum_{b=1}^{n}(\theta_b - \bar{\theta})^2 \tag{8}$$

Moreover, the standard deviation of θ can be calculated by

$$SE_{Bootstrap}(\theta) = \sqrt{Var_{Bootstrap}(\theta)} \tag{9}$$

Further, the bias can be defined as follows:

$$Bias_{Bootstrap}(\theta) = \bar{\theta} - \theta_0 \tag{10}$$

where θ_0 is the treatment effect estimated by each method from the original dataset, and the MSE can be calculated as follows:

$$MSE_{Bootstrap}(\theta^2) = Bias^2_{Bootstrap}(\theta) + Var_{Bootstrap}(\theta) \tag{11}$$

Following the above steps, we can derive such test statistics as bias and MSE. This leads to the results which are reported in Table 9. This bootstrap simulation study displays the implementation of the discussed direct matching combined method and OLS method. The model estimation with Chhaochharia and Grinstein (2009)'s dataset shows that the coefficients based on direct matching combined method generally provide a smaller effect (0.129) in comparison to OLS (0.171). The direct matching combined method improves the balanced condition between the treatment and control groups. Thus, unsurprisingly, direct matching combined method performs better than the conventional OLS technique when the underlying statistical properties are analyzed: (1) The bias for the direct matching combined method is smaller (0.031) in comparison to the bias of 0.048 of the OLS method. (2) The MSE for the direct matching combined method is also smaller (0.019), indicating that the average of the squares of the "errors" is smaller (in comparison to the MSE of 0.221 of the OLS method). The simulation results show that neglecting any imbalance problem on the covariates will saliently decrease the estimation quality of the data analysis.

3.5 Comparison between the direct matching combined method and the PSM combined method.

According to Ho et al. (2007), Sekhon (2009), and Hainmueller (2012), to deal with the self-selection problem, the PSM methods are currently widely used in observational studies for estimation of binary treatment effects based on the assumption of selection of the observables. One of the main criticisms of the PSM method is that in practice, it can be challenging to ensure that the distributions of the propensity scores between the treatment and control groups are balanced (see for example Rosenbaum and Rubin, (1983); Dehejia and Wahba, (2002)). Austin (2008) reviewed 47 studies in leading journals in medical science employing PSM and found that only two of them report the process to check the balanced condition. Diamond and Sekhon (2013) reviewed the articles published from 2000 to 2010 in the American Economic Review, Journal of Political Economy, Quarterly Journal of Economics, and The Review of Economics and Statistics on PSM. Four articles reported the mean tests for some variables in balance checking and only one article reported the tests for balance checking for all variables used in PSM.

To evaluate whether PSM combined method works better than the direct matching combined method, a comparison study between the 2 methods is conducted.[24] We follow the initial work of Rosenbaum and Rubin (1983) and Heckman, Ichimura, and Todd (1998).[25] A variety of methods exist to match the observations in the treatment group

with comparable observations in the control group. The first PSM method used is kernel matching.[26] This method is discussed in Heckman, Ichimura, and Todd (1998) and Becker and Ichino (2002).

[24] We only focus on the empirical comparison between EB and PSM in our examples. For a systematic comparison between the two methods including theory or the example, see Hainmueller (2012), and Diamond and Sekhon (2013)

[25] The PSM procedure used in this paper is discussed in the Appendix. For more on the econometrics of the PSM method, see Becker and Ichino (2002). For applications to PSM, see Vega and Winkelried (2005), de Mendonça de Guimarães e Souza (2012), Lin and Ye (2007, 2009), Chen (2016), and Samarina, Terpstra and Haan (2014).

[26] As kernel matching also uses the re-weighting technique (which is different from other matching methods, for example discarding observations in the nearest neighbor matching), this method is the most appropriate PSM method to be compared with EB (as a reweighting technique). We also compare EB with other PSM matching schemes in the robustness tests section.

3.5.1 Comparison between the direct matching combined method and PSM combined methods

3.5.1.1 Estimation of Propensity Scores

To estimate the propensity score, we use the same variables as those in the OLS method in the original natural experiment study. The logit regression result used to estimate the propensity score is reported in Table 10.[27] Most coefficients in the regression study are rather similar and significant. For example, *Sales* can negatively determine the possibility of treatment adoption (-0.188 at the 5% significant level). In contrast, some control variables (e.g., *ROA*) can positively determine the probability of treatment dummy (with a coefficient of 2.912 at the 10% significance level). Given this understanding that firm characteristics such as sales, ROA, and CEO tenure will determine the probability of whether the firm has already satisfied the regulation in 2002 (and thus whether the firm is subject to regulation by the new policy), along with the understanding that the preconditions in the compared groups can affect the results of CEO compensation rather than the increased board independence level, it provides clarity that the NYSE-NASDAQ strategic design could cast doubt on the randomness assumption for assignment of treatment. This problem is neglected in the previous literature (i.e., in Chhaochharia and Grinstein (2009)'s study).

3.5.1.2 Balanced condition in Propensity Score Matching

To examine the balanced condition in the PSM method, distribution diagrams for propensity scores are reported in Figure 5. In order to compare the PSM study, which satisfies the balanced condition, we illustrate the well-balanced condition example produced by Kirmani, Holmes, and Muir (2016) in Figure 6. In contrast to Figure 6 where

[27] We only use the year 2002 to obtain propensity scores as we must find comparable groups in 2002 before the new regulation has just been announced.

the distributions of propensity scores in the control group (bottom section) is similar to that in the treatment group (upper section), Figure 5 shows that the balanced condition is not well satisfied as the frequency of the propensity scores in the treatment group cannot match those in the control group. For example, most of the propensity scores in the treatment group range from 0.1 to 0.5, whereas most propensity scores in the control group range from 0 to 0.3. In addition, we cannot find matched observations in the control group for the treatment group with p-scores between 0.5 and 0.6. Thus, the commonly balanced condition assumed by PSM may not be satisfied and subsequent analysis conducted by this method may not be valid

3.5.1.3 Comparison between the direct matching compared method with the PSM combined method

Following Heckman, Ichimura, and Todd (1998) and Becker and Ichino (2002), new weights can be obtained from the PSM method, and the weights are combined with the regression method. The results are reported in the first row of Table 11. The diagram shows that PSM can produce higher and more significant (-0.193 at the 10% significance level) result than that estimated by EB (-0.129 not significant even at the 10% level). To evaluate the effectiveness for each method, we follow the same procedure from equation (7) to (11), and obtain the bias and MSE for each method.

The results in Table 11 show that the bias in the direct matching combined method is 0.004 in comparison to 0.034 in the PSM combined method. In addition, as shown in the table, the MSE in direct matching combined method is decreased to 0.009 after the process, controlling the information loss problem in terms of entropy measure, which is smaller than that in the PSM combined method (0.041). This finding suggests that the direct matching combined method outperforms the PSM combined technique in finite samples, which can be obtained from assigning new weights according to sample moments.

For robustness test, we capture whether our conclusion may not hold if a different PSM method is used. We use various matching schemes such as local linear regression matching, radius matching, and nearest matching, and the results are provided in Table 12.

The table illustrates the results on estimation performance from 1000 simulations, where the model design and variables' definitions are the same as before. In the model, direct matching combined method provides the lowest bias (0.031) in contrast to local linear regression matching combined method (0.036), radius matching combined method (0.048, 0.035, and 0.036) for the radius equals 0.01, 0.02, and 0.03, respectively. The estimation of MSE also squints toward direct matching combined method as its MSE is the lowest (0.02) in comparison to all the other PSMs combined methods.

Considering all the benchmarks, the "causal" effects estimated by the direct matching combined method are significantly more reliable as (1) the bias for the method is smaller in all the cases, and (2) the MSE for the method is minimum in contrast to that of the PSM combined method, which indicates that the average of the squares of the "errors" is very small. Therefore, we infer that in the context of the NYSE-NASDAQ regulation on board independence, a natural experiment design that compares the impacted and non-impacted firms to study the impact of board independence on CEO compensations with PSM combined method, is not appropriate.

3.6 Further robustness

In this subsection, we evaluate the methods on matching quality, we compute both standardized percentage bias on each covariate (locally matching) and the Rubins' B index ratio and Hotelling's T-squared generalized means test on covariates jointly (globally matching). We do this for the direct matching combined method, the PSM combined method, and then the raw data without matching scheme in original study as a comparison.

3.6.1. Further examination

We follow Rosenbaum and Rubin (1985) and evaluate the standardized percentage bias on the sample covariates before and after direct (EB) and indirect (PSM) matching. The results, that appear in Figures 7 and 8, show that the standardized percentage bias on each covariate drops for both direct matching (EB) and indirect matching (PSM) in comparison to the raw data without any matching scheme. The bias after direct matching (EB) calculation reduces to exactly zero on each covariate, which is superior than indirect matching (PSM) as some covariates are still imbalanced between the groups compared after PSM weighting.

We then examine the overall imbalance problem with Hotelling's T-squared generalized means test and Rubin's B test. Hotelling's T-squared generalized means test evaluates whether a bunch of covariates' means are equal between treatment and control groups, the null hypothesis is vectors of means are equal for the two groups. Rubin's B indicates the standardized difference of the means of the linear index of the propensity score in the treatment and control groups. The higher B ratio it is, the more imbalance problem exists between the groups compared. Rubin (2001) demonstrates that B is less than 25 for the sample compared, indicates sufficiently balanced groups. The results are reported in Table 13. Using Hotelling's T squared test, we continue to find a significant imbalance problem for the raw dataset between the treatment and control groups, as the F statistic equals 33.626 and we have almost 100 percent confidence to reject the null hypothesis that the set of covariates' means are the same. By employing indirect matching (PSM) and the indirect matching (EB) method, the values of the F statistic reduce to 0.371 and 0, respectively. The overall imbalance of the raw data, calculated by the B ratio, equals to 57.5, which implies that the groups compared are considered as sufficiently imbalanced. In indirect matching (PSM) and direct matching (EB), the B imbalance is dramatically decreased. However, the direct matching (EB) is always better than the indirect matching (PSM) as the difference in the B imbalance completely reduces to zero in contrast to PSM (with imbalance B ratio of 4.6)

4. Additional Applications: Vouchers program for Private Schooling in Colombia, NSW program and JTPA program

In this section, we offer several "natural experiment" examples that shows that if a dataset has a sound, balanced condition (but is still not perfect) in compared groups, the direct matching combined method will still be superior to the OLS (without matching step) or the PSM combined method. More specifically, we employ the direct matching combined method to reanalyse the datasets from the influential studies of Angrist et al (2002) on Vouchers program, Dehejia and Wahba (2002) on National Supported Work Demonstration (NSW) training program, and Heckman et at. (1997) on National Job Training Partnership Act (JTPA).

4.1 PACES lottery program

We follow Angrist et al (2002)'s study, who interviewed approximately 1,600 PACES applicants. They tried to ensure equal numbers of lottery winners and losers. They interviewed the applicants mostly by telephone and recorded their characteristics accordingly.

We follow the designs presented in Angrist et al (2002)'s study and regress y_i (the number of hours the applicant is working every week) on the treatment indicator Z_i(a dummy variable indicating whether the applicant won the lottery) and several control variables X_i (i.e., individual or survey characteristics such as sex and age).

$$y_i = X_i'\beta_0 + \alpha_0 Z_i + \delta_i + \varepsilon_i \qquad (12)$$

Thus far, we assume that the voucher program is randomly assigned to each candidate, as assumed by Angrist et al (2002)'s natural experiment study. In contrast to Angrist et al (2002)'s study who use the raw dataset to analyse the problem directly, we test whether the randomness assumption is reasonable.

To establish whether the applicants who won the lottery are similar to the applicants who lost, we illustrate the comparison of the characteristics between the two groups in Table 14. We illustrate in the diagram that the randomness assumption is better achieved compared to previous chapters (despite very few characteristics that are still significantly different between the compared groups).

If a few characteristics in the winners' group tend to be different from those in the control group, comparing the groups that are affected by the vouchers program with those who are not affected by the plan will still have negative impact on the natural experiment's estimation process because the two groups are still heterogeneous. Thus, the effect of the vouchers program on the number of hours winners are working per week estimated by the OLS (without a data preprocessing step in Angrist et al (2002)'s natural experiment study) might not be precise, and the direct matching combined method is still suggested.

4.2 Results

To determine whether the "vouchers program" matters, we first examine the treatment effects on the number of hours that winners are working per week. We estimate the treatment effect by employing the direct matching combined method, where the dependent variable is number of hours the candidate is working per week. To ensure that the observations of the treatment and control groups are similar and comparable, we include a set of control variables discussed in the data section

4.2.1.1 Empirical results for effects of vouchers program on the number of hours winners are working per week

The direct matching combined method calculation is reported in Table 15. Panel A reports the summary statistics before the direct matching (EB) method. For example, the mean of the age is 14.78 for the winners' group, which is smaller than that for the losers' group (14.91; although this difference is not significant).

Following the direct matching calculation process shown in the method section, we produce the weight for each observation in the control group so that the balanced condition can be sufficiently satisfied between the treatment and control groups. The density for the weight results are reported in Figure 9.

The figure shows that most of weights assigned to control group observations are approximately 0.0013. This is the same as the original weights assigned to control group. The control group has 716 observations, and each observation has a weight of 1/716, which is 0.0013. This result

indicates that the original dataset is much better with respect to the balanced condition so that the weights do not change much.

The new summary statistics after the direct matching calculation are shown in Table 15, Panel B. The table shows that the covariates are balanced and that all of the three moments are equivalent between the treatment and control group. For example, the mean, variance and skewness of gender are 0.498, 0.25, and 0.01, respectively, for both the treatment and control groups, thus indicating that the balanced condition of “gender” has been achieved.

By using the calculation process presented in the method section, we can easily obtain the treatment effects with the direct matching combined method. The estimate of the average treatment effect of the vouchers program on the number of hours that a candidate is working is reported in Table 16. The second column reports the results estimated from the OLS method. The treatment effect is -0.87 percent, which is the same as documented in Angrist et al (2002)’s study. The result means that the winners generally have 0.0087 fewer working hours per week than those in the losers’ group at the 10% significance level. To compare Angrist et al (2002)’s study, we report the treatment effect results produced by the direct matching combined method in the third column. All the control variables used in the OLS are the same as those used for the direct matching combined, and they are explained in the method section. The results provide similar conclusions that the treatment effect is -0.917 (which is significant at the 10% significance level), which means that the winners generally have 0.917 hours less than those in the losers’ group. The results are slightly different between the two groups, which may be caused by slight changes in the weights that the direct matching imposes on the control group observations. This highlights the issue that if the experiment satisfied the sound balanced condition, the direct matching combined method may provide similar (but still slightly different) conclusions than those inferred by the OLS in the original natural experiment study.

4.2.2 Simulation studies

We now extend the bootstrap simulation approach to accommodate the data in Angrist et al (2002)’s study such that the statistical properties can test for the distinctions between our proposed combined method and the regression method used in Angrist et al (2002)'s study.

We begin by following the same simulation procedures from equations (7) to (11) so that we can compare the direct matching combined method with the OLS methods in terms of the bias and MSE. The results are reported in Table 17. The table reports the bootstrap simulation results for the effect of the vouchers program on the number of hours that the candidate is working per week. The results estimated by the direct matching combined method and the OLS method are different (-0.92 and -0.87 at the 10% significance level, respectively). The raw unadjusted results estimated by the OLS have a bias of 0.02 and an MSE of 0.28.

This simulation study illustrates that the OLS may still perform worse than the EB methods when the randomness assumption is not well satisfied because of a bigger bias and MSE. This feature implies that even when a better (but not perfect) dataset is employed, the balanced condition is much better achieved (but not perfectly achieved) than the examples in the previous two chapters where the data quality is lower, and the direct matching combined method still outperforms the OLS method without matching step in the original natural experiment study, but such improvement is very limited (i.e., the direct matching combined method can only reduce MSE from 0.28 to 0.27 compared with the OLS without the matching step)

4.2.3 Comparison of the direct matching combined method and PSM combined method

We follow the same strategy as in previous chapters and apply the PSM combined method to this test with the same model and dataset used in the direct matching combined method. In Figure 10, we present the balanced condition required by PSM. In general, the overall propensity scores for both treatment and control groups are similar (although not perfect especially when the propensity scores range from 0.6 to 0.8, since the treatment group observations cannot find enough reliable observations in the control group). In doing so, the subsequent estimation may still produce a low-quality estimation although such imperfection is limited since the overall propensity scores are quite similar between compared groups and not much reweighting activities are needed.

In Table 18, we present the comparison between the PSM combined method and the direct matching combined method of Angrist et al (2002)'s dataset with the same model assumptions as in the OLS section. The direct matching combined method is used to increase the balanced condition for all covariates. The results for the PSM combined method are quite similar to the results of the direct matching combined method (-0.917 at the 10% significance level in the direct matching combined method compared with -0.937 at the 10% significance level in the PSM), which indicates that a better quality (but still not perfect) dataset requires less weighting process in the PSM combined method and the results from the method are quite similar (but still inferior) compared with that estimated from the direct matching combined method.

To evaluate the two methods' performances, we follow the same bootstrap stimulation strategy and report the results in Table 18. The propensity score matching method follow Heckman, Ichimura, and Todd (1998) and Becker and Ichino (2002). All of the benchmarks (bias and MSE) to compare the two methods are based on 1000 simulations (with replacement). In each simulation, we use the direct matching combined method and the PSM combined method to estimate the treatment effect of the vouchers program on the number of hours that each candidate is working every week. All of the control variables are the same as in the direct matching combined method and OLS method sections. The definitions of the treatment and control groups and the estimation process are provided in the method section. The definitions of each variable are presented in the Appendix. Since the direct matching combined method finds the weights to satisfy the balanced condition well, it effectively reduces the bias and MSE to 0.001 and 0.295, respectively, compared with those of the PSM. In all cases, the direct matching combined method provides the better-balanced condition required by the natural experiment, and in all cases, the method’s estimates have the lowest bias and MSE compared with the OLS and PSM combined method.

4.2.3.1 Comparison of the direct matching combined method and various PSMs combined methods

For robustness checks, we test if our main conclusion will change if we use different PSM techniques such as local linear regression matching, radius matching, or nearest neighbour matching. We present the results in Table 19.

In each simulation, we use the direct matching combined method and the various PSMs combined methods to estimate the treatment effect of the vouchers program on the number of hours each candidate is working every week. All of the control variables are the same as in the previous method sections. The definitions of the treatment and control groups and the estimation process are provided in the method section. In the model, the direct matching combined method dominates all the PSMs combined methods because it can achieve the lowest bias (0.01) and the lowest MSE (0.278)

4.2.4 Further robustness

In this subsection, we evaluate the three methods in this “financial vouchers” experiment, by calculating standardized percentage bias, Rubins’ B index ratio, and Hotelling's T-squared generalized means test.

4.2.4.1. Further examination

We follow Rosenbaum and Rubin (1985) and evaluate the standardized percentage bias of the sample covariates (i.e., age and survey type) before and after direct matching

(EB) and indirect matching (PSM) for the students who received the voucher support and the students who did not. The results in figures 15 and 16 show that both the direct matching (EB) and the indirect matching (PSM) can help reduce the imbalance in each covariate to some extent. The direct matching is able to reduce the standardized percentage bias for all the covariates to 0, but the indirect matching (PSM) cannot reduce the percentage bias more than the direct matching and the imbalance problem still exits to some degree.

In a natural experiment study, the researcher may also prefer to evaluate the balanced condition jointly with respect to all the covariates. It will be better if Hotelling's T-squared generalized means test and Rubins' B test can be included in the analysis to ensure that the global imbalance problem will not exist, as we show in Table 20.

The Hotelling’s T-squared generalized means test is reported in the middle column of the table. The column indicates that the raw data are systematically worse than the data after the indirect matching (PSM) weighting and the direct matching scheme, since the F statistic equals 1.568. This implies that we have approximately 95% confidence to reject that the set of the covariates’ means are equal between the compared groups. As would be expected from the previous examples, the direct matching effectively reduces the F statistic to 0, which is superior to the indirect matching PSM (where the F statistic equals 0.165). The table also presents the results that the direct matching method is considerably better than the PSM as measured by Rubin’s B, which is consistent with the standardized percentage bias tests.

Thus, given the data drawn from the “financial vouchers” experiment, the optimal approach is to select one matching scheme based on the standardized percentage bias tests, Hotelling’s T-squared generalized means test, and Rubin’s B ratio. The conservative approach seems to select the direct matching combined method in our case since it performs the best in all the examples that we provided based on the statistic tests which we illustrated above.

4.3 Empirical application: National JTPA Study

As most of the natural experiment studies in Accounting, Economics and Finance disciplines have the model dependent problem, at least partly, this means that the estimation result depends on the assumption of the model specification (i.e., the assumption on what types of independent variables are included in the model) (Ho et al. 2007). We make use of the "National Job Training Partnership Act (JTPA)," an influential randomized experiment to evaluate the performance of a methodology (i.e., Heckman et at. (1997) among others) to show that if our estimation results will change much with our combined method. The JTPA is devolved by the US Department of Labor to evaluate the effect of the JTPA on program participants' earnings. The JTPA randomly choses participants according to their background characteristics who were qualified to apply to the program in 16 service delivery areas in the United States of America, such that part of the qualified candidates is selected into the treatment group randomly, whereas the others are automatically left to be in the control group.

The data set can be obtained from the Upjohn Institute. The background information is included in the sub-dataset Background Information Form (BIF). The BIF contains background information such as education, training history, demographic information, etc. The outcome of interest we use is the sum of 30 months after the assignment of the training.

The sample includes 11204 observations from 1983 to 1990. Of the observations, 6102 are women and 5102 are men. The control variables include an indicator of high school graduation, an indicator of black people, an indicator of Hispanic people, an indicator of Marital status, a dummy for the

In contrast to Heckman et at. (1997)'s study who focus on the self-selection bias issue and calculate the probability that if any observation will receive the JTPA program, we combine a direct matching method with economic structural model method and figure out that our combined method has better performance on reducing the model dependency problem for the estimation of treatment effect (the treatment effect is 1051 with less independent variables and is 1049 with more independent variables), in contrast to OLS without dealing with imbalanced condition step (the treatment effect is 1117 with less independent variables, and is 970 with more independent variables) as well as a combined of PSM and OLS method (the treatment effect is 1197 with less independent variables, and is 1059 with more independent variables)

In our analysis, we run the regression from 30-month earnings on the assignment indicator D (the selection into the JTPA services or not) and several control variables (whether the candidate is black or not, high school graduate or not, etc.).

$$Earnings_i = cons + D_i + Controls_i + \varepsilon_i \tag{13}$$

We consider that the covariates in the compared groups should be similar if the natural experiment satisfies the randomness assumption. We report the comparison of covariates for the two groups in Table 21. The diagram illustrates that, for both the men and women groups, almost all the characteristics between the compared groups are still different especially for the dummy of age classification in the "men" group.

The table also reports the Hotelling's T-squared generalized means for both men and women groups. The results illustrate that there is some imbalance problem caused by the joint covariates, since the F statistic is 0.8739 in the men's group and 1.0578 in the women's group.

4.3.1 Results

To investigate whether the direct matching combined method can help to reduce the model dependency problem, as suggested by (Ho et al. 2007), we test how the estimation results change according to the two different model assumptions on covariates (one with more covariates, the other one with less covariates) with the three methods (OLS, direct matching combined method and PSM combined method).

The density for the weights results calculated by the direct matching are reported in Figure 11 for the men's group and in Figure 12 for the women's group. The figures show that much of the weights are not modified too much.

4.3.1.1 Comparison direct matching combined method with OLS and PSM combined method with respect to model dependence

To evaluate the performance of the three methods on reducing model dependence, we test how the estimation results change according to the two different model assumptions on the covariates (one with more covariates, the other one with less independent variables) with the OLS, EB and PSM.[28]

The results are reported in Table 22. The control variables include an indicator of high school graduation, an indicator of black people, an indicator of Hispanic people, an indicator of marital status, a dummy variable for the age group, and a dummy variable for unemployment. "Less variables" in the table indicates that the model only includes the indicator of JTPA assignment. "More variables" in the table indicates that the model includes the indicator of JTPA and all the other control variables. The OLS, direct matching combined method and PSM combined method follow the same model

[28] More discussion on how matching can reduce model dependence, see (Ho et al. 2007)

assumption and calculation procedure as previous chapters, and the only difference among the three methods is the weighting schemes are distinct.[29]

We are able to characterize that the direct matching combined method has better performance on reducing the model dependency problem for the estimation of the treatment effect. For the men's group, the treatment effect is 1051 with fewer independent variables and is 1049 with more independent variables. This is in contrast to the OLS (where the treatment effect is 1117 with fewer independent variables and is 970 with more covariates) and the PSM combined method (where the treatment effect is 1197 with fewer regressors and is 1059 with more independent variables). This conclusion is the same as that in the women's group.

Taken together, in contrast to the OLS and PSM combined method, the direct matching combined method provides quite stable estimation results with different model assumptions. This feature is mainly because the new weights orthogonalize the dummy variable on the JTPA assignment and all the background characteristics are quite similar between the compared groups. The results imply that the direct matching combined method can effectively reduce the model dependence problem on the raw dataset after using the weighting scheme.

4.4 Empirical application: NSW program

Concerns about if our combined method in natural experiment study is suitable to be applied in Accounting, Economics and Finance disciplines are consistent with similar methodology studies (i.e., Kullback (1968), Imbens, Spady, and Johnson (1998), Diamond and Sekhon, (2006), Abadie and Imbens (2007), Ho et al. (2007), and Qin,

[29] The direct matching combined method as well as other PSM methods have similar results as in previous chapters and not reported here.

Zhang, and Leung (2009)). We evaluate our combined method with a job training program, the National Supported Work Demonstration (NSW), which was first introduced by LaLonde (1986) and then developed by Dehejia and Wahba (2002). The idea of this experiment is to obtain a benchmark for treatment effect on increased earnings from the NSW experiment. Further, the control group in the randomized experiment is replaced by individuals (obtained from a non-experimental Current Population Survey) who are not involved in the training program and determine which method using the non-experimental data can recover the benchmark of treatment effect of job training program on salary (obtained from NSW).

We expand the NSW analysis and compare the performance of the OLS without balancing checking step with our proposed method (combing direct matching scheme with OLS) and the PSM method (combined with OLS).

The data includes 185 individuals (from NSW) who receive the training program and 15992 observations (from non-randomized CPS) who are not participating training program.

We follow the designs presented in Dehejia and Wahba (1999), and use the regression model below:

$$\mathrm{re78}_i = cons + treat_i + age_i + educ_i + black_i + hispan_i + married_i + nodegree_i + re74_i + re75_i + u74_i + u75_i + \varepsilon_i \qquad (14)$$

where the dependent variable is defined as the increased earnings in the year 1978. Treat is defined as a dummy variable that equals 1 if the observation is participating in the training program from the NSW, and otherwise it is in the control group. The controls include (1) age, (2) years being in the school: educ, (3) black: dummy for black or not, (4) hisp: dummy for Hispanic or not, (5) married: dummy for married or not, (6): nodegree: dummy for not obtained diploma or not, (7) re74: increased earnings in the year 1974, (8)re75: increased earnings in the year 1975, (9) u74: dummy for not having a job in the year 1974 nor not, and (10) u75: dummy for not having a job in the year 1975 or not. Additional details about the data and program can be obtained from Dehejia and Wahba (1999).

Hainmueller (2012) and Hainmueller and Xu (2013) focused on the covariates' balance check and showed that the EB can achieve a better-balanced condition compared with the raw dataset. They also indicate that the mean difference between the treatment and control group estimated by the EB (1761) is very close to the benchmark result (1794) obtained from the randomized data, which indicates that the EB method can efficiently recover the causal inference when the non-experimental data are employed.

We contribute to the literature and extend the analysis of Dehejia and Wahba (1999), Hainmueller (2012) and Hainmueller and Xu (2013). In contrast to their study who use the mean difference as the treatment effect, we test our proposed direct matching combined technique and compare the performance of the OLS with control variables (without matching), the direct matching combined method and the PSM combined method.[30]

Table 23 compares our estimation of the effect of job training on people's earnings with the three methods with control variables. Similar to previous examples, we follow the same steps to estimate the treatment effects of the effect of the training program on increased earnings. The literature (i.e., Dehejia and Wahba (1999)) found that the treatment effect obtained by the randomized experiment should be $1794 on average, which should be the benchmark outcome that we target. We begin by following an OLS similar to previous examples to estimate the treatment effect from equation (14). The result is reported in the second left column of table. The table

[30] We only report the kernel matching scheme. Other matching schemes provide similar estimation results and hence are not reported.

suggests that the treatment effect from the non-experimental data is $1672.426, which is far from the benchmark treatment effect obtained from the randomized experiments in previous studies (i.e., Dehejia and Wahba (1999) and Diamond and Sekhon (2013)). The evidence suggests that the OLS may not provide a precise solution if non-experimental data is used.

The direct matching combined method is reported in the third column of Table 46. The estimates in the diagram display that the effect of the NSW training program on increased income is $1776.685. The estimated result produce as very small bias compared with the benchmark provided by Dehejia and Wahba (1999), which indicates that the direct matching combined method can efficiently produce precise causal inferences when the non-experimental data are employed.

The last column illustrates that the original treatment effects estimated using the PSM combined method are generally similar with those estimated by the OLS. Having the training program generally increases earnings by 1669.661 dollars (which is significant at the 5% level). Hence, the treatment effects estimated by the PSM combined method generally provide similar conclusions as the OLS, although the magnitude and significance level are not as precise as those of the direct matching combined method.

5. Conclusion

The exogenous shock generated by natural experiment create an opportunity for researchers to obtain causality relationships. The increased awareness about the natural experiment can benefit researchers who suffer from endogenous problem by exploiting the exogenous shock of the variable of interest, but can also produce the unreliable "causality relationship" if the randomness assumption required by natural experiment does not hold. It is the latter which generates common but easy to neglect concerns among

Accounting, Finance and Economic researchers.

This study evaluates the issue that if the randomness criterial holds and if not, how the problem of unbalanced covariates between treatment and control groups should be addressed to improve the causal inference.

More specifically, we focus on the imbalance problem in the nature experiment studies that scholars may ignore in academic journals. We review 372 empirical natural experiment studies in Economics, Finance, and Accounting journals, and show that the majority papers do not check the first moment, higher moments, and joint imbalanced problems.

In order to figure out what can be wrong if the balanced condition is not satisfied as required by natural experiment, we produce Monte Carlo simulations, where covariates in treatment and control groups are distinct and the randomness assumption does not hold, as the major problem we argue that scholars may ignore in practice. The results present that the estimations of the original imbalanced dataset implemented by Ordinary Least Square (OLS), a commonly used method in natural experiment study, suffers from the model dependence problem since with different model assumptions (with respect to different number of the regressors included in the model) the estimation results become much more volatile compared with the results after we deal with the imbalanced condition problem via the combined method step.

In addition, we reanalyse the examples by using PSM combined method. PSM methods are commonly used for estimation of binary treatment effects based on the assumption of the selection of observables. One of the main criticisms of the PSM method is that, in practice, it can be challenging to ensure that the distributions of the propensity scores of the control and treatment groups are balanced (see for example Rosenbaum and Rubin, (1983, 1985); Dehejia and Wahba, (2002) for tests to check for balance). We investigate the balanced condition required by the PSM and extend the analysis in our studies to show that PSM may not produce both local balanced condition (if each individual covariate is the same between compared groups) and global balanced condition (if all the covariates are balanced jointly between treatment and control groups) if the analyzed data quality is low (the natural experiment is not random).We provide the concluding remarks that estimation performance after PSM matching (combined with OLS) is even worse than the OLS (used in original study) without matching step, since the Mean Squared Error (MSE) for PSM method is much higher than that of the OLS method.

We show in the text with an intuitive example on when the PSM combined method may perform worse than the original OLS study without matching step: PSM combined method initially does quite well regarding its estimation performance when the balanced condition is met in random dataset since its estimated parameter (1.981) is the same as OLS (1.981). However, as the balanced condition required by PSM deteriorates, the PSM combined method's performance suffers compared with OLS without matching step, and the potential MSE with PSM combined method goes up dramatically compared with that from OLS.

In our reviewed 372 natural experiment studies, conditional on the Propensity Score Matching (PSM) method included in the papers, only 10.53% of the papers check the balanced condition required by the PSM method, in doing so, the potential estimation problem may be a concern as we illustrate in the Monte-Carlo section and the empirical study section.

We further examine different natural experiment studies claiming causal inference in influential economic, financial and accounting journals, and compare the estimation performance between their original results (without matching step) and the results after dealing with the imbalanced condition problem with our combined method (direct matching combined with regression).

Our first influential example is to use the Sarbanes–Oxley Act (SOX) as a natural experiment. For instance, concerning the corporate board structure is an endogenous variable, an experiment design proposed by Chhaochharia and Grinstein (2009) employs SOX as a natural experiment to study the cause inference of board independence on CEO compensation. Such a “natural experiment” is widely used in the literature: For example, see Duchin, Matsusaka, and Ozbas (2010); Chen (2014); Chen, Cheng, and Wang (2015); and Armstrong, Core, and Guay (2014), among others.

Chhaochharia and Grinstein (2009) explore the fact that SOX requires NYSE and NASDAQ to propose the regulation that all of the listed firms should have more than 50% independent board of directors. Despite its strength of design, that is, firms which did not satisfy this requirement were serendipitously forced to increase their board independence level (thus, the variation in board structure of the affected firms appears exogenous), this experiment lacks balance checking in terms of the characteristics between the groups compared in the study. Indeed, a part of the predicament in this design is that the assignment of a treatment variable to the subjects is not random as the evidence shows that the characteristics are significantly different among the groups compared. To overcome this challenge, we deal with the imbalanced problem and illustrate that the treatment effect is insignificant and 25% lower than the report in the original study where the imbalanced condition problem exits. To evaluate the statistical properties of these methods, we implement bootstrap simulation tests on this experiment design and show that our proposed direct matching combined method leads to lower bias, and a lower MSE in comparison to the OLS method (used in the original natural experiment study) and the commonly used matching scheme (i.e., PSM method).

Our second influential example is provided by Angrist et al. (2002) who investigate the causal effect of using school vouchers (distributed randomly by lotteries) on the number of hours that the winners of the lotteries work every week. Angrist et al. (2002)

conclude that the winners work less than their peers who lose in the lotteries. Although the randomness assumption is still a concern since Rubin's B ratio is 30.4, which indicates that the samples compared are not sufficiently and jointly balanced with respect to the background characteristics, the quality of the dataset is much better compared with previous case since most of the covariates are quite similar between compared groups. Hence, the estimation process requires fewer matching activities, in doing so, the estimation results after dealing with the imbalanced condition problem become less volatile

We evaluate the performance of those methods based on the precision of coefficients. Although such a comparison does not include all the statistical properties, we focus on the bias and MSE, and the results show that our proposed direct matching combined method dominates OLS (without matching step) and PSM combined method. Nevertheless, because the data quality is much better than the previous two examples, the estimation results after dealing with the imbalanced condition problem become less volatile, in doing so, the data quality matters with respect to the estimation performance which is consistent with our Monte-Carlo study.

As most of the natural experiment studies in Accounting, Economics and Finance disciplines have the model dependent problem, at least partly, this means that the estimation result depends on the assumption of the model specification (i.e., the assumption on what types of independent variables are included in the model) (Ho et al. 2007). We make use of the "National Job Training Partnership Act (JTPA)," an influential randomized experiment to evaluate the performance of a methodology (i.e., Heckman et at. (1997) among others) to show that if our estimation results will change much with our combined method. The JTPA is devolved by the US Department of Labor to evaluate the effect of the JTPA on program participants' earnings. The JTPA randomly choses participants according to their background characteristics who were qualified to apply to the program in 16 service delivery areas in the United States of America, such that part of the qualified candidates is selected into the treatment group randomly, whereas the others are automatically left to be in the control group.

In contrast to Heckman et at. (1997)'s study who focus on the self-selection bias issue and calculate the probability that if any observation will receive the JTPA program, we combine a direct matching method with OLS method and figure out that our combined method has better performance on reducing the model dependency problem for the estimation of treatment effect (the treatment effect is 1051 with less independent variables and is 1049 with more independent variables), in contrast to OLS without dealing with imbalanced condition step (the treatment effect is 1117 with less independent variables, and is 970 with more independent variables) as well as a combined of PSM and OLS method (the treatment effect is 1197 with less independent variables, and is 1059 with more independent variables)

Concerns about if our combined method in natural experiment study is suitable to be applied in Accounting, Economics and Finance disciplines are consistent with similar methodology studies (i.e., Kullback (1968), Imbens, Spady, and Johnson (1998), Diamond and Sekhon, (2006), Abadie and Imbens (2007), Ho et al. (2007), and Qin, Zhang, and Leung (2009)). We evaluate our combined method with a job training program, the National Supported Work Demonstration (NSW), which was first introduced by LaLonde (1986) and then developed by Dehejia and Wahba (2002). The idea of this experiment is to obtain a benchmark for treatment effect on increased earnings from the NSW experiment. Further, the control group in the randomized experiment is replaced by individuals (obtained from a non-experimental Current Population Survey) who are not involved in the training program and determine which method using the non-experimental data can recover the benchmark of treatment effect of job training program on salary (obtained from NSW).

We expand the NSW analysis and compare the performance of the OLS without balancing checking step with our proposed method (combing direct matching scheme with OLS) and the PSM method (combined with OLS).

In this design, the estimates by OLS without matching step (1672.426) and PSM combing with OLS method (1669.661) are significantly lower than the threshold (1794) set by Dehejia and Wahba (2002) in comparison to the direct matching combing with OLS estimation (1776.685).

In this study, we provide recommendations to the researchers in the natural experiment area, and propose to apply the direct matching combined method which enables the creation of balanced compared groups that is easy to apply and provide better causal inference sine it has a wide variety of advantages regarding statistical properties than the conventional OLS method and the PSM combined method. Such process is suggested to be considered by Accounting, Finance and Economic researchers when they need to deal with the non-randomness problem in natural experiment studies.

In all cases, we thank the authors for making data available that are easier for interested researchers to replicate. None should be faulted, particularly for an unawareness of the new framework we propose here, which is based on a process developed many years after their original studies.

References

Abadie, A, G., Imbens, G., 2002. Simple and bias-corrected matching estimators for average treatment effects. Working paper. Available at: http://www.nber.org/papers/t0283.

Akey, P, 2015. Valuing Changes in Political Networks: Evidence from Campaign Contributions to Close Congressional Elections. Review of Financial Studies 28,3188-3233

Angrist, J., Krueger, B, A., 1991. Does Compulsory School Attendance Affect Schooling and Earnings? The Quarterly Journal of Economics 106(4), 979–1014.

Angrist, J., Bettinger, E., Kremer, M.,2006. Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. American Economic Review 96(3), 847-862.

Angrist, J; Bettinger, E; Bloom, E; King, E and Kremer, M., 2002 Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. American Economic Review 92 (5), 1535-1558.

Armstrong, S, C., Core, E, J., Guay, R, W., 2014. Do independent directors cause improvements in firm transparency? Journal of Financial Economics 113(3), 383-403.

Atanasov, V., Black, B., 2016. Shock-Based Causal Inference in Corporate Finance and Accounting Research. Critical Finance Review 5, 207-304.

Austin, P, C., 2008. A Critical Appraisal of Propensity Score Matching in the Medical Literature Between 1996 and 2003. Statistics in Medicine 27 (12), 2037–2049.

Alberto, A.,Matthew, M,C., Martin .2017 .Endogenous Stratification in Randomized Experiments. Working PAPER. NBER

Becker, S.O., Ichino, A., 2002. Estimation of average treatment effects based on propensity scores. The STATA Journal 2, 358-377.

Bettinger,E., Kremer,M., Saavedra, J,E., 2010. Are educational Vouchers only redistributive? Orcid Economic Journal 120(546), 204-228.

Campbell, D,T., and Stanley, J,C.,1963. Experimental and quasi-experimental designs for research. Houghton Mifflin Company, Boston.

Campbell, D, T., Boruch, R., 1975. Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six ways in which quasiexperimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennet & A. A. Lumsdaine (Eds.), Evaluation and Experiment, Academic Press, New York

Chemmanur, T., Tian, X., 2013. Do anti-takeover provisions spur corporate innovation? Working paper. Available at:

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1572219

Chen, D., 2014. The Non-monotonic effect of board independence on credit ratings. Journal of Financial Services Research 45(2), 145-171.

Chen,W., 2016. Does inflation targeting work well? Evidence from CEE countries. Acta Oeconomica 66(3), 375-392

Chen, X., Cheng, Q., Wang, X., 2015.Does increased board independence reduce earnings management? Evidence from recent regulatory reforms. Review of Accounting Studies 20(2), 899-933.

Chhaochharia, V., Grinstein, Y., 2009. CEO compensation and board structure. The Journal of Finance 64(1), 231–261.

Cohen, L., Coval, J., Malloy, C., 2011. Do powerful politicians cause corporate downsizing? Journal of Political Economy 119, 1015–1060.

Colin, A., Trivedi, P.K., 2009. Microeconometrics Using Stata, first ed. Stata Press Publication, College Station, TX.

Cressie, N. A. 1998. Aggregation and interaction issues in statistical modeling of spatiotemporal processes. Geoderma, 85 (2-3), 133-140.

de Mendonça, H.F., de Guimarães e Souza, G.J., 2012. Is inflation targeting a good remedy to control inflation? Journal of Development Economics 98, 178-191.

Dehejia, R. H., Wahba, S., 1999.Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association 94, 1053-1062

Dehejia, R. H., Wahba, S., 2002. Propensity Score-Matching Methods For Non-experimental Causal Studies.Review of Economics and Statistics 84,151-161.

Deming, E., Stephan,F.,1940. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. The Annals of Mathematical Statistics 11(4), 427-444

Diamond, A., Sekhon, J., 2006. Genetic Matching for Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Working Paper Available at:

https://ideas.repec.org/a/tpr/restat/v95y2013i3p932-945.html

Diamond, A., Sekhon, J.S., 2013. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Review of Economics and Statistics 95, 932-945.

Diamond, J., Robinson, J., 2010.Natural Experiments of History.The Belknap Press of Harvard University Press: Cambridge,MA.

Duchin, R., Matsusaka, G, J., Ozbas, O., 2010. When are outside directors effective? Journal of Financial Economics 96(2), 198-214.

Dunning, T., 2008. Improving Causal Inference: Strengths and Limitations of Natural Experiments. Political Science Quarterly 61 (2), 282–93.

Dunning, T., 2010. Endogenous Oil Rents. Comparative Political Studies 43 (3), 379-410

Dunning, T., 2012. Natural Experiments in the Social Sciences: A Design-Based Approach. Cambridge University Press: Cambridge, United Kingdom.

Dravis, F., 2007. The Role of Independent Directors after Sarbanes-Oxley.1st ed. Chicago: American Bar Association

Fowler, A., Hall, B,A., 2015. Congressional seniority and pork: A pig fat myth? European Journal of Political Economy 40, 42-56

Freedman, D., Pisani, R., Purves, R., 1997. Statistics. 3rd Ed. New York: Norton.

Glynn, A,N.,Kashin,K., 2018.Front-door versus back-door adjustment with unmeasured confounding: Bias Formulas for Front-door and Hybrid Adjustments with Application to a Job Training Program. Journal of the American Statistical Association, forthcoming.

Goldman, E, Rocholl, J, So, J, 2013. Politically Connected Boards of Directors and The Allocation of Procurement Contracts 17(5), 1617-1648

Gropper, D,M.,Jahera, J,S., Park, J,C.,2013. Does it help to have friends in high places? Bank stock performance and congressional committee chairmanships. Journal of Banking and Finance 37(6), 1986-1999

Gregory, R., McNulty, E, J., Krasno, S,J., 2009. Observing the counterfactual? The search for political experiments in nature. Political Analysis 1,341–57.

Gross, C., Konigsgruber, R.,Pantzalis, C.,Perotti, P., 2016. The financial reporting consequences of proximity to political power. Journal of Accounting and Public Policy 35(6),609-634

Hainmueller, J., 2012. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. Political Analysis 20, 25-46.

Hainmueller, J., Xu, Y., 2013. Ebalance: a Stata package for entropy balancing. Journal of Statistical Software 54, 1-18.

Heckman, J., Ichimura, H., and Todd, P. (1997). Matching as an econometric evaluation estimator evidence from evaluating a job training program. Review of Economic Studies, 64, 605–654

Heckman, J, .J., Ichimura, H., Todd, P., 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261-294.

Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1999). The Economics and Econometrics of Active Labor Market Programs. In Ashenfelter, O. and Card, D., editors, Handbook of Labor Economics, Volume III. North Holland

Heckman,J,J., and Urzua, S., (2010). Comparing IV With Structural Models: What Simple IV Can and Cannot Identify. Journal of Econometrics, 156(1),27-37

Hermalin, B, E., Weisbach, S, M., 1998. Endogenously Chosen Boards of Directors and Their Monitoring of the CEO. American Economic Review, 88(1), 96–118.

Hermalin, B, E., Weisbach, S, M., 2003. Boards of Directors as an Endogenously Determined Institution: A Survey of the Economic Literature. Federal Reserve Bank of New York Economic Policy Review 9(1), 7–26.

Hellerstein, J., and G. Imbens. 1999. Imposing moment restrictions from auxiliary data by weighting. The Review of Economics and Statistics 81,1–14.

Hirano, K., G. Imbens, and G. Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–89.

Ho, D.E., Imai, K., King, G., Stuart, E.A., 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15, 199-236.

Huber, M.,Laffers, L.,Mellace, G.,2017. Sharp IV bounds on average treatment effects on the treated and other populations under endogeneity and noncompliance. Journal of Applied Econometrics 32(1), 56-79

Iacus, S., G. King, and G. Porro. 2009. Causal inference without balance checking: Coarsened exact matching. Mimeo Harvard University.

Imbens, W,G., Spady,H,Richard., Johnson, P., 1998. Information Theoretic Approaches to Inference in moment condition models 66(2), 333-357

Imbens G., 2010. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua. Journal of Economic Literature, 48(2),399-423

Kullback, S., 1959. Statistics and Information Theory, 1st ed. Wiley& Sons, Inc, New York, NY.

Kullback,S., 1968. Probability Densities with Given Marginals. The Annals of Mathematical Statistics 39(4), 1236-1243

Kirmani, H,B., Holmes, V, M., Muir, D, A., 2016. Long-Term survival and freedom from reintervention after off-pump coronary artery bypass grafting, A Propensity-Matched Study. Circulation 134(17), 1209-1220

Kostovetsky, L.,2015. Political capital and moral hazard. Journal of Financial Economics 116(1), 144-159

LaLonde, R., 1986. Evaluating the econometric evaluations of traing programs with experimental data. American Economic Review 76(4), 604-620

Lin, S., Ye, H., 2007. Does inflation targeting really make a difference? Evaluating the treatment effect of inflation targeting in seven industrial countries. Journal of Monetary Economics 54, 2521-2533.

Lin, S., Ye, H., 2009. Does inflation targeting make a difference in developing countries? Journal of Development Economics 89, 118-123.

Nannicini,T.,2007. Simulation-based sensitivity analysis for matching estimators. Stata Journal 7(3), 334-350

Porter, M., 1992. Capital disadvantage: America's failing capital investment system. Harvard Business Review 70, 65-82.

Peter, R, M., Kenneth,R,T.,Alexey,G.,2007. Using state administrative data to measure program performance. The Review of Economics and Statistics 89(4), 761-783

Pearl, J., 2010. The foundations of causal inference. Sociological Methodology 40(1),75–149.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. The Annals of Statistics 22, 300–325

Qin, J., Zhang, B., Leung, D., 2009. Empirical Likelihood in Missing Data Problems. Journal of the American Statistical Association 104, 1492–1503.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41-55.

Rosenbaum, P.R., Rubin, D.B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 39(1), 33-38.

Rubin, D., 2006. Matched Sampling for Causal Effects. Cambridge University Press, Cambridge.

Samarina, A., Terpstra, M., De Haan, J., 2014. Inflation targeting and inflation performance: a comparative analysis. Applied Economics 46, 41-56.

Schennach, M, S., 2007. Point estimation with exponentially titled empirical likelihood. The Annals of Statistics 35(2), 634-672

Sekhon, J.S., 2009. Opiates for the matches: matching methods for causal inference. Annual Review of Political Science 12, 487-408.

Sekhon, J. S., Titiunik, R., 2012. When natural experiments are neither natural nor experiments. American Political Science Review 106(1), 35-57.

Stefano,M,I., Gary,K., Giuseppe., 2012.Causal Inference Without Balance Checking: Coarsened Exact Matching. Political Analysis, 20(1), 1-24.

Smith, A, J., Todd, P,E.,2001. Reconciling conflicting evidence on the performance of propensity score matching methods. American Economic Review, 91(2), 112-118

Snyder, A.J., Welch, I., 2017. Do powerful politicians really cause corporate downsizing? Journal of Political Economy, 125(6), 2225-2231.

Uribe, C., Murnane, R,J., Willett, J,B.,Somers, M,A., 2006. Expanding school enrollment by subsidizing private schools: Lessons from Bogota. Comparative education review 50(2), 241-277

Vega, M., Winkelried, D., 2005. Inflation targeting and inflation behavior: a successful story? International Journal of Central Banking 1, 153-175.

Wooldridge, M,J., 2009. Introductory econometrics: a modern approach. South-Western Cengage Learning, Mason, Ohio

Zaslavsky, C., 1988. Integrating mathematics with the study of cultural traditions.6th Annual International Conference on Mathematical Education, Budapest, Hungary. (ED 303 540)

Zhao, Z., 2004. Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. Review of Economics and Statistics 86, 91–107.

Table 1 Sample articles on natural experiment

The information in the table is obtained from the Web of Science database and reports article numbers across years and journals which imply or implicitly imply the natural experiment as discussed in chapter two. We exclude the papers focusing on Regression Discontinuity Design (RDD) approach, Instrument Variable (IV) approach, and the time series models approach so that the papers satisfy the method settings according to the method description.

Journal	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018(H)	Total
American Economic Review	1	1	1	1	1	1	1	1	1	1	5	3	3	0	4	3	2	1	3	34
Econometrica	0	0	0	0	0	0	0	2	1	1	0	1	0	0	0	1	1	1	0	8
Financial Management	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	2	2	1	2	10
Journal of Accounting and Economics	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	2	2	2	9
Journal of Accounting Research	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2	3	0	1	1	9
Journal of Banking and Finance	0	0	0	0	2	1	1	1	0	2	0	1	3	0	4	4	3	4	3	29
Journal of Corporate Finance	0	0	0	0	0	1	0	1	1	1	1	0	2	2	2	2	2	5	5	25
Journal of Development Economics	1	0	1	1	0	0	1	1	2	1	1	1	1	2	5	3	0	4	1	26
Journal of Finance	0	0	0	1	0	0	0	1	0	1	0	1	0	6	2	5	3	3	1	24
Journal of Financial Economics	0	0	3	1	1	0	1	1	1	0	1	1	3	5	3	3	3	5	3	35
Journal of Financial Intermediation	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	1	1	2	2	10
Journal of Financialand QuantitativeAnalysis	0	0	0	0	0	1	0	0	0	0	2	0	1	0	2	1	6	3	1	17
Journal of Labor Economics	0	1	1	1	0	0	1	1	1	1	1	0	1	0	2	1	0	1	0	13
Journal of Law and Economics	1	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	8
Journal of Political Economy	0	0	0	0	0	1	1	0	0	1	2	2	0	1	0	0	5	0	0	13
Review of Finance	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	3	1	2	1	11
The Accounting Review	0	0	0	0	0	0	0	1	1	1	2	1	0	3	3	1	1	1	2	17
The Quarterly Journal of Economics	1	1	1	1	1	1	1	2	1	0	1	0	0	2	1	2	1	0	1	18
The Review of Economics and Statistics	0	1	2	2	0	1	1	0	2	1	2	5	4	5	0	2	3	1	1	33
The Review of Financial Studies	0	0	0	0	0	0	0	0	1	1	1	3	2	1	0	5	3	6	0	23
Total	4	5	9	9	5	8	8	12	13	13	21	20	25	31	33	43	40	44	29	372

Table 2: A review of balanced condition problem in empirical natural experiment literature

The table illustrates the balanced condition check results for our sample. It provides the information on the relative ratio of the balance checked papers and the non-checked articles. The information in the table is obtained from the Web of Science database.

Strategy	Yes	No
Checked First moment Balanced condition of Natural Experiment	152	220
(Percentage)	40.86%	59.14%
Checked Higher moments (second, third and fourth moments)	65	307
Balanced condition of Natural Experiment (Percentage)	17.47%	82.53%
Checked Global (Jointly) Balanced condition of Natural Experiment	31	341
(Percentage)	8.33%	91.67%
Include PSM methods	38	334
(Percentage)	10.22%	89.78%
If PSM is included, is PSM Balanced Condition checked?	4	34
(Percentage)	10.53%	89.47%

Table 3 Balance checking for covariates between treatment and control groups

The sample includes 40,000 observations. The data generator process is explained in method section. The L statistics is following Iacus, King, and Porro (2008). A multivariate $L1$ distance, univariate $L1$ distances, difference in means and empirical quantiles difference are reported. The $L1$ measures are computed by coarsening the data according to breaks and comparing across the multivariate histogram.

Balance Checking for covariates between treatment and control groups (Monte Caro Simulation)

Variable			Control Group			Treatment Group	
	$L1$	mean	min	0.25	0.5	0.75	max
y_{1i}	0.2833	3	3	3	3	3	3
y_{2i}	0.2827	3	3	3	3	3	3

Table 4 Estimates of the Monte Carlo Simulation Study.

The data generator process is explained in method section. The direct matching (EB) calculation process in explained in method section. All the three methods use the same model assumption, dataset, except the weighting scheme is different.

Compare direct matching (combined with OLS) with OLS (without matching scheme) and PSM (combined with OLS) in Monte Caro Simulation

Benchmark Settings	Benchmark Treatment Effect (Unadjusted)	Direct matching combined with OLS (Unadjusted)	Direct matching combined with OLS (Adjusted)	PSM combined with OLS (Unadjusted)	PSM combined with OLS (Adjusted)	OLS (without matching) (Unadjusted)	OLS (without matching) (Adjusted)
Original Treatment Effects	0.044	0.036	0.035	0.019	-0.006	5.993	-0.008
Bias	0	0.008	0.009	0.025	0.050	5.949	0.052
No of observations	20,000	20,000	20,000	20,000	20,000	20,000	20,000

Table 5 Compare PSM with OLS in the Monte Carlo Simulation Study.

We create one random dataset, with two covariates for both control group and treatment group (both groups with 10,000 observations) from a uniform distribution (0,10). We also create one non-random dataset, with two covariates for the control group (with 10,000 observations) from a uniform distribution (0,10) and two covariates for the treatment group (with 10,000 observations) from a uniform distribution (8,18). All the other settings are following the same procedure as in the section 2.3.

Panel A: Random dataset results

Comparation between PSM method and OLS method (Random Dataset)

Benchmark Settings	OLS	PSM(Radius 0.01)	PSM(Radius 0.03)	PSM(Kernel Matching)
Treatment Effects	1.981	1.981	1.981	1.981
Observations	20,000	20,000	20,000	20,000

Panel B: Non-random dataset results

Comparation between PSM method and OLS method (Non-Random Dataset)

Benchmark Settings	OLS	PSM(Radius 0.01)	PSM(Radius 0.03)	PSM(Kernel Matching)
Treatment Effects	1.958	1.942	1.955	1.956
MSE	0.001	0.003	3.925	3.926
Observations	20,000	20,000	20,000	20,000

Table 6: Balance checking for covariates in the study of board independence

The sample includes 807 firms in the period 2000-2005. The data are collected according to the instructions from Chhaochharia and Grinstein (2009) and can be obtained from the Supplementary Appendix on the Journal of Finance website. Sales is the natural log of company sales. ROA is the natural log of one plus net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1). Stock Returns is the natural log of the annual gross stock return (dividend reinvested), measured in year (t−1). Tenure is the number of years the CEO served with the firm. More details are explained in the Appendix. The right "Difference" column provides the mean difference between treatment and control groups. Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels according to the T statistics.

Balance Checking for covariates between treatment and control groups

Variable	Control Group					Treatment Group					Difference
	Obs	Mean	Std.Dev.	Min	Max	Obs	Mean	Std.Dev.	Min	Max	
Sales*Dummy(00-02)	4,074	3.87	4.002	0	12.41	768	3.598	3.726	0	11.804	0.272*
Sales*Dummy(03-05)	4,074	3.936	4.063	0	12.636	768	3.715	3.84	0	11.892	0.221
ROA*Dummy (00-02)	4,074	0.024	0.054	-0.597	0.399	768	0.031	0.071	-1.209	0.283	-0.007***
ROA *Dummy (03-05)	4,074	0.019	0.083	-3.781	0.294	768	0.028	0.054	-0.314	0.316	-0.009***

Stock Returns*Dummy (00-02)	4,074	0.031	0.304	-1.852	2.089	768	0.043	0.324	-1.556	1.551	-0.012
Stock Returns *Dummy (03-05)	4,074	0.048	0.259	-2.198	1.318	768	0.055	0.257	-1.35	1.34	-0.007
CEO Tenure	4,074	1.897	0.692	0	3.892	768	2.261	0.874	0	3.912	-0.364***

Table 7 Summary of direct matching (EB) results for board independence study (After Weighting)

The sample includes 807 firms in the period 2000-2005. The data are collected according to the instructions from Chhaochharia and Grinstein (2009) and can be obtained from the Supplementary Appendix on the Journal of Finance website. Sales is the natural log of company sales. ROA is the natural log of one plus net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1). Stock Returns is the natural log of the annual gross stock return (dividend reinvested), measured in year (t−1). Tenure is the number of years the CEO served with the firm. The direct matching (EB) process is explained in the methodology section. More details on variables are recorded in the Appendix.

Summary of Entropy balancing results for board independence study (After Weighting)

	Treatment Group			Control Group		
Variable	mean	variance	skewness	mean	variance	skewness
Sales*Dummy(00-02)	3.598	13.88	0.203	3.598	13.88	0.203
Sales*Dummy(03-05)	3.715	14.75	0.191	3.715	14.75	0.191
ROA*Dummy (00-02)	0.031	0.005	-6.24	0.031	0.005	-6.24
ROA *Dummy (03-05)	0.028	0.003	0.422	0.028	0.003	0.42
Stock Returns*Dummy (00-02)	0.043	0.105	0.494	0.043	0.105	0.494
Stock Returns *Dummy (03-05)	0.055	0.066	-0.307	0.055	0.066	-0.308
CEO Tenure	2.261	0.764	-0.084	2.261	0.764	-0.084

Table 8 Estimates of the board independence effect on CEO compensation

The sample includes 807 firms in the period 2000-2005. The dependent variable is defined as the natural log of CEO compensation, Dummy (Noncompliant Board02) is coded as 1 if the firm did not comply with the regulation in the year 2002, and as 0 otherwise. Dummy (03-05) is set to 1 if the observation is in the period 2003-2005, and 0 otherwise. Dummy ('00-'02) equals 1 if the observation is in the period 2000-2002, and 0 otherwise. Controls include (1) Sales*Dummy (00-02), (2) Sales*Dummy(03-05), (3) ROA*Dummy (00-02), (4) ROA *Dummy (03-05), (5) Stock Returns*Dummy (00-02), (6) Stock Returns *Dummy (03-05), (7) CEO Tenure, (8) Firm fixed effect, and (9) Industry-year fixed effect, which is defined as Fama-French 48 Industry factor times year dummy variable. Sales is the natural log of company sales. ROA is the natural log of one plus net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1). Stock Returns is the natural log of the annual gross stock return (dividend reinvested), measured in year (t−1). Tenure is the number of years the CEO served with the firm. The estimation process is provided in the method chapter. Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels. The real number of observations might have been decreased for specific tests because of missing values, perfect predictions, or multi-collinearity problems.

Estimates of the board independence effect on CEO compensation

Dependent Variable: Ln(Total CEO Compensation)	OLS (in the original natural experiment study)	Direct matching combined method
Independent Variables:	Coefficient	Coefficient
Dummy(Noncompliant Board(02)*Dummy(03-05))	-0.171*	-0.129
Sales*Dummy(00-02)	0.359***	0.144
Sales*Dummy(03-05)	0.325***	0.024
ROA*Dummy(00-02)	0.167	-0.035
ROA*Dummy(03-05)	0.21	0.963

Stock Returns*Dummy(00-02)	0.118***	0.063
Stock Returns*Dummy(03-05)	0.294***	0.172
CEO Tenure	-0.02	-0.064
Firm fixed effect	Yes	Yes
Industry-year fixed effect	Yes	Yes
Constant	Yes	Yes
Number of observations	4842	4842
Adjusted-R Square	0.643	0.697

Table 9 Comparison between the direct matching combined method and the OLS method (in the original natural experiment study)

The table reports the simulation results for the effect of board independence in the CEO compensation study. All of the benchmarks (bias and MSE) are based on 5000 bootstrap simulations (with replacement). In each simulation, we use the direct matching combine method and the OLS method to estimate the effect of board independence on CEO compensation. All of the control variables are the same between the EB method and OLS method and are explained in the experiment design section. The calculation process for each benchmark (bias and MSE) is explained in the simulation section. Definitions for each variable are explained in Appendix A. The real number of simulations might be decreased because of no solutions for weights in a specific EB simulation.

Comparation between the direct matching combined method and the OLS method (in the original experiment study)

Benchmark Settings	Direct matching Combined method	OLS method (in the original experiment study)
Original Treatment Effects	-0.129	-0.171
Bias	0.031	0.048
MSE	0.019	0.221
No of simulations	1000	1000

Table 10: Logit regression results used to estimate propensity score in board independence study

The sample includes 807 firms in the year 2002. The dependent variable is a Dummy (Noncompliant Board). If a firm did not comply with the new regulation in 2002, then this firm is required by the new policy to increase its board independence level, and Dummy (Noncompliant Board) equals 1; otherwise, dummy (Noncompliant Board) equals 0. The baseline model includes control variables such as Sales (the natural log of company sales), ROA (the natural log of one plus net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1)), Stock Returns (the natural log of the annual gross stock return (dividend reinvested), measured in year (t−1)), Tenure (the number of years the CEO served with the firm), and Industry (Fama-French 48 Industry factor). *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Logit regression results used to estimate propensity score in board independence study

Dependent Variable: Dummy (Noncompliant Board)	
Independent Variables:	Coefficient
Sales	-0.188**
ROA	2.912*
Stock Returns	-0.301
CEO Tenure	0.724***
Industry	0.002
Constant	Yes
Number of observations	807
Pseudo-R Squrare	0.067

Table 11: Comparison between the direct matching combined method and the PSM combined method

The table reports the simulation results for the effect of board independence in the CEO compensation study. All of the benchmarks (bias and MSE) are based on 5000 bootstrap simulations (with replacement). In each simulation, we use the direct matching combine method and PSM combined method to estimate the effect of board independence on CEO compensation. All of the control variables are the same between the 2 methods. The calculation process for each benchmark (bias and MSE) is explained in the simulation section. Definitions for each variable are explained in Appendix A. The real number of simulations might be decreased because of no solutions for weights in a specific EB simulation.

Comparation between the direct matching combined method and the PSM combined method

Benchmark Settings	Direct matching combined method	PSM combined method
Original Treatment Effects	-0.129	-0.193*
Bias	0.004	0.034
MSE	0.009	0.041
No of simulations	1956	5000

Table 12: Comparison between the direct matching combined method and various PSMs combined method

The table reports the simulation results for the effect of board independence in the CEO compensation study. All of the benchmarks (bias and MSE) are based on 1000 bootstrap simulations (with replacement). In each simulation, we use the direct matching combined method and PSM combined method to estimate the effect of board independence on CEO compensation. All of the control variables are the same for each method. The calculation process for each benchmark (bias and MSE) is explained in the simulation section. Definitions for each variable are explained in Appendix A. The real number of simulations might be decreased because of no solutions for weights in a specific EB simulation.

Comparison between the direct matching combined method and various PSMs combined method

Benchmark Settings	Direct matching combined method	Local Linear Regression Matching combined method	Radius Matching combined method (Radius=0.01)	Radius Matching combined method (Radius=0.02)	Radius Matching combined method (Radius=0.03)	Nearest Matching combined method
Original Treatment Effects	-0.129	-0.216	-0.180	-0.193	-0.197	-0.307
Bias	0.031	0.036	0.048	0.035	0.036	0.075
MSE	0.020	0.071	0.045	0.042	0.042	0.056
No of simulations	1000	1000	1000	1000	1000	1000

Table 13:Hotelling's T-Squared generalized means test and Rubin's B test

The table reports the Hotelling's T-squared generalized means test and Rubins' B test on the raw data, the data with weights after indirect matching (PSM), and the data with weights after direct (EB) matching. In Hotelling's T-squared generalized means test evaluates whether a set of means equal between two groups, the Null Hypothesis is vectors of means are equal for the two groups. Rubins' B indicates the standardized difference of the means of the linear index of the propensity score in the treatment and control groups. Rubin (2001) demonstrates that B is less than 25 for the compared sample indicates a sufficiently balanced compared group. The asterisk indicates the compared groups are sufficiently imbalanced.

Hotelling's T-squared generalized means test and Rubins' B test

Statistics of the test	Hotelling's T-squared generalized means test (Null Hypothesis:Vectors of means are equal for the two groups)				Rubin's B
Raw Data	F(7,4834):	33.626	Prob>F (7,4834):	0	57.5*
Data After indirect matching	F(7,4821)	0.371	Prob > (7,4821)	0.920	4.6
Data After direct matching	F(7,4576):	0	Prob>F(7,4576):	1	0

=

Table 14: Comparison of characteristics between lottery winners and losers

The descriptive statistics table reports the variables for the full sample in the effect of vouchers program on the time winners are working per week. The covariates include: type of survey, city of the survey, access to phone, age, gender, year of application, month of the interview, strata of residence. The left column reports the summary statistics for lottery winners' group. The middle column provides the results for lottery losers' group. The right column gives the mean difference between the two groups (mean of the variable for the treatment group minus mean of the variable for the control group). Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels according to the T statistics. Data are hand collected by Angrist et al (2002), and they can be obtained from Joshua Angrist's website at MIT.

Balance Checking for covariates in the study of voucher program on number of hours that winner is working per week

Variable	Treatment Group				Control Group				Difference
	Obs	Mean	Std.Err.	Std Dev	Obs	Mean	Std.Err.	Std Dev	
Type of survey	816	0.447	0.017	0.498	761	0.428	0.018	0.495	-0.019
City of survey	816	0.121	0.011	0.327	761	0.114	0.012	0.318	-0.007
Access to phone	816	0.949	0.008	0.221	761	0.941	0.009	0.236	-0.008
Age	816	14.781	0.059	1.677	761	14.912	0.060	1.664	0.131
Gender	816	0.498	0.018	0.500	761	0.490	0.018	0.500	-0.007
Year 1995	816	0.717	0.016	0.451	761	0.739	0.016	0.440	0.022

Year 1997	816	0.178	0.013	0.382	761	0.168	0.014	0.374	-0.009
Inverview Month (1)	816	0.015	0.004	0.120	761	0.017	0.005	0.130	0.002
Inverview Month (2)	816	0.025	0.005	0.155	761	0.029	0.006	0.168	0.004
Inverview Month (3)	816	0.322	0.016	0.468	761	0.283	0.016	0.451	-0.04*
Inverview Month (4)	816	0.176	0.013	0.381	761	0.167	0.014	0.373	-0.010
Inverview Month (5)	816	0.042	0.007	0.200	761	0.074	0.009	0.261	0.032***
Inverview Month (6)	816	0.082	0.010	0.275	761	0.114	0.012	0.318	0.032**
Inverview Month (7)	816	0.216	0.014	0.412	761	0.202	0.015	0.402	-0.013
Inverview Month (8)	816	0.100	0.011	0.301	761	0.084	0.010	0.278	-0.016
Inverview Month (9)	816	0.009	0.003	0.092	761	0.012	0.004	0.108	0.003
Inverview Month (10)	816	0.005	0.002	0.070	761	0.003	0.002	0.051	-0.002
Inverview Month (11)	816	0.006	0.003	0.078	761	0.008	0.003	0.089	0.002

Inverview Month (12)	816	0.002	0.002	0.049	761	0.008	0.003	0.089	0.005
Strata of residence (1)	816	0.138	0.012	0.346	761	0.145	0.013	0.352	0.006
Strata of residence (2)	816	0.549	0.017	0.498	761	0.553	0.018	0.497	0.004
Strata of residence (3)	816	0.137	0.012	0.344	761	0.127	0.012	0.334	-0.010
Strata of residence (4)	816	0.009	0.003	0.092	761	0.003	0.002	0.051	-0.006
Strata of residence (5)	816	0.001	0.001	0.035	761	0.001	0.001	0.036	0.000
Strata of residence (ms)	816	0.165	0.013	0.372	761	0.171	0.014	0.377	0.005

Table 15: Summary of direct matching Results: Before and After Weighting

The descriptive statistics table reports the variables for the full sample in the effect of vouchers program on the time winners are working per week. The covariates include: type of survey, city of the survey, access to phone, age, gender, year of application, month of the interview, strata of residence. The summary of statistics reflects the observations before and after the entropy balancing for the board independence study. Treatment group observations are candidates who win the lottery. Otherwise are control group observations. The EB calculation is explained in Method section. Data are hand collected by Angrist et al (2002), and they can be obtained from Joshua Angrist's website at MIT.

Panel A:Summary of direct matching results for impact of vouchers program (Before Weighting)

	Treatment Group			Control Group		
Variable	mean	variance	skewness	mean	variance	skewness
Type of survey	0.447	0.248	0.212	0.428	0.245	0.289
City of survey	0.121	0.107	2.320	0.114	0.101	2.424
Access to phone	0.949	0.049	-4.060	0.941	0.056	-3.738
Age	14.780	2.812	0.294	14.910	2.770	0.246
Gender	0.498	0.250	0.010	0.490	0.250	0.039

Year 1995	0.717	0.203	-0.963	0.739	0.193	-1.085
Year 1997	0.178	0.146	1.686	0.168	0.140	1.774
Inverview Month (1)	0.015	0.015	8.063	0.017	0.017	7.454
Inverview Month (2)	0.025	0.024	6.150	0.029	0.028	5.623
Inverview Month (3)	0.322	0.219	0.760	0.283	0.203	0.966
Inverview Month (4)	0.177	0.146	1.697	0.167	0.139	1.787
Inverview Month (5)	0.042	0.040	4.587	0.074	0.068	3.266
Inverview Month (6)	0.082	0.075	3.044	0.114	0.101	2.424
Inverview Month (7)	0.216	0.169	1.383	0.202	0.162	1.482
Inverview Month (8)	0.101	0.091	2.658	0.084	0.077	2.997
Inverview Month (9)	0.009	0.009	10.660	0.012	0.012	9.031
Inverview Month (10)	0.006	0.006	12.660	0.008	0.008	11.130

Inverview Month (11)	0.002	0.002	20.120	0.008	0.008	11.130
Strata of residence (1)	0.139	0.119	2.093	0.145	0.124	2.022
Strata of residence (2)	0.549	0.248	-0.197	0.553	0.248	-0.214
Strata of residence (3)	0.137	0.119	2.108	0.128	0.111	2.234
Strata of residence (4)	0.009	0.009	10.660	0.003	0.003	19.430
Strata of residence (5)	0.165	0.138	1.801	0.171	0.142	1.749

Panel B: Summary of Entropy balancing results for impact of vouchers program (After Weighting)

	Treatment Group			Control Group		
Variable	mean	variance	skewness	mean	variance	skewness
Type of survey	0.447	0.248	0.212	0.447	0.248	0.212

City of survey	0.121	0.107	2.320	0.121	0.107	2.320
Access to phone	0.949	0.049	-4.060	0.949	0.049	-4.060
Age	14.780	2.812	0.294	14.780	2.812	0.293
Gender	0.498	0.250	0.010	0.498	0.250	0.010
Year 1995	0.717	0.203	-0.963	0.717	0.203	-0.963
Year 1997	0.178	0.146	1.686	0.178	0.146	1.686
Inverview Month (1)	0.015	0.015	8.063	0.015	0.015	8.063
Inverview Month (2)	0.025	0.024	6.150	0.025	0.024	6.150
Inverview Month (3)	0.322	0.219	0.760	0.322	0.219	0.761
Inverview Month (4)	0.177	0.146	1.697	0.176	0.146	1.698
Inverview Month (5)	0.042	0.040	4.587	0.042	0.040	4.587
Inverview Month (6)	0.082	0.075	3.044	0.082	0.075	3.044

Inverview Month (7)	0.216	0.169	1.383	0.216	0.169	1.383
Inverview Month (8)	0.101	0.091	2.658	0.101	0.091	2.658
Inverview Month (9)	0.009	0.009	10.660	0.009	0.009	10.660
Inverview Month (10)	0.006	0.006	12.660	0.006	0.006	12.660
Inverview Month (11)	0.002	0.002	20.120	0.002	0.002	20.100
Strata of residence (1)	0.139	0.119	2.093	0.139	0.119	2.094
Strata of residence (2)	0.549	0.248	-0.197	0.549	0.248	-0.197
Strata of residence (3)	0.137	0.119	2.108	0.137	0.119	2.108
Strata of residence (4)	0.009	0.009	10.660	0.009	0.009	10.660
Strata of residence (5)	0.165	0.138	1.801	0.165	0.138	1.801

Table 16: Estimates of the vouchers program on the number of hours the candidate is working per week

The table reports the treatment effects for the full sample in the effect of vouchers program on the time winners are working per week. The covariates include: type of survey, city of the survey, access to phone, age, gender, year of application, month of the interview, strata of residence. The summary of statistics reflects the observations before and after the entropy balancing for the board independence study. Treatment group observations are candidates who win the lottery. Otherwise are control group observations. The EB calculation is explained in Method section. Data are hand collected by Angrist et al (2002), and they can be obtained from Joshua Angrist's website at MIT. The estimation process is provided in the method section. Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels. All the units in the table are in percentile. The real number of observations might have been decreased for specific tests because of missing values, perfect predictions, or multi-collinearity problems.

Estimates of the vouchers program on the No of working hours for candidate per week

Dependent Variable: No of working hours per week	OLS method (natural experiment)	Direct matching combined method
Independent Variables:	Coefficient	Coefficient
Dummy(winner)	-0.87*	-0.917*
Type of survey	-3.667***	-4.082***
City of survey	-2.123*	-1.875

Access to phone	1.828	2.300
Age	1.938***	2.018***
Gender	3.811***	3.709***
Year 1995	-2.234	-2.201**
Year 1997	-0.349	0.000
Inverview Month (1)	4.657	4.711*
Inverview Month (2)	2.550	2.758
Inverview Month (3)	1.336	1.413
Inverview Month (4)	0.891	1.146
Inverview Month (5)	0.846	1.347
Inverview Month (6)	1.722	1.597
Inverview Month (7)	4.160	4.761***

Inverview Month (8)	5.037	5.455***
Inverview Month (9)	7.507	8.039**
Inverview Month (11)	0.910	1.423
Inverview Month (12)	3.014	7.087
Strata of residence (1)	5.323	3.008
Strata of residence (2)	4.354	2.094
Strata of residence (3)	4.282	2.025
Strata of residence (4)	0.729	0.000
Strata of residence (ms)	6.629	4.353**
Constant	-31.726***	-30.921***
Number of observations	1,577	1,577
Adjusted-R Squrare	0.109	0.1288

Table 17: Comparison between the direct matching combined method and the OLS method (in original natural experiment study)

The table reports the treatment effects for the full sample in the effect of vouchers program on the time winners are working per week. The covariates include: type of survey, city of the survey, access to phone, age, gender, year of application, month of the interview, strata of residence. The summary of statistics reflects the observations before and after the entropy balancing for the board independence study. Treatment group observations are candidates who win the lottery. Otherwise are control group observations. The EB calculation is explained in Method section. Data are hand collected by Angrist et al (2002), and they can be obtained from Joshua Angrist's website at MIT. The bias is the difference between estimation result on treatment effect from different methods and the benchmark. The MSE are based on 1000 bootstrap simulations (with replacement). In each simulation, we use the direct matching combined method and OLS methods to estimate the effect of program on working hours. All of the control variables are the same between the 2 methods and are explained in the experiment design section.

Comparison between the direct matching combined method with the OLS method (in original natural experiment study)

Benchmark Settings	Direct matching combined method	OLS Method
Original Treatment Effects	-0.92*	-0.87*
Bias	0.01	0.02
MSE	0.27	0.28

No of simulations	1,000	1,000

Table 18: Comparison of the direct matching combined method and PSM combined method

Table reports the results of the comparison of the direct matching combined method and PSM combined method. The propensity scores matching methods follow Heckman, Ichimura, and Todd (1998) and Becker and Ichino (2002). All of the benchmarks (bias, MSE and confidence intervals) to compare the two methods are based on 1000 simulations (with replacement). In each simulation, we use the direct matching combined method and PSM combined method to estimate the treatment effect of vouchers program on number of hours per candidate is working every week. All of the control variables are the same as in method sections. The definitions of the treatment and control groups and the estimation process are provided in the method section. The definitions of each variable are presented in Appendix.

Comparison of the direct matching combined method and PSM combined method

Benchmark Settings	Direct matching combined method	PSM combined method
Original Treatment Effects	-0.917*	-0.937*
Bias	0.001	0.005
MSE	0.295	0.305
No of simulations	1000	1000

Table 19: Comparison between Direct matching combined method and various PSMs combined method

The table reports the simulation results for the effect of financial vouchers on the number of hours a student may work per week. The propensity score matching methods follow Heckman, Ichimura, and Todd. (1998) and Becker and Ichino (2002). All of the benchmarks (bias, MSE and confidence intervals) to compare the two methods are based on 1000 simulations (with replacement). In each simulation, we use the direct matching combined method and PSM combined method to estimate the treatment effect of vouchers program on number of hours per candidate is working every week. All of the control variables are the same as method section. The definitions of the treatment and control groups and the estimation process are provided in the method section. The real number of simulations might be decreased because of no solutions for weights in a specific EB simulation.

Comparison between direct matching combined method and various PSMs combined method

Benchmark Settings	Direct matching combined method	Local Linear Regression Matching combined method	Radius Matching combined method	Nearest Matching combined method
Original Treatment Effects	-0.917	-1.035	-0.889	-0.926
Bias	0.010	0.082	0.017	0.039
MSE	0.278	0.955	0.284	0.292
No of simulations	1000	1000	1000	1000

Table 20: Hotelling's T-squared generalized means test and Rubins' B test

The table reports the Hotelling's T-squared generalized means test and Rubins' B test. In Hotelling's T-squared generalized means test evaluates whether a set of means equal between two groups, the Null Hypothesis is vectors of means are equal for the two groups. Rubins’ B indicates the standardized difference of the means of the linear index of the propensity score in the treatment and control groups. Rubin (2001) demonstrates that B is less than 25 for the compared sample indicates a sufficiently balanced compared group. The asterisk indicates the compared groups are sufficiently imbalanced.

Hotelling's T-squared generalized means test and Rubins' B test

Statistics of the test	Hotelling's T-squared generalized means test (Null Hypothesis:Vectors of means are equal for the two groups)				Rubin's B
Raw Data	F(23,1553):	1.568	Prob> F(23,1553):	0.042	30.4*
Data After indirect matching	F(23,1551)	0.165	Prob> F(23,1551):	1	9.9
Data After direct matching	F(23,1553):	0	Prob> F(23,1553):	1	0.2

Table 21: Balance checking for covariates between treatment and control groups

The sample includes 11204 observations from 1983 to 1990, 6102 of them are women, 5102 of them are men. The control variables include: Indicator of high school graduate, indicator of black people, indicator of Hispanic, indicator of Married, dummy for age group, dummy for unemployment. The right "Difference" column provides the mean difference between treatment and control groups. Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels according to the T statistics.

Panel A: Balance Checking for covariates between treatment and control groups (Men)

Variable	Control Group					Treatment Group					Difference
	Obs	Mean	Std.Dev.	Min	Max	Obs	Mean	Std.Dev	Min	Max	
hsorged	1703.000	0.694	0.445	0.000	1.000	3399.000	0.693	0.447	0.000	1.000	0.002
black	1703.000	0.254	0.435	0.000	1.000	3399.000	0.254	0.435	0.000	1.000	-0.001
hispanic	1703.000	0.093	0.291	0.000	1.000	3399.000	0.099	0.299	0.000	1.000	-0.006
married	1703.000	0.338	0.462	0.000	1.000	3399.000	0.360	0.469	0.000	1.000	-0.023
wkless13	1703.000	0.395	0.467	0.000	1.000	3399.000	0.403	0.468	0.000	1.000	-0.008
class_tr	1703.000	0.189	0.392	0.000	1.000	3399.000	0.209	0.407	0.000	1.000	-0.02*
ojt_jsa	1703.000	0.503	0.500	0.000	1.000	3399.000	0.504	0.500	0.000	1.000	-0.001

Hotelling's T-squared generalized means test

H0:	Vectors of means are equal for the two groups	F(13,5088):	0.8739	Prob>F(13,5088):	0.5806

Panel B: Balance Checking for covariates between treatment and control groups (Women)

Variable	Control Group					Treatment Group					Difference
	Obs	Mean	Std.Dev.	Min	Max	Obs	Mean	Std. Dev	Min	Max	
hsorged	2014.000	0.704	0.442	0.000	1.000	4088.000	0.729	0.431	0.000	1.000	-0.024*
black	2014.000	0.257	0.437	0.000	1.000	4088.000	0.268	0.443	0.000	1.000	-0.011
hispanic	2014.000	0.125	0.330	0.000	1.000	4088.000	0.117	0.321	0.000	1.000	0.008
married	2014.000	0.208	0.388	0.000	1.000	4088.000	0.225	0.400	0.000	1.000	-0.017
wkless13	2014.000	0.519	0.472	0.000	1.000	4088.000	0.518	0.468	0.000	1.000	0.001
class_tr	2014.000	0.387	0.487	0.000	1.000	4088.000	0.382	0.486	0.000	1.000	0.005
ojt_jsa	2014.000	0.382	0.486	0.000	1.000	4088.000	0.371	0.483	0.000	1.000	0.011

Hotelling's T-squared generalized means test

H0:	Vectors of means are equal for the two groups	F(14,6087):	1.0578	Prob>F(13,5088):	0.3915

Table 22: Estimation results on treatment effect with OLS, direct matching combined method, and PSM combined method

The sample includes 11204 observations from 1983 to 1990, 6102 of them are women, 5102 of them are men. The control variables include: Indicator of high school graduate, indicator of black people, indicator of Hispanic, indicator of Married, dummy for age group, dummy for unemployment. Less variables indicates that the model only include indicator of JTPA assignment. More variables indicate that the model include both indicator of JTPA and all the other control variables. Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels according to the T statistics.

Panel A: Estimates of the treatment effect on Men's earning

Dependent Variable:30 month earning	Ordinary Least Squares method	Ordinary Least Squares method	Direct matching combined method	Direct matching combined method	PSM combined method	PSM combined method
Independent Variables:	Coefficient (Less Variables)	Coefficient (More Variables)	Coefficient (Less Variables)	Coefficient (More variables)	Coefficient (Less Variables)	Coefficient (More Variables)
Dummy (treatment)	1116.585*	969.921*	1050.802*	1048.911*	1196.628**	1059.043**

Panel B: Estimates of the treatment effect on Women's earning

Dependent Variable:30 month earning	Ordinary Least Squares method	Ordinary Least Squares method	Direct matching combined method	Direct matching combined method	PSM combined method	PSM combined method
Independent Variables:	Coefficient (Less Variables)	Coefficient (More Variables)	Coefficient (Less Variables)	Coefficient (More variables)	Coefficient (Less Variables)	Coefficient (More Variables)
Dummy (treatment)	1242.557***	1139.456***	1205.382***	1204.045***	1165.63***	1123.649***

Table 23: Estimates of the NSW training program on increased income.

The data are collected from Dehejia and Wahba (1999). The dependent variable is defined as increased earnings in the year 1978, treat is defined as a dummy variable that equals 1 if the observation is receiving training program from NSWDP, otherwise in the control group. Controls include (1) age, (2) years being in the school: educ, (3) black: dummy for black or not, (4) hisp: dummy for Hispanic or not, (5), married: dummy for married or not, (6): nodegree: dummy for not obtained diploma or not, (7) re74: increased earnings in the year 1974, (8)re75: increased earnings in the year 1975, (9) u74: dummy for not having a job in the year 1974 nor not, (10) u75: dummy for not having a job in the year 1975 nor not. Additional details about the data and program can be obtained from Dehejia and Wahba (1999). The estimation process is provided in the method section. Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels. The real number of observations might have been decreased for specific tests because of missing values, perfect predictions, or multi-collinearity problems.

Estimates of the training program on raised income

Dependent Variable:re78	Ordinary Least Squares method	Direct matching combined method	PSM Combined Method
Independent Variables:	Coefficient	Coefficient	Coefficient
Treat	1672.426***	1776.685***	1669.661**
Control Variables	Yes	Yes	Yes
constant	Yes	Yes	Yes

Figure 1: Frequency distributions of propensity scores

The figure reports the frequency distribution of propensity scores estimated by logit regression method. The propensity score matching methods follow Heckman, Ichimura, and Todd. (1998) and Becker and Ichino (2002). The dependent variable in logit model is a Dummy (Treatment). The distribution in red colour is for treatment group and the distribution in blue colour is for control group.

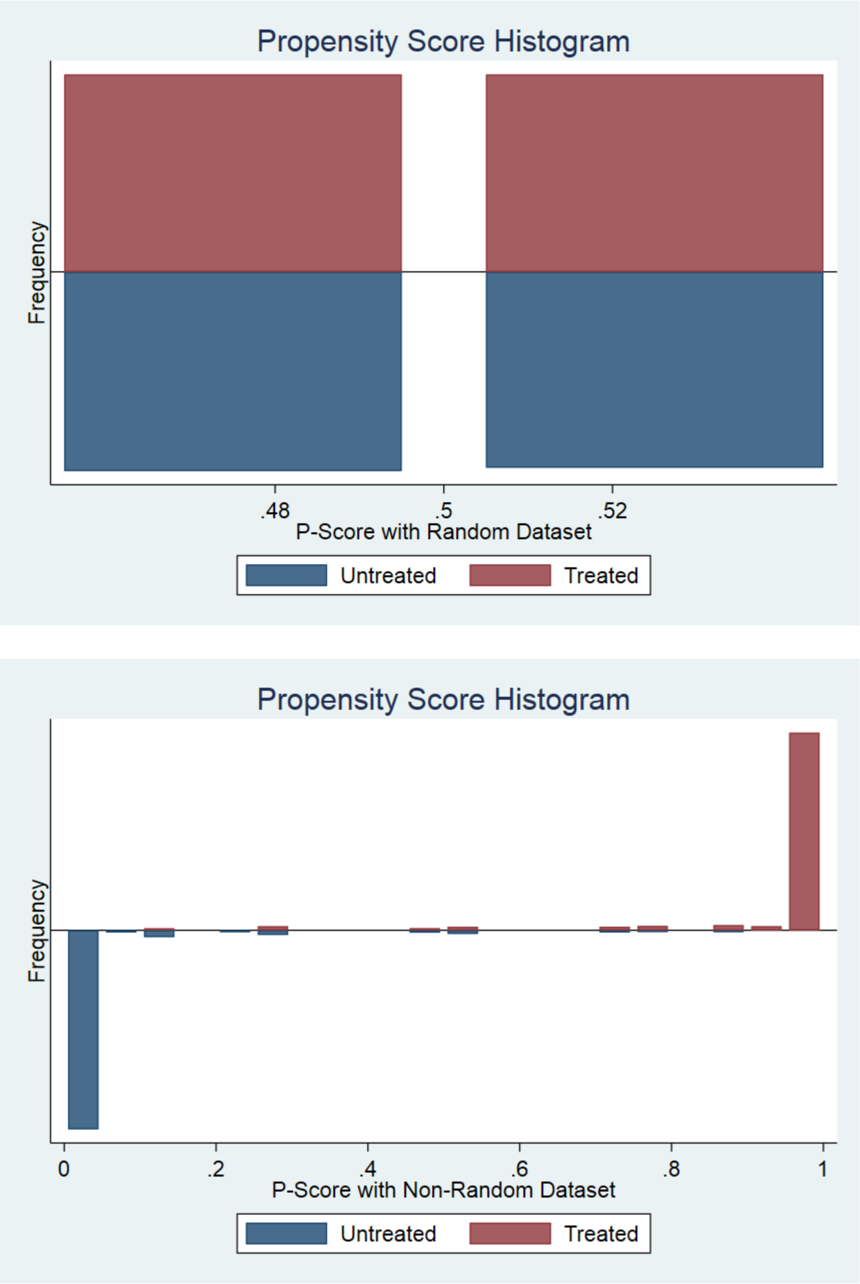


Figure 2 Density function with and without weighting operation process

The figure reports an example of density function for a covariate with and without weighting operation process. The left column reports the raw density function without weighting operation. The right column reports the covariate density function after weighting operation. The red curve represents the density function of the covariate in treatment group. The blue curve represents the density function of the covariate in control group.

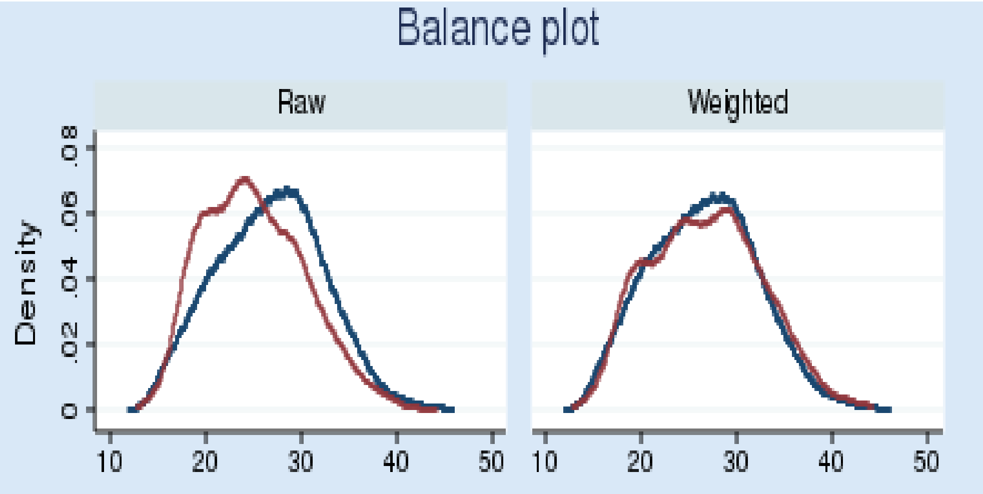


Figure 3 Changes in total board independence ratio from 2003 to 2005 of sample firms

The sample consists all the observations we collect from institutional Shareholder Services (formerly RiskMetrics). The board information including independence classification is obtained from Institutional Shareholder Services (formerly RiskMetrics). The compliance ratio shows the percentage of the sample firms which have more than 50% independent directors in the board.

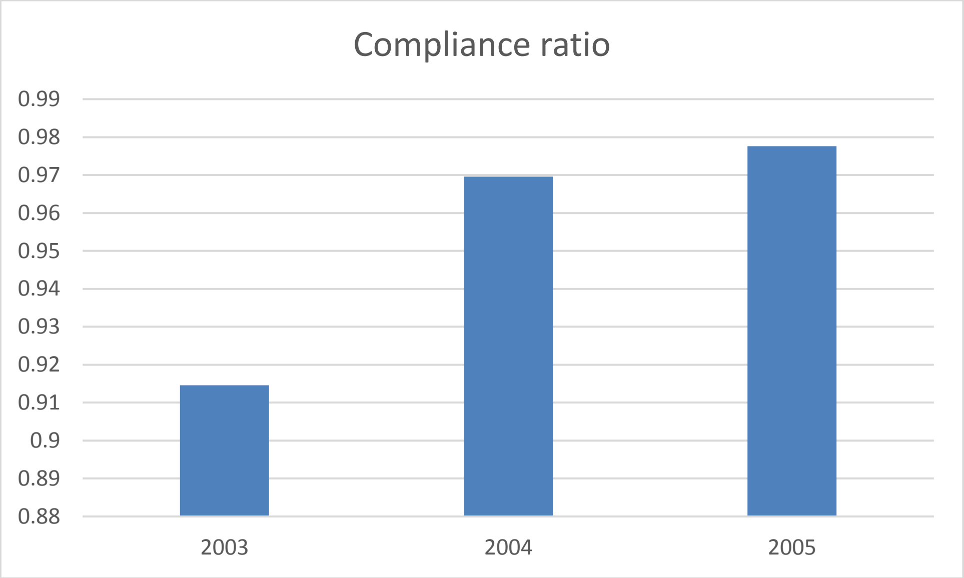


Figure 4. Density for weights obtained from direct matching (EB) calculation

The figure reports the density of weights obtained from direct matching (EB) calculation. The sample includes 807 firms in the period 2000-2005. The data are collected according to the instructions from Chhaochharia and Grinstein (2009) and can be obtained from the Supplementary Appendix on the Journal of Finance website. The EB process is explained in the methodology section.

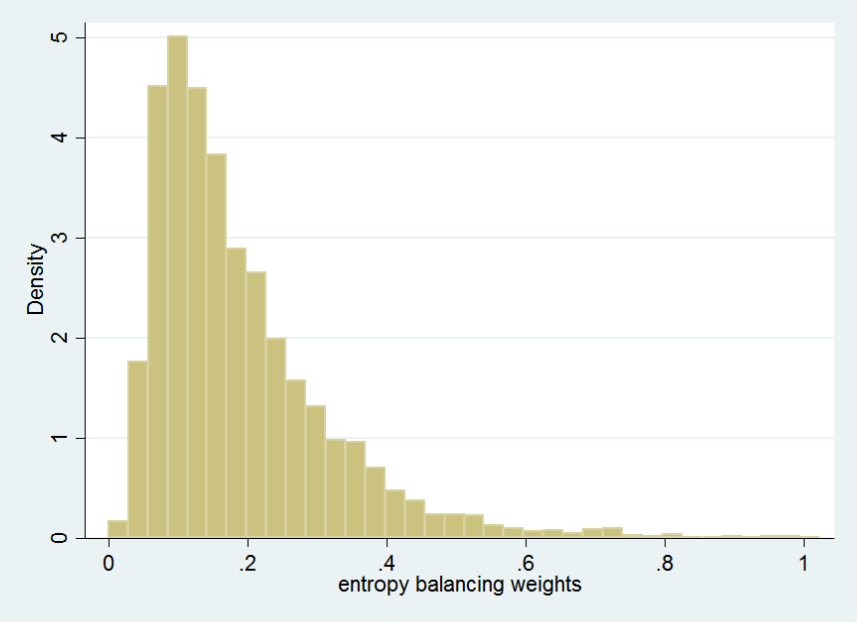


Figure 5 Balanced condition in Propensity Score Method Study

The figure reports the frequency distribution of propensity scores estimated by a logit model in the board independence study. The propensity scores matching methods follow Heckman, Ichimura, and Todd. (1998) and Becker and Ichino (2002). The dummy variable for treatment equals 1 if the firms are forced to increase board independence level in 2002, and dummy of treatment equals 0 if firms have satisfied regulation before 2002 and thus are not affected by the policy.To control for firm characteristics, we use (1) Sales (the natural log of company sales), (2) ROA (the natural log of one plus net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1)), (3) Stock Returns (the natural log of the annual gross stock return (dividend reinvested), measured in year (t−1)), (4) Tenure (the number of years the CEO served with the firm), and (5) Industry (Fama-French 48 Industry factor).

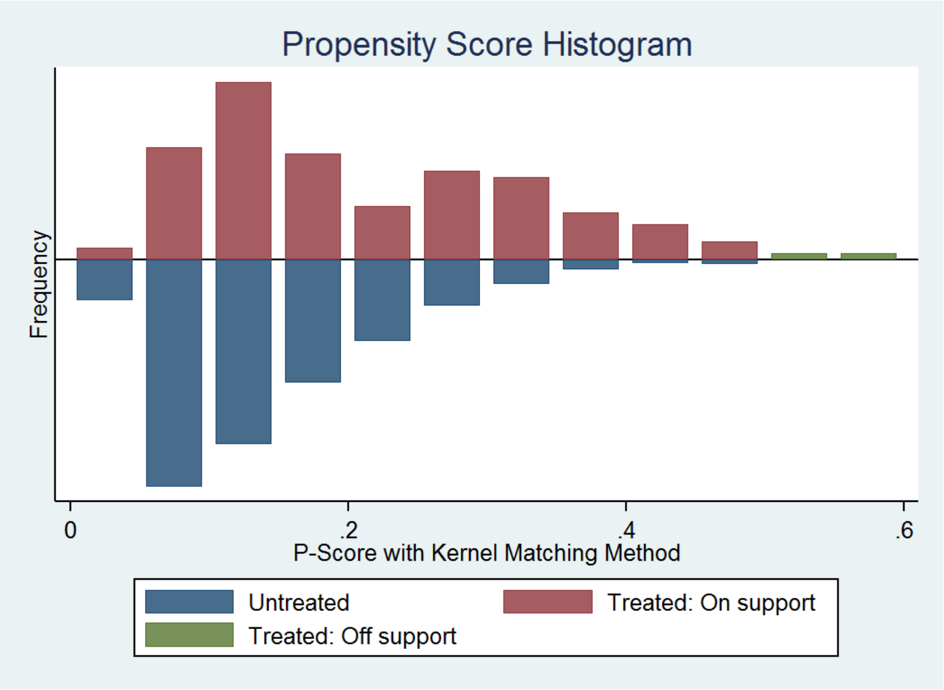


Figure 6 Example of Propensity Score Method Study which satisfies the Balanced condition

The figure reports the example of PSM that satisfies the balanced condition between compared groups (Kirmani, Holmes and Muir (2016)). The bottom section indicates the distribution for control group. The upper section shows the distribution for treatment group. The propensity score matching methods follow Heckman, Ichimura, and Todd. (1998) and Becker and Ichino (2002).

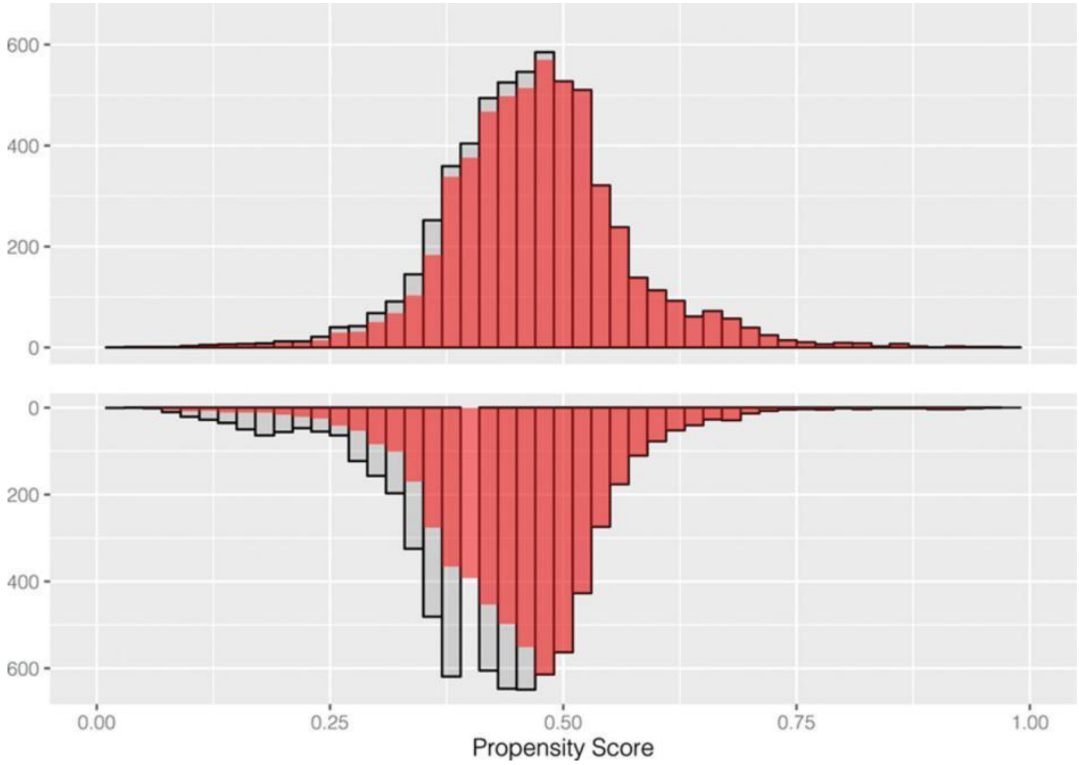


Figure 7 Standardized percentage bias across covariates before and after direct matching (EB) calculation

The figure reports the standardized percentage bias on the sample covariates before and after EB matching for treatment and control groups. The formula can be obtained from Rosenbaum and Rubin (1985). Sales is the natural log of company sales. ROA is the natural log of one plus net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1). Stock Returns is the natural log of the annual gross stock return (dividend reinvested), measured in year (t−1). ltenure is the number of years the CEO served with the firm.

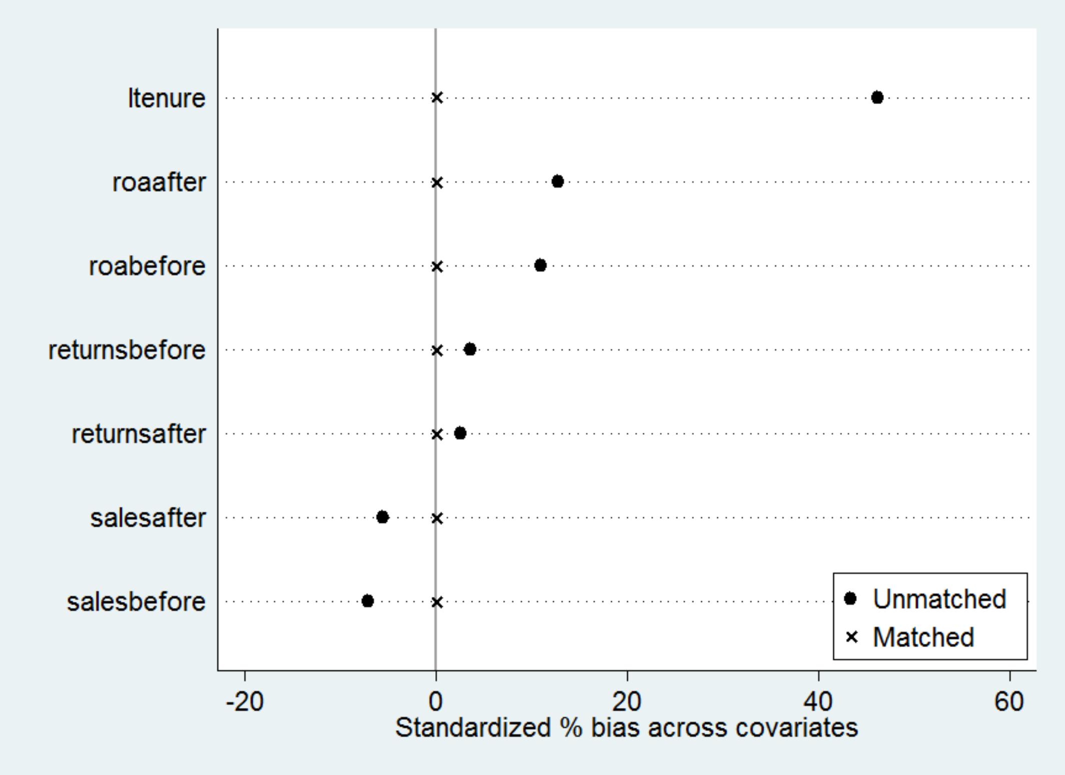


Figure 8: Standardized percentage bias across covariates before and after indirect (PSM) matching calculation

The figure reports the standardized percentage bias on the sample covariates before and after PSM matching for treatment and control groups. The formula can be obtained from Rosenbaum and Rubin (1985). Sales is the natural log of company sales. ROA is the natural log of one plus net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1). Stock Returns is the natural log of the annual gross stock return (dividend reinvested), measured in year (t−1). ltenure is the number of years the CEO served with the firm.

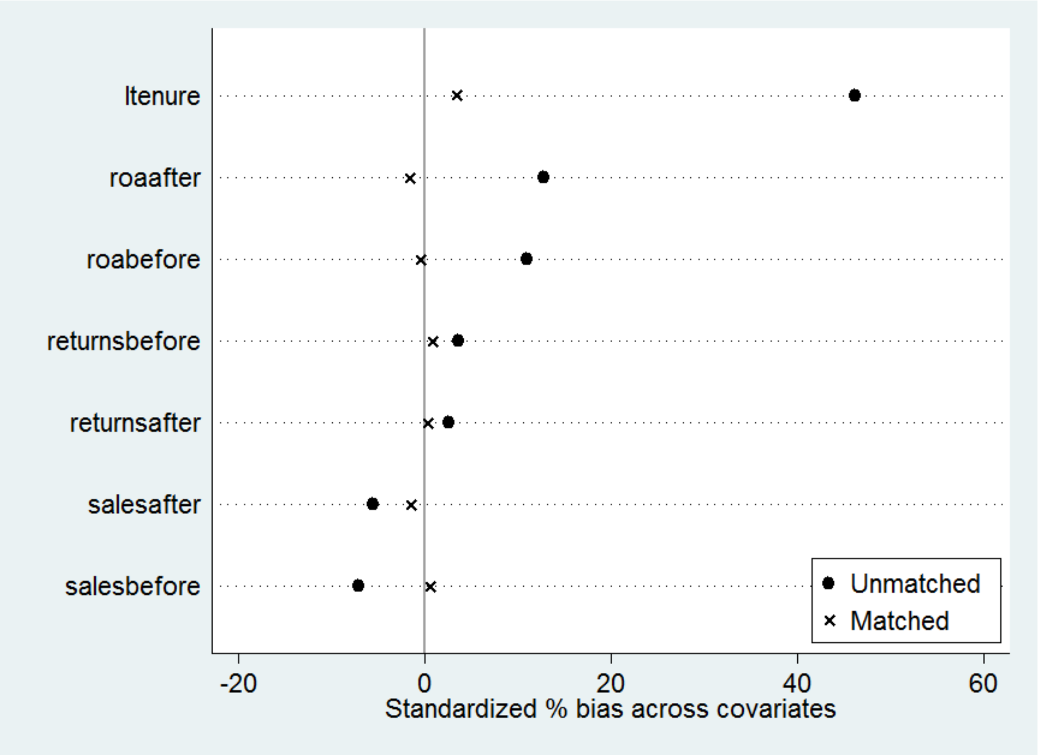


Figure 9: Kernel Density estimate for weights obtained from direct matching (EB) calculation

The figure reports the density of weights obtained from the direct matching (EB) calculation. The covariates include: type of survey, city of the survey, access to phone, age, gender, year of application, month of the interview, strata of residence. The summary of statistics reflects the observations before and after the entropy balancing for the board independence study. Treatment group observations are candidates who win the lottery. Otherwise are control group observations. The direct matching (EB) calculation is explained in Method section. Data are hand collected by Angrist et al (2002), and they can be obtained from Joshua Angrist's website at MIT.

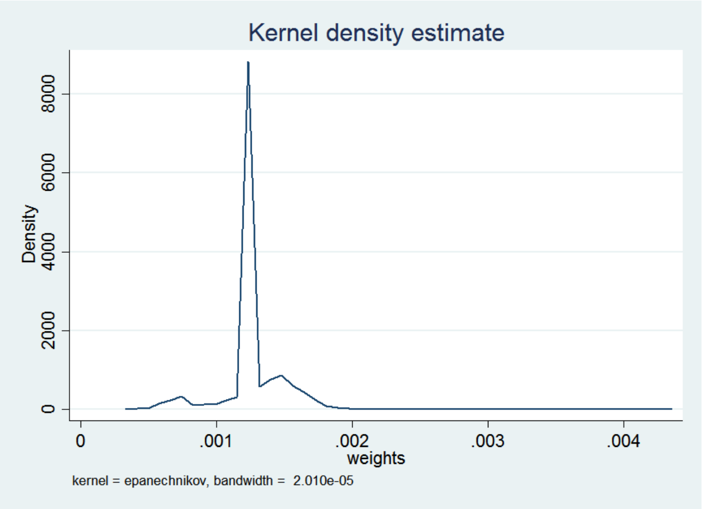


Figure 10: Frequency distributions of propensity scores

The figure reports the frequency distribution of propensity scores estimated by logit regression method. The propensity score matching methods follow Heckman, Ichimura, and Todd. (1998) and Becker and Ichino (2002). The dependent variable in logit model is a Dummy (Treatment). Definition for treatment and control groups is provided in the data and sample section. The covariates include: type of survey, city of the survey, access to phone, age, gender, year of application, month of the interview, strata of residence. Data are hand collected by Angrist et al (2002), and they can be obtained from Joshua Angrist's website at MIT. The distribution in red color is for treatment group and the distribution in blue color is for control group.

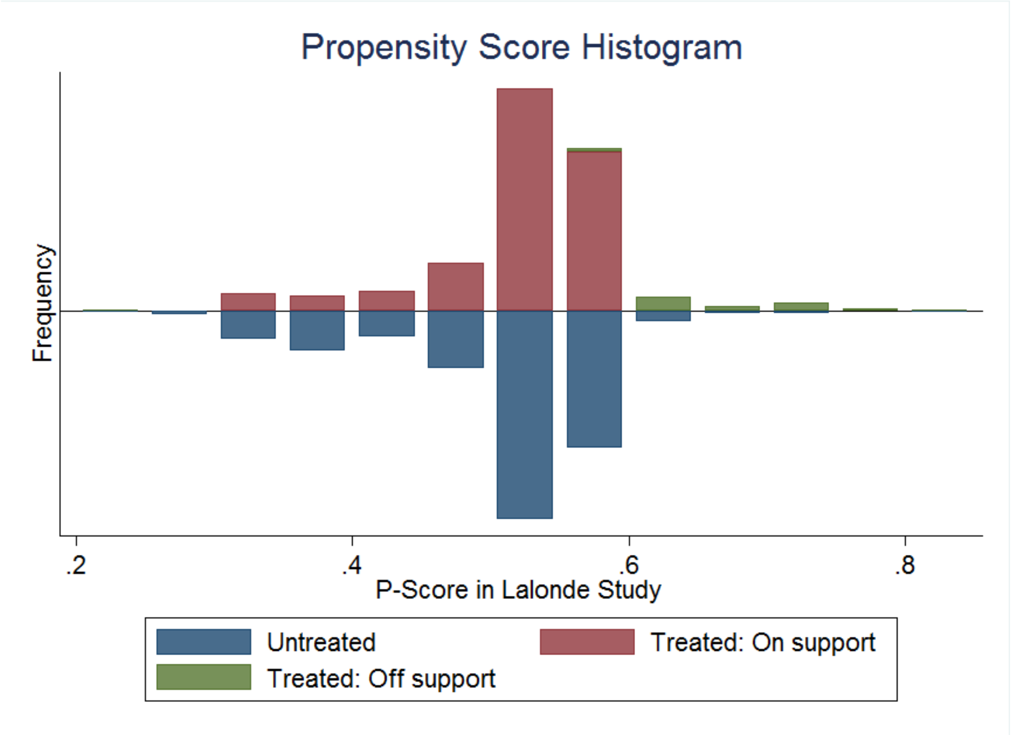


Figure 11: Standardized percentage bias across covariates before and after direct matching (EB) calculation

The figure reports the standardized percentage bias on the sample covariates before and after EB matching for treatment and control groups. The formula can be obtained from Rosenbaum and Rubin (1985). The covariates include: type of survey, city of the survey, access to phone, age, gender, year of application, month of the interview, strata of residence. Data are hand collected by Angrist et al (2002), and they can be obtained from Joshua Angrist's website at MIT.

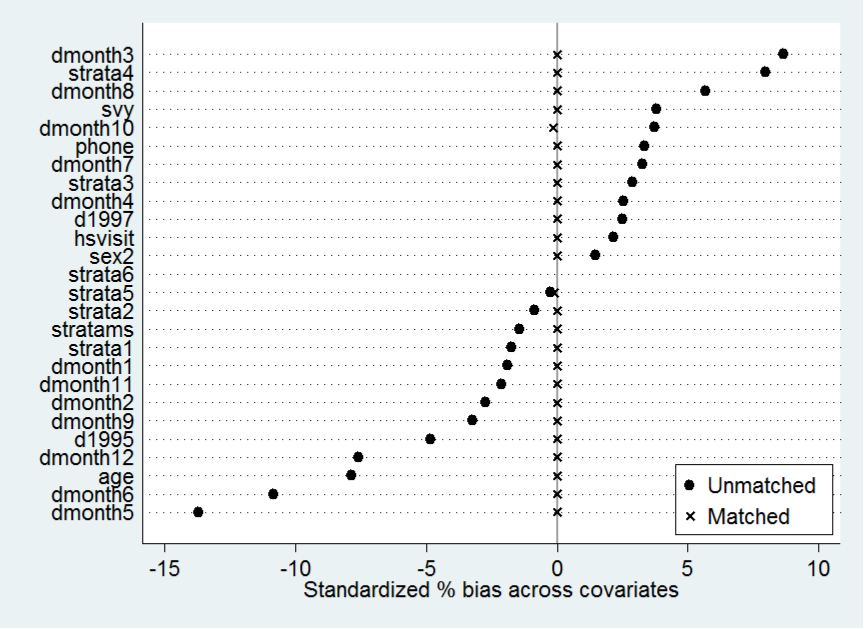


Figure 12: Standardized percentage bias across covariates before and after PSM calculation

The figure reports the standardized percentage bias on the sample covariates before and after PSM matching for treatment and control groups. The formula can be obtained from Rosenbaum and Rubin (1985). The covariates include: type of survey, city of the survey, access to phone, age, gender, year of application, month of the interview, strata of residence. Data are hand collected by Angrist et al (2002), and they can be obtained from Joshua Angrist's website at MIT.

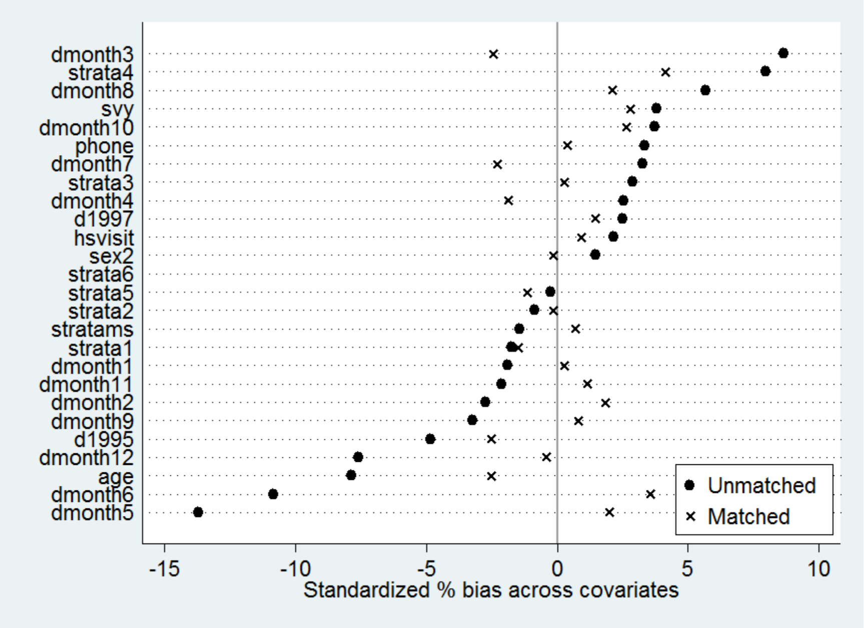


Figure 13: Density for weights obtained from direct matching (EB) calculation for men’s group

The figure reports the density of weights obtained from EB calculation. The sample includes 11204 observations from 1983 to 1990, 6102 of them are women, 5102 of them are men. The control variables include: Indicator of high school graduate, indicator of black people, indicator of Hispanic, indicator of Married, dummy for age group, dummy for unemployment. The right “Difference” column provides the mean difference between treatment and control groups.

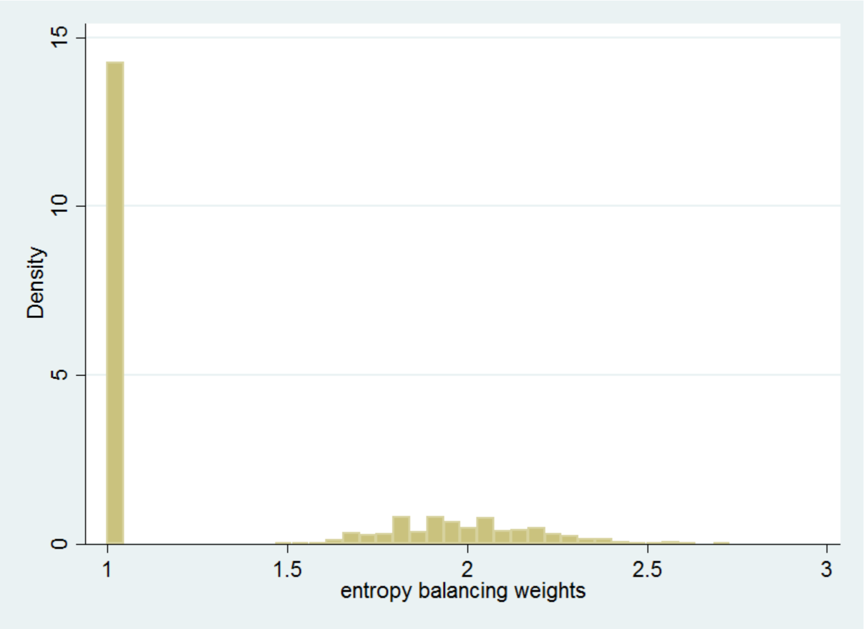
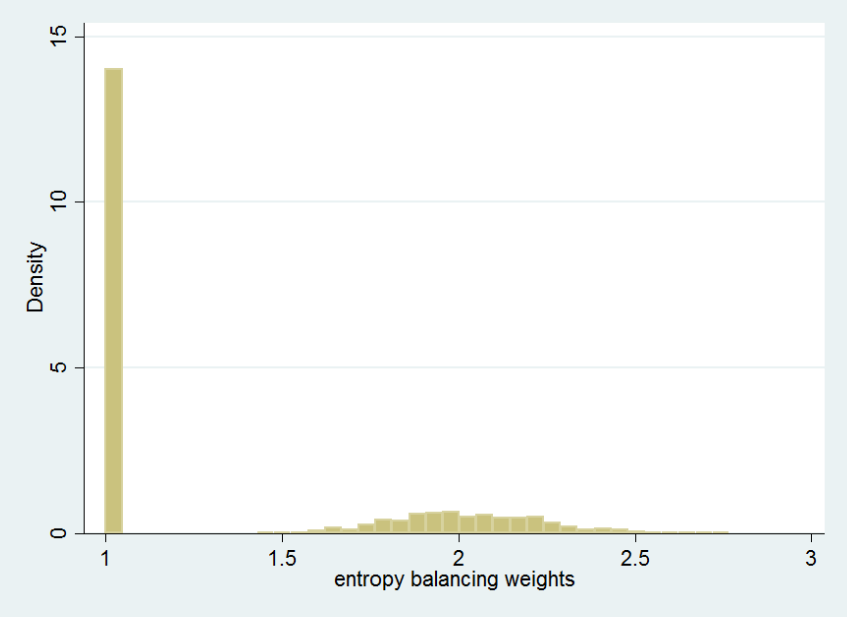


Figure 14: Density for weights obtained from direct matching (EB) calculation for women's group

The figure reports the density of weights obtained from EB calculation. The sample includes 11204 observations from 1983 to 1990, 6102 of them are women, 5102 of them are men. The control variables include: Indicator of high school graduate, indicator of black people, indicator of Hispanic, indicator of Married, dummy for age group, dummy for unemployment. The right "Difference" column provides the mean difference between treatment and control groups.



Internet Appendices

Appendix A: Mechanism of estimation of problems in a non-random environment and the framework for EB methodology

A.1 Bias issues induced by OLS if intervention of treatment is not random

In this subsection, we analyze the estimation bias of the OLS technique in the natural experiment design when covariates are not balanced between the groups compared. More specifically, suppose a model we employ in the natural experiment study is described as follows:

$$z_i = \theta_0 + \theta_d d_i + \theta_1 y_{1i} + \varepsilon_i \tag{A1}$$

where θ_0, θ_d, and θ_1 are the parameters which are unknown, d_i is the dummy variable capturing the observations that receive the intervention (treatment) or not, and ε_i is the non-observed disturbance errors.[31] Following equation (A1), a random sample with n units can be obtained, $\{(d_i, y_{1i}, z_{i,}): i=1,2,....,n\}$. We suppose that in the random sample, all the regressors are not constant and that they do not have perfect linear relationships. Suppose that the intervention d_i is not randomly assigned to each unit in the sample, such that y_{1i} are distinct between treatment and control groups. We assume that an unknown factor (i.e.,y_{2i}), which not only induces the non-randomness problem of intervention d_i, but also impacts z_i is not included in equation (1).[32] In this case, we miss the regressor y_{2i}, which should be in the true (population) model.

$$z_i = \hat{\theta}_0 + \hat{\theta}_d d_i + \hat{\theta}_1 y_{1i} + \hat{\theta}_2 y_{2i} + \eta_i \tag{A2}$$

Suppose we are interested in $\hat{\theta}_d$, the treatment effect of d_i on z_i. For example, z_i is the exam score, $d_i$is an indicator capturing whether a student receives a challenging elective training program or not, y_{1i} is the number of hours a student studies before the exam, and y_{2i} is the learning ability which will determine whether or not the student is willing to select the training program and will simultaneously impact the student's exam score. To obtain an unbiased estimator of $\hat{\theta}_d$, we should regress z_i on d_i, y_{1i}, and y_{2i}, but due to unawareness or unavailability, we only run a regression of z_i on d_i and y_{1i}, that is, equation (A1).

[31] For simplification purpose, we only include two independent variables, one can increase the number of regressors in the real applications and the conclusion will not change.

[32] If the randomness assumption does not hold, some reason must exist so that the candidates in the treatment group will receive this intervention for a reason and the candidates in the control group will not receive intervention for a reason. Suppose we do not know the reason and it is missed in the model. We assume that the unknown reason is factor y_2, which is in equation (2).

Consider equation (A1) again. The parameter θ_d estimated by OLS is as follows[33]:

$$\theta_d = \frac{\sum_{i=1}^{n} \gamma_{id} z_i}{\sum_{i=1}^{n} \gamma_{id}^2} \tag{A3}$$

where γ_{id} are the residuals from the regression of d_i on y_{1i}. To calculate the numerator in equation (A3), we can use equation (A2) to plug into (A3). Owing to the properties of residuals in an OLS setting, γ_{id} should have a mean of zero and no correlations with y_{1i}. Likewise, η_i has a zero mean and is not correlated with d_i, y_{1i}, and y_{2i}. As γ_{id} is just the linear combination of d_i and y_{1i}, it turns out to be that η_i has no correlation with γ_{id}.

Hence, we have the following:

$$\sum_{i=1}^{n} \gamma_{id} z_i = \sum_{i=1}^{n} \gamma_{id}\left(\widehat{\theta_0} + \widehat{\theta_d} d_i + \widehat{\theta_1} y_{1i} + \widehat{\theta_2} y_{2i} + \eta\right) = 0 + \widehat{\theta_d} \sum_{i=1}^{n} \gamma_{id} d_i + 0 +$$
$$\widehat{\theta_2} \sum_{i=1}^{n} \gamma_{id}\, y_{2i} + 0 = \widehat{\theta_d} \sum_{i=1}^{n} \gamma_{id} d_i + \widehat{\theta_2} \sum_{i=1}^{n} \gamma_{id}\, y_{2i} \tag{A4}$$

As d_i can be calculated by its fitted value plus the residuals as follows: $\widehat{d_i} + \gamma_{id}$

$$\sum_{i=1}^{n} \gamma_{id} d_i = \sum_{i=1}^{n} \gamma_{id}(\widehat{d_i} + \gamma_{id}) = \sum_{i=1}^{n} \widehat{\gamma_{id}} d_i + \sum_{i=1}^{n} \gamma_{id}^2 = 0 + \sum_{i=1}^{n} \gamma_{id}^2 = \sum_{i=1}^{n} \gamma_{id}^2 \tag{A5}$$

When we plug (A4) and (A5) back to (A3), the parameter of θd can be obtained as follows:

$$\theta_d = \frac{\sum_{i=1}^{n} \gamma_{id} z_i}{\sum_{i=1}^{n} \gamma_{id}^2} = \widehat{\theta_d} + \widehat{\theta_2} \frac{\sum_{i=1}^{n} \gamma_{id} y_{i2}}{\sum_{i=1}^{n} \gamma_{id}^2} = \widehat{\theta_d} + \widehat{\theta_2} \delta_1 \tag{A6}$$

where δ_1 is the slope for d_i if we regress y_{2i} on d_i and y_{1i}. Hence, we have the following:

$$\mathrm{E}(\theta_d) = \mathrm{E}\left(\widehat{\theta_d} + \widehat{\theta_2}\delta_1\right) = \mathrm{E}\left(\widehat{\theta_d}\right) + \mathrm{E}\left(\widehat{\theta_2}\right)\delta_1 = \widetilde{\theta_d} + \widetilde{\theta_2}\delta_1 \tag{A7}$$

This infers that the bias of θ_d is

$$\mathrm{Bias}(\theta_d) = \mathrm{E}(\theta_d) - \widetilde{\theta_d} = \widetilde{\theta_2}\delta_1 \tag{A8}$$

Equation (A8) shows that as long as the missed factor y_2 exits and y_2 is correlated to the intervention indicator d ($\delta_1 \neq 0$) as well as the dependent variable z ($\widetilde{\theta_2} \neq 0$), the bias exists and its magnitude equals $\widetilde{\theta_2}\delta_1$.

1.2 Estimation problems for the variance of the OLS estimators if intervention of the treatment is not random

Consider again that the intervention d is not random and if a factor y_1 impacts the treatment indicator (d) and the dependent variable (z) but is neglected. We are interested to produce the correct variance for the estimator $\widehat{\theta_d}$ from the true model

[33] More details on the derive, see Wooldridge (2009).

$$z_i = \widehat{\theta_0} + \widehat{\theta_d} d_i + \widehat{\theta_1} y_{1_i} + \varepsilon_i \tag{A9}$$

Suppose that the error term ε_i has a constant variance σ^2 conditional on the values of regressors, that is

$$\text{Var}(\varepsilon_i | d_i, y_{1i}) = \sigma^2 \tag{A10}$$

Similar to equation (A3), the parameter, θ_d, estimated by OLS from equation (A9) is

$$\theta_d = \frac{\sum_{i=1}^{n} \gamma_{id} z_i}{\sum_{i=1}^{n} \gamma_{id}^2} \tag{A11}$$

where γ_{id} are the residuals from the regression of d_i on y_{1_i}. Then, with the OLS properties on residuals, we have

$$\textstyle\sum_{i=1}^{n} \gamma_{id} = 0 \, , \sum_{i=1}^{n} \gamma_{id}\, y_{1i} = 0 \tag{A12}$$

When we plug (A5), (A9), and (A12) back to (A11), then we have

$$\theta_d = \widehat{\theta_d} + \frac{\sum_{i=1}^{n} \gamma_{id} \varepsilon_i}{\sum_{i=1}^{n} \gamma_{id}^2} \tag{A13}$$

Under the random sampling condition, given the regressors d and y_1 (suppose we can call them Y), ε should be independent of Y. Given the condition of explanatory variables Y, γ_{id} are not random. Based on equation (A13), we have

$$\text{Var}(\theta_d | Y) = \frac{\sum_{i=1}^{n} \gamma_{id}^2 Var(\varepsilon_i | Y)}{\left(\sum_{i=1}^{n} \gamma_{id}^2\right)^2} = \frac{\sum_{i=1}^{n} \gamma_{id}^2 \sigma^2}{(\sum_{i=1}^{n} \gamma_{id}^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} \gamma_{id}^2} \tag{A14}$$

As γ_{id} by definition are the residuals obtained from the regression of d$_i$ on y$_{i1}$, we have

$$\textstyle\sum_{i=1}^{n} \gamma_{id}^2 = SSR_d = TSS_d \frac{SSR_d}{TSS_d} = TSS_d \left(1 - 1 + \frac{SSR_d}{TSS_d}\right) = TSS_d(1 - R_d^2) \tag{A15}$$

Hence,

$$\text{Var}(\theta_d) = \frac{\sigma^2}{TSS_d(1 - R_d^2)} \tag{A16}$$

As d_i is not randomly assigned to each observation owing to an unobserved factor y_{i1} which impacts both d_i and z_i, we mistakenly obtain the variance of parameter of θ_d from the equation

$$z_i = \widetilde{\theta_0} + \widetilde{\theta_d} d_i + \varepsilon_i \tag{A17}$$

rather than the correct choice equation (A9). As such, we have

$$\text{Var}\left(\widetilde{\theta_d}\right) = \frac{\sigma^2}{TSS_d} \tag{A18}$$

By comparing (A16) and (A18),[34] we can infer that the true variance of the parameter θ_d can

[34] The derivation for equation (18) is very similar to the derivation for equation (16) when only one explanatory variable is included in the model, so that we do not repeat.

never be obtained correctly as long as the missed factor y_{1i} exits and have correlations with the intervention indicator d_i (such that R_d^2 is between 0 and 1).

As the OLS standard errors are calculated by the variance of the parameter (i.e., the incorrect value in equation (A18) rather than the correct value in equation (A16)), the outcome on the confidence interval as well as the t-statistic used to determine the significance level for the parameter is no longer valid. In doing so, the statistical inference in a non-random intervention case study can be misleading.

Appendix B: Committee chairmanship appointment as a natural experiment for the study of its effect on firm-level R&D investment

B.1 Design

Cohen, Coval, and Malloy (2011) use the changes in chairmanship in an influential congress committee to identify the causal effect of powerful politicians on firms' R&D investments[35]. To be appointed as a chairman of congressional committees, the candidate must be the most senior member of the party in the relevant congressional committee. Such a chairmanship appointment can occur if (1) the previous chairman in the committee resigns during his (or her) term, or (2) there are changes in the controlling party that has just won the new congressional election, which then induces the congressional committee to be reorganized. As (1) and (2) are challenging to forecast before the appointment, Cohen, Coval, and Malloy (2011) argue that such new "chairmanship appointments" are an exogenous event. In addition, they use the OLS method and find a negative relationship between the shocks from chairmanship appointment and firm-level R&D investments. They interpret this relationship as empirical support that a powerful chairmanship can cause a "crowd out" effect with respect to firm investment.

We follow the designs presented in Cohen, Coval, and Malloy (2011) and use the following regression model:[36]

[35] According to Cohen, Coval, and Malloy (2011), the top three influential congress committees include the Finance Committee, the Veterans Affairs Committee, and the Appropriations Committee. In this study, we only include the Finance Committee Study (Top 1 Committee study) as the relevant data are provided by Cohen, Coval, and Malloy (2011) on their website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607

[36] There is a debate over the effects of politician shocks. For example, Cohen, Coval, and Malloy (2011) find that powerful politicians will crowd out a firm's R&D, but Snyder and Welch (2015) argue that this decrease is caused by outliers. However, none of the two studies provide an efficient method to address the randomness assumption of the natural experiment problem rather than relying on the researchers' own knowledge to one or more special events (i.e., the oil crisis during 1980's) which provides us incentive to use the EB method to detect the problem more efficiently.

$$FirmR\&D_{it} = Shock\_Chairman_{it} + Controls_{it} + Firm\_FE_I + Year\_FE_t + \varepsilon_{it} \quad \text{(B1)}$$

where the dependent variable is defined as a firm's R&D expense (lagged 1 year) scaled by firm assets, Shock_Chairman is defined as a dummy variable that equals 1 if the firm is located in the state where a powerful congressional chairman holds office, but otherwise equals 0. The controls include (1) lagged Q (Q ratio in previous year), (2) cash flow (scaled by the asset in the previous year), and (3) lagged leverage (leverage ratio in the previous year). All the tests include firm-level fixed effect and year-fixed effects.[37]

B.2 Is the assignment to treatment really exogenous?

Thus far, our methodology framework assumes that the chairmanship appointment is randomly assigned to each state as assumed by Cohen, Coval, and Malloy (2011). To establish whether firms located in the state where a powerful committee chairman holds office are similar to firms without politician shock, we follow the same dataset as in Cohen, Coval, and Malloy (2011) and illustrate the comparison of firm characteristics between treatment and control groups in Table B1. We illustrate in the diagram that the randomness assumption is not convincing as the table shows that almost all firm characteristics (e.g., Lag Q, Cash flow) are significantly different between the groups compared. For example, the mean of the variable Lag Q for the treatment group is 3.087, which is 0.257 greater than that in the control group at the 1% significance level.[38]

If firms in the shocked group tend to be significantly different from those of firms in the control group, comparing the groups that are affected by the powerful politician with those who are not affected by the shock will potentially bias the results toward finding a causality relationship as the two groups are heterogeneous and not comparable. Thus, the evidence of the decrease in firms' R&D investment estimated by OLS (without a data preprocessing step) may not be attributable to the powerful politician, but rather to heterogeneous pre-conditions in each group.

Nevertheless, it is possible to find a plausible group of weights assigned to the control

[37] Additional details about the data and program can be obtained from the website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607.

[38] In this chapter, we report the aggregate level results. In an unreported table, we also compare the balanced condition results in each term of office when the powerful politician sits as a chairman of a congress committee and the results are similar.

Table B1: Balance checking for covariates in the study of the effect of a powerful chairman on firm's R&D

The data are collected according to Cohen, Coval, and Malloy (2011) and can be obtained from website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607. Controls include (1) lagged Q (Q Ratio in previous year), (2) cash flow (scaled by asset in previous year), (3) lagged leverage (leverage ratio in previous year), (4) Year. More details are explained in Appendix. The right "Difference" column provides the mean difference between the two groups. Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels according to the T statistics.

Balance Checking for covariates in the study of impact of powerful chairman on firm's R&D

Variable	Treatment Group					Control Group					Difference
	Obs	Mean	Std.Dev	Min	Max	Obs	Mean	Std.Dev	Min	Max	
Lag Q	2,173	3.087	4.803	-12.587	34.035	75,425	2.830	4.422	-12.951	52.984	-0.257***
Lag Leverage	2,173	0.349	0.251	0.006	0.988	75,425	0.347	0.247	0.004	0.988	-0.003
CashFlow_Scaled	2,173	-0.007	0.280	-1.370	0.539	75,425	0.016	0.259	-1.927	0.590	0.023***
Year	2,173	1,993.276	3.835	1,970	2,006	75,425	1,991.236	10.503	1,967	2,008	-2.04***

group such that, after controlling for them with the EB method, the assignment of politician shock can be shown to be orthogonal to the background characteristics.

The results after the EB calculation for Top 1 committee study (defined in footnote 27) are shown in Table B2. The table shows that the covariates are balanced and that all first, second, and third moments are the same between the two groups for original dataset obtained from Cohen, Coval, and Malloy (2011). For instance, the mean, variance, and skewness of Lag Q are 3.087, 23.07, and 3.296 respectively, in both treatment and control groups, implying that the balanced condition on Lag Q is well satisfied. One can improve the randomness assumption by assuring that the covariates in compared groups are balanced in higher moments, which makes the assumption of randomness required by nature experiment far more reasonable.

Following the calculation process shown in the method section, we produce the weight for each observation in the control group, such that the balanced condition can be sufficiently satisfied between the treatment and control groups. The density for weights' results is reported in Figure B1. The figure shows that considerable number of weights are modified from their original values, the weights assigned to control group observations range from 0 to 0.13, and no observations are assigned with extreme weights (i.e., weights near 1),[39] which implies that our result will not be biased owing to specific observations with large assigned weights.[40]

B.3 Results

B.3.1 Main results for the effect of powerful chairmanship on firm's R&D investment

To show the results for the effect of a powerful committee chairman on a firm's R&D, we initially use the regression method without a data-preprocessing step as in Cohen, Coval, and Malloy (2011). The result is reported in Table B3, left column. The results show that the chairmanship shock can decrease the R&D ratio (the coefficient is -0.005 at the 5%

[39] In order to compare and interpret the weights more quickly, the default program in stata "ebalance" scaled the weights, such that the original weights in the control group $1/n_2$ equals 1 and the sum of the total weights in the control group equals the total weights in the treatment group (n_1), where n_1 and n_2 are the number of observations in the treatment and control groups respectively.

[40] In the EB calculation, we try to produce the weights to ensure that the moments are well matched between the treatment and control groups. However, if we could not find an exact match with respect to the moments, we could set a tolerance level. In this study, the tolerance level is set at 1.5%.

Table B2 Summary of Entropy balancing results for effect of the powerful chairman on firm's R&D (After Weighting)

The data are collected according to Cohen, Coval, and Malloy (2011) and can be obtained from website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607. Controls include (1) lagged Q (Q ratio in previous year), (2) cash flow (scaled by asset in previous year), (3) lagged leverage (leverage ratio in previous year), (4) Year. More details are explained in the Appendix. The EB process is explained in the methodology section. More details on variables are recorded in the Appendix.

Summary of Entropy balancing results for impact of powerful chairman on firm R&D (After Weighting)

	Treatment Group			Control Group		
Variable	mean	variance	skewness	mean	variance	skewness
Lag Q	3.087	23.070	3.296	3.087	23.070	3.296
Lag Leverage	0.349	0.063	0.601	0.349	0.063	0.601
CashFlow_Scaled	-0.007	0.078	-2.473	-0.007	0.078	-2.473
Year	1,993	15	-1	1,993	15	-1

Figure B1 Histogram for weights obtained from EB calculation

The figure reports the density of weights obtained from EB calculation. The data are collected according to Cohen, Coval, and Malloy (2011) and can be obtained from website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607. Controls include (1) lagged Q, (2) cash flow (scaled by asset in previous year), (3) lagged leverage, (4) Year. More details are explained in the Appendix. The EB process is explained in the methodology section. More details on variables are recorded in the Appendix.

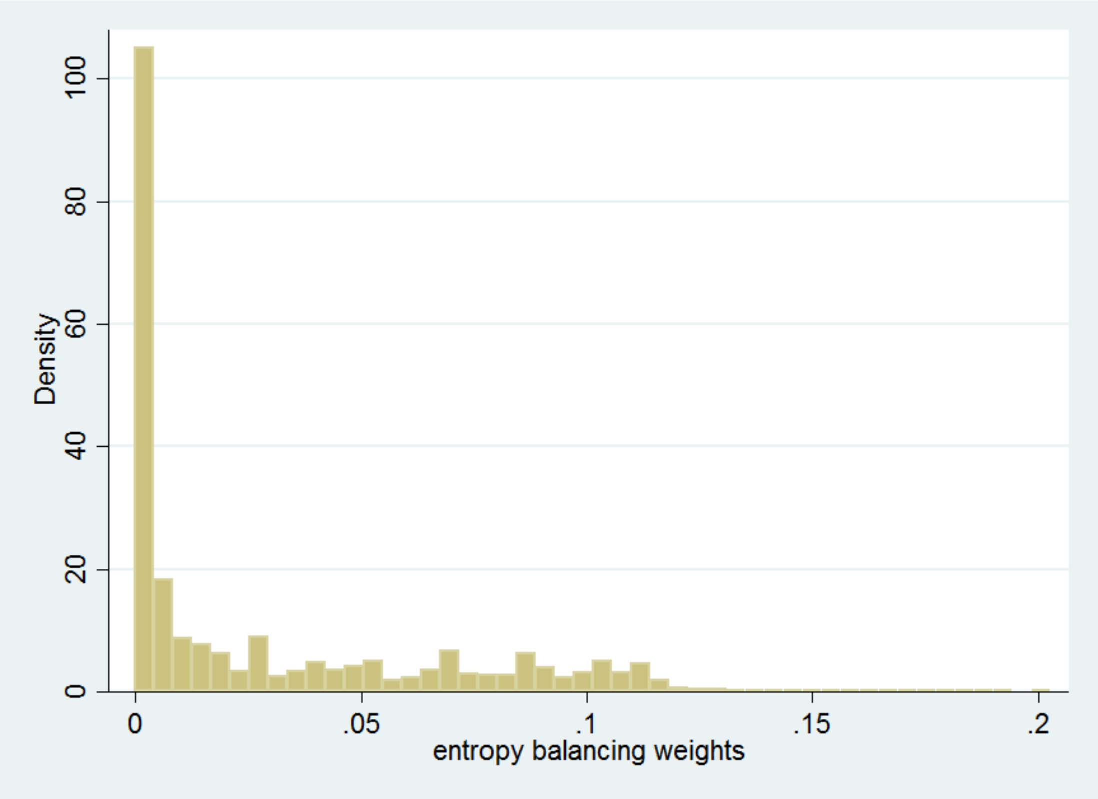


Table B3 Estimates of the powerful chairman's effect on firm's R&D

The data are collected according to Cohen, Coval, and Malloy (2011) and can be obtained from website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607. The dependent variable is defined as firm R&D expense scaled by (lagged 1 year) firm assets, Dummy(Shock_Chairman) is defined as a dummy variable that equals 1 if the firm locates in the state where a powerful congressional chairman holds office; otherwise, the variable equals 0. Controls include (1) lagged Q (Q ratio in previous year), (2) cash flow (scaled by asset in previous year), and (3) lagged leverage (leverage ratio in previous year). All tests include firm-level fixed effect and year-fixed effects.

The definition for each variable is explained in the Appendix. The estimation process is provided in the method section. Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels. The real number of observations might have been decreased for specific tests because of missing values, perfect predictions, or multi-collinearity problems.

Estimates of the powerful chairman's effect on firm's R&D

Dependent Variable:R&D/LagA	Ordinary Least Squares method	Entropy Balancing Method (combined with OLS method)
Independent Variables:	Coefficient	Coefficient
Dummy (Chairmanship Shock)	-0.005**	0.002
Lag Q	0.002***	0.002***
Lag Leverage	-0.046***	-0.048***
CashFlow_Scaled	-0.156***	-0.163***
Year fixed effect	Yes	Yes
Firm fixed effect	Yes	Yes
Constant	Yes	Yes
Number of observations	77598	77598
Adjusted-R Squrare	0.772	0.803

significance level).[41] Thus, the evidence shows that the chairman appears to "cause" a decrease in the firm's R&D.

Table B3, right column, reports the estimated coefficients after the EB data reweighting process for equation (19). We ensure that the same method and model is used as in Cohen, Coval, and Malloy (2011)'s study, except that we assign weights in control groups, such that the firm characteristics are as similar as possible with respect to the first, second, and third moments between the treatment and control groups, such that the assignment of politician shock is credibly exogenous to the background covariates. The estimates from the diagram show that the coefficients estimated in this EB test appeared very different and that the powerful chairman has no "causality" relationship with the firm's R&D investment (the coefficient is only 0.002 and not significant). In other words, this EB test shows that if we address the imbalance problem, there are no significant causality effects on a firm's R&D investment by comparing firms affected by a powerful politician with those without such a shock.[42]

B.3.2 Simulation studies

We now extend the bootstrap simulation approach to accommodate the data in Cohen, Coval, and Malloy (2011)'s study, such that statistical properties can test for the distinctions between EB, the regression method used in Cohen, Coval, and Malloy (2011)'s study. We begin by following the same simulation procedures as in the examples in the main paper so that we can compare EB methods with OLS methods in terms of bias and MSE. The results are reported in Table B4. The table reports the bootstrap simulation results for the effect of powerful committee chairman on a firm's R&D. The results estimated by the EB matching method and OLS method are different (0.002 not significant and -0.005 at the 5% significant

[41] We use the same data from the Cohen, Coval, and Malloy (2011), but we modify the coding according to the instructions of Snyder and Welch (2015) as the original dataset has errors in the coding of powerful chairmanship shock. For more details, see Snyder and Welch (2015). Nevertheless, our replicated results are qualitatively similar to those in the original study in Cohen, Coval, and Malloy (2011).

[42] Snyder and Welch (2015) argue that the effect is due to a special event. It is challenging for researchers to identify each potential event for a given study. Hence, we propose an automatic and direct method to detect the unbalanced problem and assign small weights (and minimize the information loss simultaneously) to the observations in the control group, if they are rather different from the treatment group observations.

Table B4: Comparison between Entropy Balancing method and OLS method

The table reports the simulation results for the effect of a powerful chairman on the firm's R&D investment. All of the benchmarks (bias and MSE) are based on 5000 bootstrap simulations (with replacement). In each simulation, we use the EB and OLS methods to estimate the effect of a powerful chairman on the firm's R&D investment. All of the control variables are the same between the EB method and the OLS method and are explained in the experiment design section. The calculation process for each benchmark (bias and MSE) is explained from method section in the main paper. Definitions for each variable are explained in the Appendix. The real number of simulations might be decreased because of no solutions for weights in a specific EB simulation.

Comparation between Entropy Balancing method and OLS method

Benchmark Settings	Entropy Balancing method	OLS Method
Original Treatment Effects	0.002	-0.005**
Bias	0	-0.001
MSE	12×10^{-6}	17×10^{-6}
No of simulations	4064	5000

level, respectively). The EB method appears to yield more reliable results as follows: (1) The bias for the EB matching method is almost zero, indicating that there is nearly no bias for this method in comparison to the bias of -0.001 for OLS. (2) The EB method can reduce much more MSE in comparison to the OLS method.

B.3.3 Comparison between the entropy balancing method and propensity score matching method for the example from Cohen, Coval, and Malloy (2011)

As discussed, to deal with the self-selection problem, PSM methods are currently widely used in observational studies for estimation of binary treatment effects based on the assumption of selection on observables. To estimate the propensity score, we use the same variables as those in the EB method and the OLS method from Cohen, Coval, and Malloy (2011)'s study. The logit regression result used to estimate the propensity score is reported in Table B5. Most coefficients in the regression study are significant to forecast the propensity of being selected in the treatment group. For example, Lag Q can positively determine the possibility of treatment adoption (the coefficient equals 0.001 at the 5% significant level). In contrast, some control variables (i.e., Cash flow) can negatively determine the probability of the treatment dummy (the coefficient is -0.238 at the 1% significant level). The results suggest that firm characteristics, such as Lag Q, Lag leverage, or Cash flow will determine the preconditions of whether the firm can self-select to be in the state where a powerful politician holds office, rather than the endogenous event argued by Cohen, Coval, and Malloy (2011)

B.3.3.1 Balanced condition problem in propensity score matching

With respect to the balanced condition in PSM, Figure B3 reports the propensity score distribution in Cohen, Coval, and Malloy (2011)'s study.[43] The figure shows that the balanced condition is not well satisfied as the frequency of the propensity scores in the treatment group cannot match those in the control group well. For example, during the periods after the political shocks, the propensity scores in the treatment group (approximately 0.03 or ranging from 0.06

[43] In this section, we report the aggregate level results. In an unreported table, we also compare the results in each term when the powerful politician sits as a chairman of a congress committee and the results are similar. This can be provided if requested.

Table B5. Logit regression results used to estimate propensity score in firm R&D

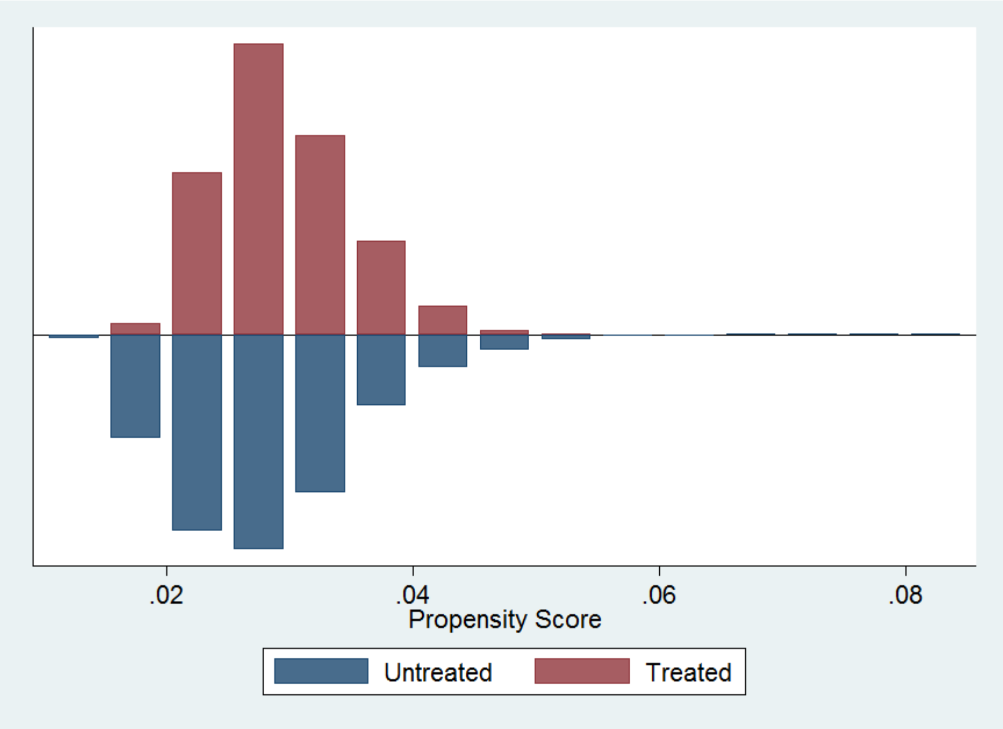
The data are collected according to Cohen, Coval, and Malloy (2011) and can be obtained from website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607. The dependent variable is a Dummy (Chairman Schok) that equals 1 if the firm locates in a state where a powerful congressional chairman holds office; otherwise, the variable equals 0. Controls include (1) lagged Q (Q ratio in previous year), (2) cash flow (scaled by asset in previous year), and (3) lagged leverage (leverage ratio in previous year)., (4) year, Asterisks indicate significance at the 0.01 (***), 0.05 (**), and 0.10 (*) levels.

Logit regression results used to estimate propensity score in firm R&D study

Dependent Variable: Dummy (Chairmanship Shock)	
Independent Variables:	Coefficient
Lag Q	0.001**
Lag Leverage	0.326***
CashFlow_Scaled	-0.238***
Year	0.026***
Constant	Yes
Number of observations	77,598
Pseudo-R Squrare	0.007

Figure B3 Frequency distributions of propensity scores

The figure reports the frequency distribution of propensity scores estimated by logit regression method. The propensity scores matching methods follow Heckman, Ichimura, and Todd. (1998) and Becker and Ichino (2002). The data are collected according to Cohen, Coval, and Malloy (2011) and can be obtained from website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607. Controls include (1) lagged Q (Q ratio in previous year), (2) cash flow (scaled by asset in previous year), (3) lagged leverage (leverage ratio in previous year), (4) Year. The distribution in red color is for treatment group and the distribution in blue color is for control group.



to 0.08) cannot find sufficient overlaps of propensity scores in the control group in the same range. Thus, the common support condition assumed by PSM may not be satisfied well and subsequent analysis conducted by this method may not be valid.

We follow our proposed EB method by combining the new weights into the regression model as in equation (B1). To compare the estimating performance of our proposed matching framework with that of the commonly used PSM method, we follow the same procedure from examples in the main paper and obtain the bias and MSE for each round of simulation. Table B6 summarizes our method and a PSM framework that ignores information loss due to discarding or reweighting the original dataset. The diagrammed results illustrate that the bias in the EB method is zero in comparison to 0.001 in the PSM method. With respect to the MSE problem, the EB method lowers the MSE more in comparison to the PSM method.

B.3.3.2 Comparison between the entropy balancing method and various propensity score matching methods

For robustness checks, we test if our conclusion will not hold if different PSM methods are employed. We use various matching schemes, such as local linear regression matching, radius matching, and nearest matching, and the results are in Table B7. The table illustrates the results on estimation performance from 1000 simulations, where the model design and variables' definitions are the same as in Cohen, Coval, and Malloy (2011). In the model, EB can effectively reduce bias as well as MSE.

B.4 Further robustness

In this subsection, we evaluate the methods on matching quality. We compute both standardized percentage bias on each covariate (locally matching), and jointly on the Rubin's B index ratio and Hotelling's T-squared generalized means test on the covariates (globally matching). We undertake this for the weighted data after EB matching, the weighted data after PSM matching, and then the raw data without any weighting scheme as a comparison.

Table B6: Comparison between Entropy Balancing method and Propensity Score Matching method

The table reports the simulation results for effect of a powerful chairman on the firm's R&D. All of the benchmarks (bias and MSE) are based on 5000 bootstrap simulations (with replacement). In each simulation, we use the EB and PSM methods to estimate the effect of the committee chairman on the firm's R&D. All of the control variables are the same between the EB method and the PSM method. The calculation process for each benchmark (bias and MSE) is explained in the simulation section. Definitions for each variable are explained in the Appendix. The real number of simulations might be decreased because of no solutions for weights in a specific EB and PSM simulation.

Comparation between Entropy Balancing method and Propensity Score Matching method

Benchmark Settings	Entropy Balancing method	Propensity Score Matching Method
Original Treatment Effects	0.002	-0.003
Bias	0.000	-0.001
MSE	12×10^{-6}	17×10^{-6}
No of simulations	4064	5000

Table B7 Comparison between Entropy Balancing method and various Propensity Score Matching methods

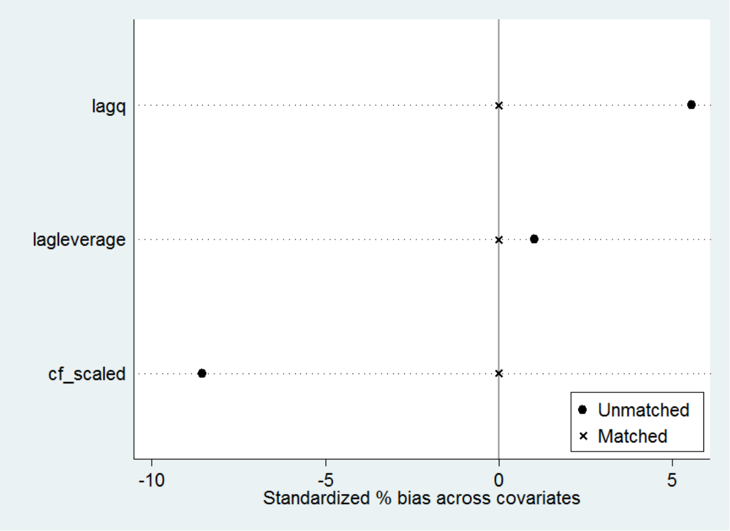
The table reports the simulation results for the effect of politician shock on firm's R&D. The propensity scores matching methods follow Heckman, Ichimura, and Todd. (1998) and Becker and Ichino (2002). The data are collected according to Cohen, Coval, and Malloy (2011) and can be obtained from website http://www.hbs.edu/faculty/Pages/item.aspx?num=40607. Controls include (1) lagged Q (Q ratio in previous year), (2) cash flow (scaled by asset in previous year), (3) lagged leverage (leverage ratio in previous year), (4) Year. The calculation process for each benchmark (bias and MSE) is explained in the simulation section. Definitions for each variable are explained in Appendix. The real number of simulations might be decreased because of no solutions for weights in a specific EB simulation.

Comparation between Entropy Balancing method and Various Propensity Score Matching methods

Benchmark Settings	Entropy Balancing method	Local Linear Regression Matching	Radius Matching	Nearest Matching
Original Treatment Effects	0.002	-0.005	-0.001	-0.005
Bias	6.27×10^{-4}	37.75×10^{-4}	6.64×10^{-4}	0.003
MSE	1×10^{-5}	11×10^{-5}	3×10^{-5}	7×10^{-5}
No of simulations	1000	1000	1000	1000

Figure B4 Standardized percentage bias across covariates before and after EB calculation

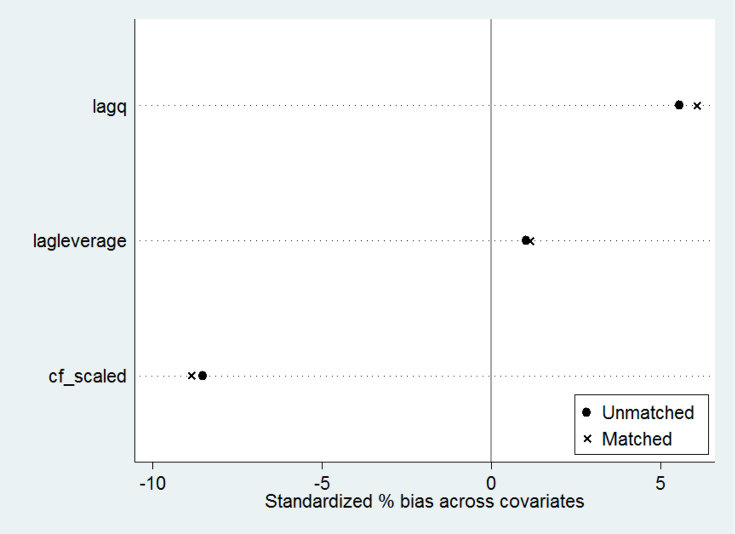
The figure reports the standardized percentage bias on the sample covariates before and after EB matching for treatment and control groups. The formula can be obtained from Rosenbaum and Rubin (1985). firm characteristics include (1) lagged Q (Q ratio in previous year), (2) cash flow (scaled by asset in previous year), (3) lagged leverage (leverage ratio in previous year)



.

Figure B5 Standardized percentage bias across covariates before and after PSM calculation

The figure reports the standardized percentage bias on the sample covariates before and after PSM matching for treatment and control groups. The formula can be obtained from Rosenbaum and Rubin (1985). firm characteristics include (1) lagged Q (Q ratio in previous year), (2) cash flow (scaled by asset in previous year), (3) lagged leverage (leverage ratio in previous year)



B.4.1. Further examination

The results in Figures B4 and B5 summarize the standardized percentage bias on the sample covariates before and after EB (and PSM) matching. Without the weighting scheme, raw data have appreciably higher imbalance problem on individual covariates. However, using the EB reweighting scheme, the differences in the covariates between the treatment and control groups decrease to zero. This suggests that the EB has power to reduce the imbalance problem caused by each covariate. However, the results in the figure show that after the PSM weighting scheme, the data has more imbalance problem in comparison to the raw data without matching since the standardized percentage bias on the sample covariates after PSM is increasing. The worse matching result after PSM may be caused by the imbalanced propensity score itself, as shown in figure, or more importantly, PSM may not be appropriate to be used in the natural experiment design as it tries to match globally rather than locally. The latter is one requirement of the randomness assumption in a natural experiment.

We then examine the overall imbalance problem with Hotelling's T-squared generalized means test and Rubins' B test. The middle column of Table B8 presents the estimates of Hotelling's T-squared generalized means test. The null hypothesis is that the vectors of the means are equal for the two groups. Given the results presented in the table, EB has successfully reduced the F statistic from 7.812 to 0, which indicates that the weighted data after the EB matching achieve the randomness assumption required by the natural experiment. Meanwhile, the PSM result is worse than the raw data result because the PSM increases the F statistic from 7.812 to 75.257. To provide insight into why the PSM has such poor matching and estimation properties, the literature on the PSM provides potential directions for the PSM. For example, for the logit model misspecification direction, see (Heckman et al. (1997, 1998,1999)). For structures applying the PSM, see (Iacus, King and Porro (2009)). For the unconfoundedness assumption check, see Nannicini (2007). Overall, the tables and figures show that the EB significantly reduces the imbalance (both locally and globally) problem, which does not occur with the commonly used PSM.

Table B8 Hotelling's T-squared generalized means test

The table reports the Hotelling's T-squared generalized means test and Rubins' B test for the raw data, the data with weights after PSM matching, and the data with weights after EB matching. In Hotelling's T-squared generalized means test evaluates whether a set of means equal between two groups, the Null Hypothesis is vectors of means are equal for the two groups.

Hotelling's T-squared generalized means test

Statistics of the test	Hotelling's T-squared generalized means test (Null Hypothesis: Vectors of means are equal for the two groups)				Rubin's B
Raw Data	F(3,77594):	7.812	Prob>F(3,77594):	0	10.1
Data After PSM matching	F(3,77594)	75.257	Prob>F(3,77594):	0	10.8
Data After EB matching	F(3,77594)	0	Prob> (3,77594)	1	0

Appendix C: Propensity Score Matching

Propensity score matching methods are used following the initial work of Rosenbaum and Rubin (1983) and Heckman, Ichimura, and Todd (1998). For more on the econometrics of the method see Becker and Ichino (2002). For applications see Vega and Winkelried (2005); de Mendonça and de Guimarães e Souza (2012); Lin and Ye (2007, 2009) and Samarina, Terpstra, and De Haan (2014).

Consider the SOX example of the main paper,

$$ATT = E[Y_{i1}^{k}|D_i = 1] - E[Y_{i0}^{k}|D_i = 1] \tag{C1}$$

which presents the simplest version of the average treatment effect on the treated (ATT) concept. Y_i^k is the outcome for observation i and variable k and D_i is a dummy variable capturing whether or not treatment has occurred. Here

Di=

1 if the observation is in treatment group

to increase the board independence according to the SEC regulation

0 if if the observation is in control group

The term $Y_{i1}^{k}|D_i = 1$ is the economic outcome of variable k given that i is a firm that was forced to increase board independence, and $Y_{i0}^{k}|D_i = 1$ is the counterfactual. That is, the second term in equation (20) measures what the economic outcome of variable k would be for firm i if firm i was not complying with the new SEC regulation. There are two difficulties in estimating the ATT using equation (20). First, this second term $Y_{i0}^{k}|D_i = 1$ in practice is not observed since it is not possible to observe the economic outcomes in a firm complying with the SEC regulation that has not actually complied. The assumption needed for the matching method to be realized is the conditional independence assumption, or unconfoundedness (Dehejia and Wahba, (2002)).

This is

$$Y_0, Y_1 \perp D | P(X) \tag{C2}$$

where X are a set of covariates which are not affected by being treated so that the outcomes are independent of the treatment (whether or not to comply with the new SEC regulation) so that

$$E[Y_{i0}|D_i = 1, X_i] = E[Y_{i0}|D_i = 0, X_i] = E[Y_{i0}|X_i] \qquad \text{(C3)}$$

Equation (20) is then re-expressed as

$$ATT = E[Y_{i1}|D_i = 1, X_i] - E[Y_{i0}|D_i = 0, X_i] \qquad \text{(C4)}$$

The second problem that arises in estimating the ATT occurs when the number of covariates in X increases resulting in a dimensionality problem. Rosenbaum and Rubin (1983) propose a method to match the treated group with the control group using their propensity scores in this circumstance (see Lin and Ye, 2007, 2009; and de Mendonça and de Guimarães e Souza, 2012). These probabilities (or propensity scores) are commonly estimated using probit or logit models. Here, equation (23) is rewritten as

$$ATT = E[Y_{i1}|D_i = 1, p(X_i)] - E[Y_{i0}|D_i = 0, P(X_i)] \qquad \text{(C5)}$$

This is the propensity score matching estimator. There are a variety of methods to match the treated firms with the comparable control group firms which mainly arise because P (X_i) is continuous. The matching methods may include: nearest neighbor matching; radius matching; kernel matching (with Mahalanobis matching method).[44] The matching methods are summarized in Heckman, Ichimura, and Todd (1998) and Becker and Ichino (2002).

In order to estimate treatment effect by PSM, previous literature proposes several methods. In this paper, according to Heckman, Ichimura, and Todd. (1998), Kernel Matching method will be employed.

Let c be the set of units in control groups, and T be the set of units in treated group. Let Y_i(C) and Y_j(T) be the observed outcomes for control units and treated units respectively. Let C(i) be the set of units in control group matched to the unit i in treated group with estimated propensity score of P_i.

The estimator of kernel matching is given by:

$$T^K = \frac{1}{N^T}\sum_{i \in T}\{Y_i^T - \frac{\sum_{j \in c} Y_J^C G(\frac{P_j - P_i}{h_n})}{\sum_{k \in c} G(\frac{P_k - P_i}{h_n})}\} \qquad \text{(C6)}$$

Where h_n is the smoothing parameter which set specific bandwidth and G is function (Gaussian) for kernel. In kernel matching method, one can match each treated element based on kernel-weighted

[44] For the Mahalanobis matching method, see Rubin (1978).

average of control unit outcomes. The standard errors can be exclusively achieved by bootstrapping procedure

Similar approaches on nearest neighbor matching; radius matching and Local Linear Regression Matching are summarized in Heckman, Ichimura, and Todd (1998) and Becker and Ichino (2002).

Appendix D: Definitions of the variables in SOX study

Definitions of the variables in board independence study

Variable	Definition
Ln(Total CEO Compensation)	The natural log of CEO compensation
Dummy(Noncompliant Board’02)	Dummy(Noncompliant Board’02) is coded as 1 if the firm did not comply with the Regulation in the year 2002 and as 0 otherwise
Sales	The natural log of company sales
ROA	The natural log of one plus net income before extraordinary items and discontinued operations divided by the book value of assets—all measured in (t−1)
Stock Returns	The natural log of the annual gross stock return (dividend reinvested), measured in year (t−1)
CEO Tenure	The number of years the CEO served in the firm
Dummy(’03-’05)	Dummy(’03-’05) is set to 1 if the observation is in the periods 2003-2005 and 0 otherwise
Dummy(‘00-‘02)	Dummy (’00-’02) equals 1 if the observation is in the periods 2000-2002 and 0 otherwise
Industry-year fixed effect	Industry-year fixed effect which is defined as Fama-French 48 Industry factor times year dummy variable

Appendix E: Definitions of the variables in Powerful politician study

Definitions of the variables in chairmanship appointment study

Variable	Definition
Dummy (Chairmanship Shock)	Firm R&D expense scaled by (lagged 1 year) firm assets
Dummy(Shock_Chairman)	A dummy variable which equals 1 if the firm locates in the state where a powerful congress chairman represents otherwise equals 0
Lag Q	Q ratio in previous year
Lag Leverage	Leverage ratio in previous year
CashFlow_Scaled	Cash flows scaled by (lagged 1 year) firm assets