Appendix B. PSM Details

Propensity Score Matching

PSM analysis will be performed via the following four steps:

Step 1: Identify the pre-treatment characteristics that will be used in the propensity score model to match fellows and unfunded applicants. These characteristics will include variables that both predict receiving fellowships and that might affect the outcomes of interest and will be culled from NSF extant data and from survey data. Degree granting institutions, GPA, gender, and time to PhD are all candidates for pre-treatment characteristics.

Step 2: Fit a logistic model that predicts the probability of being awarded a fellowship based on pre-treatment characteristics. Use the coefficients from this model to estimate the propensity score for each individual, which represents the probability of receiving a fellowship. Finally, we identify and exclude from impact analyses those individuals who are outside of the "common support" group – the range of common scores across fellows and unfunded applicants. Enforcing the common support is important to ensure the similarity of the matched non-awardees to awardees.[1]

Step 3: Use the estimated propensity scores to create matched sets of fellows and unfunded applicants. Propensity scores can be utilized in a number of ways, including matching, stratification, weighting, and regression adjustment.[2] Use stratification (also called interval matching) as our primary method, which entails constructing a number of propensity score strata by dividing all treatment and comparison group members who are in the common support into subgroups of equal size based on the propensity scores. Determine subgroups or number of strata, standard practice is often five (Rosenbaum and Rubin, 1983). This method because it allows for the inclusion of the largest number of cases and does not impose a functional form (e.g., linear) on the relationship between propensity to participate and treatment effect.

Step 4: Test whether there are any differences between the awardees and non-awardees within each propensity score strata. There are several ways of performing this analysis. One way is using a t-test for each pre-treatment characteristic.[3] Another is using an F-test to jointly test whether the awardees are similar to the non-awardees in each propensity score stratum which

---

[1] Rosenbaum and Rubin, 1983 and Kaliendo and Copeining, 2008

[2] Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, 71(4): 1161-89; Morgan S.L. and Harding D. J. (2006). "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice." *Sociological Methods & Research*, 35(1), 3–60; and Abadie, A., & Imbens, G. W. (2009). Matching on the Estimated Propensity Score. NBER Working Paper.

[3] Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics*, 84(1): 151-61; Agodini, Roberto, and Mark Dynarski. 2004. "Are Experiments the Only Option? A Look at Dropout Prevention Programs." *Review of Economics and Statistics*, 86(1): 180-94.

takes the correlation between the matching characteristics.[4] As these tests are sensitive to sample size (*i.e.*, they tend fail to detect sizable differences in small samples, but detect slight differences in larger samples), these will be supplemented using standardized differences.[5] The standardized difference of a matching characteristic between awardees and non-awardees in a given propensity score stratum is calculated using:

$$(1) \qquad B_{X,S} = \frac{|\overline{X}_{T,S} - \overline{X}_{C,S}|}{\sqrt{\frac{1}{2}\sigma^2{}_{X,T} + \frac{1}{2}\sigma^2{}_{X,C}}}$$

Where:

X denotes the variable of interest;
S denotes the stratum;
T denotes the treatment group, and C denotes the comparison group;
$\overline{X}_{T,S}$ and $\overline{X}_{C,S}$ denote the treatment and comparison group mean of X in stratum S; and
$\sigma^2{}_{X,T}$ and $\sigma^2{}_{X,C}$ denote the overall variance of X in the treatment and comparison group, respectively.

Standardized differences larger than 0.15 will be considered to be suggestive evidence of treatment-comparison group unbalance with respect to the corresponding variables. If statistical balance is not achieved across treatment and comparison groups in each stratum, the logistic model used in Step 2 will be modified by including interactions and higher-order terms of the unbalanced characteristics and repeat Steps 2 through 4 until satisfactory balance is achieved.

[4] Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). "Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs?" *Review of Economics and Statistics* , 86, 156-179

[5] Morgan, S.L., & Winship, C. (2008) "Counterfactuals and causal inference: Methods and principles for social Research" New York: Cambridge University Press.

Estimation of Impacts

Following the matching, the impact of the EAPSI and IRFP programs will be estimated by comparing fellows' outcomes to those of their comparison group to determine what fellows' expected outcomes would have been had they not received an EAPSI and IRFP award.

Estimation of the impacts

After creating the propensity score strata, a multivariate regression model will be used to estimate the impact of the program of interest. This regression model will employ a number of matching characteristics and other variables that are hypothesized to affect the outcomes of interest as covariates. The inclusion of the matching characteristics in this model will give us the chance to get a "doubly-robust" impact estimate since they will have been used twice: both in the propensity score model and in the estimation of impacts.[6] The following is a prototypical regression model that will be used to estimate the program impacts:[7]

$$(2) \qquad Y_i = \beta_0 + \sum_{j=1}^{4} \beta_j S_i^j + \sum_{j=1}^{5} \beta_{(4+j)} S_i^j T_i + \sum_{n=1}^{N} \beta_{(9+n)} X_i^n + \varepsilon_i$$

Where:

Y_i is the outcome of interest for individual i,

T_i is the treatment indicator for individual i (1=treatment , 0=comparison group),

S_i^j is the indicator (dummy) variable for the j^{th} propensity score stratum. As mentioned above, we will use five propensity score strata; hence the prototypical model includes four strata indicators (j=1,2,…,4) and the fifth stratum is set to be reference stratum whose indicator is not included in the model,

X_i^n is the n^{th} (n=1,2,…,N) covariate for individual i (such as gender, age, etc.) that are grand-mean centered, and

ε_i is the usual error term for individual i.

Interpretation of the coefficients in the model is as follows:

β_0 is the mean value of the outcome for the non-awardees in the reference (fifth) propensity score stratum,

β_j (j=1,2,…,4) is the difference between the mean value of the outcome of the non-awardees in the j^{th} stratum and the reference stratum,

β_{4+j} (j=1,2,…,5) is the impact estimate (i.e., the covariate adjusted difference between the outcomes of the awardees and non-awardees) for the j^{th} stratum, and

---

[6] Ho D.E., Imai K., King G., and Stuart E. A. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political Analysis*. 2007; 15: 199–236.; Morgan S.L. and Harding D. J. (2006). "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice." *Sociological Methods & Research*, 35(1), 3–60.

[7] For illustrative purposes, we present the impact model for continuous outcomes. For binary outcomes, we will fit a logistic model which is structured similarly to the model in Equation 1.

β_{9+n} (n=1,2,…,N) is the estimated overall relationship between the n^{th} covariate and the outcome controlling for other covariates.

As seen, the model in Equation 2 allows for the estimation of separate treatment effect estimates for each propensity score stratum. More specifically, the estimate of coefficient β_{4+j}, $\hat{\beta}_{4+j}$ is the impact estimate for the j^{th} (j=1, 2,…, 5) stratum. In order to calculate an overall treatment effect estimate, the stratum-specific estimates are aggregated as follows:

(3) $$TE = \sum_{j=1}^{5} P_j \hat{\beta}_{4+j}$$

Where P_j is the proportion of treatment group members in the j^{th} stratum, which is used to weight the strata-specific impact estimates.[8] Standard error of the overall treatment effect estimate can be then calculated as:

(4) $$\text{Std Error}(TE) = \sqrt{P^T VCV(\hat{\beta})P}$$

Where

P is a 5x1 vector that holds P_j (j=1,2,…,5), and

$VCV(\hat{\beta})$ is the portion of the variance-covariance matrix of the estimated impact model that holds the estimates of the variances of and covariances between the stratum-specific impact estimates.

Estimated coefficients from the impact model and the overall impact estimate will be presented as well as their corresponding standard errors and p-values. Hence, for dichotomous outcomes, impact estimates will be presented in the form of percentage points. For continuous outcomes, overall impact estimates in "effect size" units (e.g., Hedges' *g*) will also be presented. The effect size for an impact estimate will be calculated as:

(5) $$ES = \frac{TE}{PooledSD}$$

Where

TE is calculated as shown in Equation 3, and

[8] Stratum-specific treatment effect estimates can be aggregated to yield an overall impact estimate in a number of ways. The method chosen here—weighing the estimate for each stratum by the proportion of treatment group members in that stratum—is widely used (Morgan S.L. and Harding D. J. 2006. "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice." *Sociological Methods & Research*, 35(1), 3–60; Caliendo, Marco and Sabine Kopeinig. 2007. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys*, 22(1): 31-72).

(6) $$PooledSD = \sqrt{\frac{(N_t - 1)S_t^2 + (N_c - 1)S_c^2}{(N_t - 1) + (N_c - 1)}}$$

Where

N_t = sample size of treatment group,

N_c = sample size of comparison group,

S_t^2 = variance of the outcome for treatment group (unadjusted), and

S_c^2 = variance of the outcome for comparison group (unadjusted).