Appendix E: Sampling Plan and Estimates of Differences

This Appendix presents details of the sampling plan for the S-STEM recipient survey and the details of the quasi-experimental comparison of S-STEM recipients to a propensity-score-matched comparison group of participants in the Beginning Postsecondary Students (BPS) Longitudinal Study (see description below). Before presenting the sampling and analysis plans, we begin with an overview of the analyses proposed for the study.

Overview of the Analyses

The evaluation of S-STEM will draw on extant data as well as require new data collection efforts. This package seeks OMB approval for the new data collection efforts, which include a survey of S-STEM principal investigators and a survey of S-STEM scholarship recipients. The research questions of this study will be addressed through a combination of descriptive, relational, benchmarking, and quasi-experimental comparative analyses. An overview of each of these is presented below, and additional details including the advantages and limitations of the evaluations approach, are discussed in Supporting Statement B.

Descriptive analyses of strategies of S-STEM projects will describe the ways in which S-STEM projects (i.e., grantee institutions) recruit and retain students in STEM fields, allocate scholarship funds, and provide educational and support programming for scholarship recipients. Data from the S-STEM Monitoring System will be analyzed to describe variation in student support activities (e.g. academic support, career counseling, recruitment, research opportunities) offered as part of S-STEM. PI surveys and interviews during site visits will further probe the activities that are offered as part of the S-STEM program and other supports that are available to S-STEM students.

Relational analyses of associations between strategies of S-STEM projects and outcomes will explore the relationships between S-STEM program services and supports and outcomes of interest. The PI web surveys will provide data on program characteristics and the recipient web surveys will provide data on recipient outcomes for these analyses.

Benchmark comparisons of S-STEM recipient educational and academic support experiences will use items and data from the NSSE survey to provide a non-matched comparison group against which to benchmark selected outcomes for currently-enrolled S-STEM recipients. The variables that will be used in the benchmark comparison include the items from the 2011 NSSE survey.[1] These include measures of students' allocation of time and effort to curricular and co-curricular activities and interactions with faculty members.

Quasi-experimental comparative analyses of S-STEM recipients to a matched comparison group will provide estimates of effects on key program outcomes. We will use propensity-score-matching (PSM) to compare responses of S-STEM recipients to a comparison group of participants in the Beginning Postsecondary Study (BPS) surveys (NPSAS:04, BPS:04/09). Propensity score matching allows a comparison of the S-STEM scholarship recipients (treatment group) to BPS respondents

[1] We have secured both permission to use NSSE 2011 survey items and a data license from the Center for Postsecondary Research at the University of Indiana School of Education.

(comparison group) selected based on their similarity to the S-STEM scholarship recipients.[2] PSM is a common quasi-experimental design approach that has been shown to produce unbiased estimates of the difference between the treatment and comparison group.[3] A detailed exposition of these methods is presented in Supporting Statement B and Appendices.

The treatment group for this quasi-experiment will consist of S-STEM scholarship recipients. The sample of S-STEM recipients will be restricted to those enrolled in an associate's or bachelor's degree program who received an S-STEM scholarship from either a two- or four-year institution that was awarded an S-STEM grant between 2006 and 2010. The comparison group will consist of BPS survey respondents. Within each S-STEM awardee institution,[4] S-STEM recipients will be matched to BPS survey respondents from the same institution on student level characteristics (including receipt of financial aid, academic information, demographic characteristics, and discipline). If the matching is not possible within an awardee institution, we will match students from institutions with similar institutional characteristics measured in the Integrated Postsecondary Education Data System (IPEDS).[5]

Sampling Plans

Below, we provide detailed sampling plans for the S-STEM PI Survey and S-STEM Recipient Survey. (No sampling plan for site visits is provided; site visits will be conducted at a purposive sample of S-STEM awardee institutions.)

From 2006 to 2010, the S-STEM program granted 513 S-STEM awards. Of these, 19 provide S-STEM scholarships only to graduate students and will be excluded from the sampling frame. The evaluation will examine the remaining 494 S-STEM awards that provide scholarships to undergraduate recipients (see Exhibit E.1). Given that earlier Abt studies have achieved location rates of 75%[6] and response rates of 80%,[7] this study has set a target response rate of 80%. We also include a plan for nonresponse bias analysis (see section B.4 and Appendix G), per OMB guidance[8] in the event that an 80% response rate is not achieved.

---

2 J. D. Angrist, "Estimating the labor market impact of voluntary military service using social security data on military applicants," *Econometrica*, 66 (1998): 249-288. 1998; J. Heckman, H. Ichimura, J. Smith, and P.Todd, "Characterizing selection bias using experimental data," *Econometrica*, 66 (1998): 1017-1098.

3 Rosenbaum and Rubin, "Reducing bias in observational studies"; Heckman et al., "Characterizing selection bias using experimental data"; Thomas D. Cook, William R. Shadish, and Vivian C. Wong, "Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons," *Journal of Policy Analysis and Management*, 27(4) (2008), 724-750.

4 *Examples* of these characteristics are: state and geographic information, sector of institution (public, private, non-profit), level of institution (2 - year vs. 4 year), historically black college, degree of urbanization, Carnegie classification, cost of attendance, selectivity of the institution, enrollment size, enrollment characteristics

5 S-STEM recipients and BPS respondents will be matched on selected characteristics during their first year of enrollment – either at their S-STEM institution (S-STEM recipient) or at their first-ever post-secondary institution (BPS respondents).

6 Abt Associates Inc., Needs assessment of the NIGMS Research Supplements to Promote Diversity in Health Related Research: Final Report, April, 2009.

7 Abt Associates Inc., CAREER, GK-12, and NSF-International studies.

8 Office of Management and Budget Standards and Guidelines for Statistical Surveys, September 2006. http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf

PI Survey Sample

There were 513 S-STEM awards made from 2006 through 2010, of which 19 are excluded because they give scholarships to graduate students only. This leaves 494 eligible awards in the sampling frame; of these, there are 483 unique PIs (11 had more than one S-STEM award). We propose to survey the census of 483 unique PIs.

We propose a census of PIs because extant data are not available on characteristics or models of the S-STEM projects, which vary with respect to recruitment and selection strategies, educational opportunities and support services for scholarship recipients. The lack of data makes it difficult to divide this population into homogenous subgroups to obtain reasonable strata from which to sample, and a simple random sample could potentially leave out programs that are unique in nature and not provide precise estimates of the population. Because we propose to survey the census of PIs, we do not present a sampling plan. All 483 unique PIs will be invited to participate in the PI survey.

Recipient Survey Sample

Of the 494 eligible S-STEM awards, 462 had student-level data in the monitoring system. These awards were made to a total of 377 unique IHEs (n=75 two-year IHEs and n=302 four-year IHEs). These two-year IHEs had a population of 5,477 undergraduate S-STEM recipients and the four-year IHEs had a population of 14,391 undergraduate S-STEM recipients (these numbers include those who are currently or were formerly enrolled) for a total of 19,868 eligible recipients in the sampling frame (see Exhibit E.1).

We will select the census of two-year awardee IHEs and a sample of 3,074 S-STEM scholarship recipients from within these 75 IHEs. From four-year awardee IHEs we will select a sample of 166 IHEs and within these we will select a sample of 5,146 S-STEM scholarship recipients. The total number of S-STEM recipients who will be invited to complete the recipient survey is 8,220. (In analyses, we will compare recipients selected from within an IHE to a matched comparison group of students who attended the same IHE and were participants in the BPS:04/09 survey – for which we will use extant data.[9] An overall estimate of differences (in outcomes) between S-STEM recipients and BPS respondents will be calculated by averaging across differences observed within each IHE.)

The analytic sample size estimates for the two- and four-year IHE recipient samples are based on a desired minimum detectable effect size (MDE) of a 0.075 for continuous outcomes (such as time to degree), and corresponding minimum detectable differences (MDDs) of between 2.3 and 3.8 percentage points for dichotomous outcomes (such as "earned degree" versus "did not earn degree").[10] Previous literature relevant to this study has shown that the typical effect size for similar continuous outcomes ranges from 0.075 to 0.2 and from 5 to 20 percentage points for dichotomous outcomes(e.g., Crisp et al., 2009; Eagan et al, 2010; Dowd & Coury, 2006; Ishitani,

[9] If the matching is not possible within an awardee institution, we will match students from institutions with similar institutional characteristics using data from the Integrated Postsecondary Education Data System (IPEDS). Examples of these characteristics are: geographic location of institution, sector of institution (public, private, non-profit), level of institution (2 - year vs. 4 year), historically black college/university, degree of urbanization, Carnegie classification, cost of attendance, selectivity of the institution enrollment size, and other enrollment characteristics.

[10] The MDE (used for continuous variables) is expressed as a percent of the standard deviation of the outcome, and the MDD (used for dichotomous variables) is expressed as a percentage point difference in the mean value of the outcome.

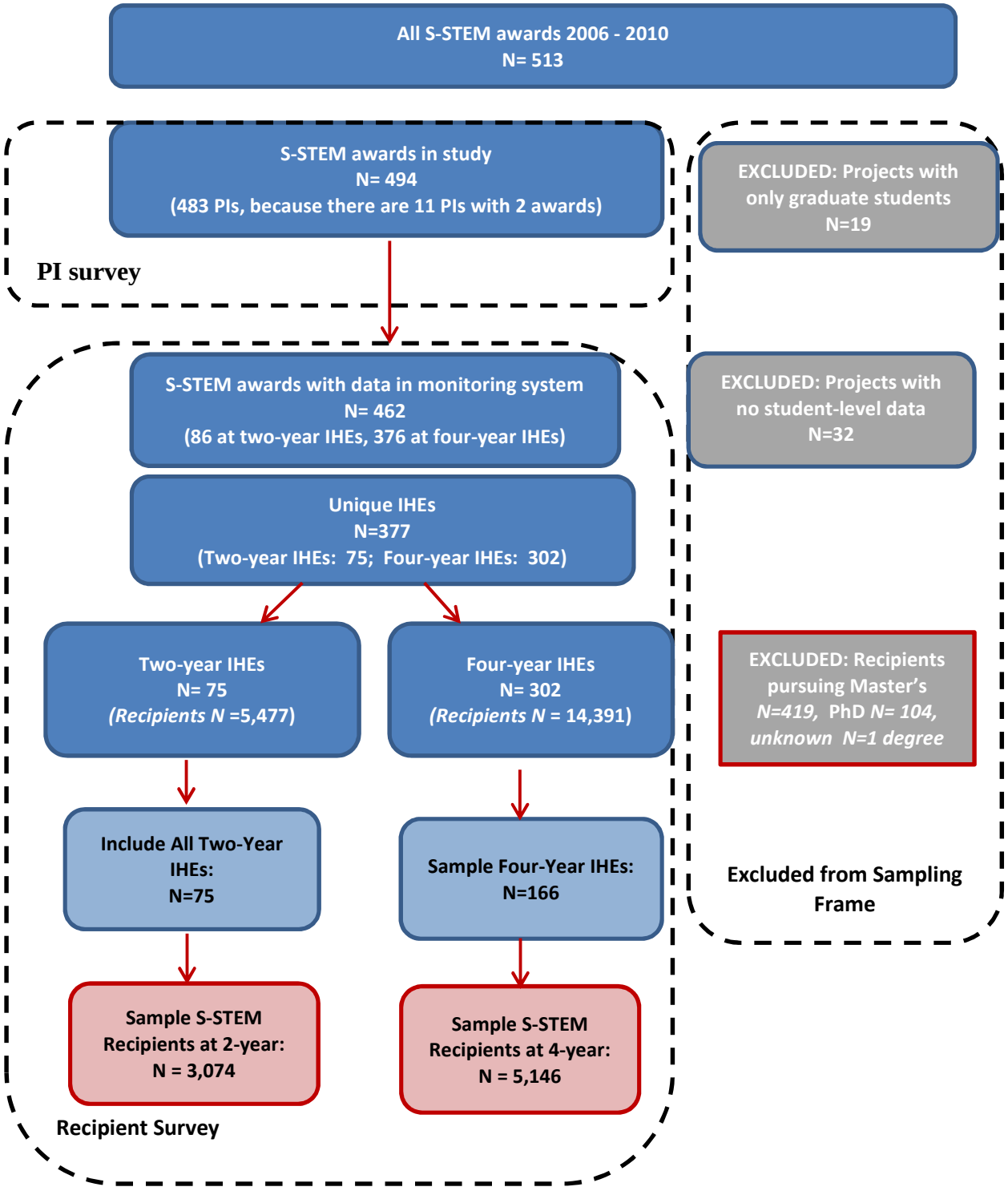
2012; Melguizo & Dowd, 2006). Based on this literature, the proposed evaluation is designed to detect MDEs of 0.075 which corresponds to MDDs of between 2.3 and 3.8 percentage points. Sampling calculations are based on the following assumptions:

- Significance level (alpha) = 0.05;
- Power = 80 percent
- The variance of effect size of the outcome across S-STEM awardee institutions is zero (this assumption is consistent with a fixed effects model for the treatment variable).
- The proportion of variation in outcomes explained by institutional-level covariates (reported symbolically as B) is approximately 0.1[11] and that the proportion of variation explained by individual-level covariates (R-squared) is 0.2.[12]
- The number of units in the constructed comparison group will equal the number of units in the "treatment" group, namely the S-STEM scholarship recipients.

[11] Dowd & Coury, 2006; Melguizo & Dowd, 2006 show that selectivity of school explains 10 percent of persistence and graduation rates.

[12] Adelman, C. (1999). Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment. Washington, DC: U.S. Department of Education. Retrieved September 20, 2012 from http://www2.ed.gov/pubs/Toolbox/Title.html.

Exhibit E.1: Study Samples for PI and S-STEM Recipients for Surveys



The proposed analysis will compare outcomes for two groups of respondents within each selected IHE: S-STEM scholarship recipients who respond to the Recipient Survey; and a comparison group of respondents who attended the same IHE and completed the BPS:04/09 survey (for which extant data are available). Thus, the necessary sample size is based on the total number of cases in the treatment and comparison groups combined, for each IHE and the total number of selected IHEs.

Exhibit E2 shows the required number of IHEs and respondents per IHE needed to detect a range of minimum detectable effect sizes. Highlighted rows show the relevant MDE and the corresponding analytic sample sizes needed for the recipient samples at two- and four-year IHEs: for the sample of recipients from two-year IHEs, Exhibit E2 shows that 84 IHEs, each with an average of 50 individuals (treatment and comparison groups combined), would provide the power needed to detect an MDE of .075 for continuous outcomes and that 138 four-year IHEs, each with an average of 30 individuals, would be needed to detect an MDE of 0.075.

Exhibit E.2: Analytic sample sizes of IHEs and respondents per IHE needed for a given minimum detectable effect size[a]

Minimum Detectable Effect Sizes (MDEs)	Analytic Sample Size (S-STEM awardee IHEs)	Mean analytic sample size per S-STEM awardee (Treatment + Comparison groups)	Mean analytic sample size of S-STEM recipients (Treatment group only)
0.10	40	60	30
0.10	47	50	25
0.10	58	40	20
0.10	79	30	15
0.10	115	20	10
0.075	69	60	30
0.075	**84**	**50**	**25**
0.075	103	40	20
0.075	**138**	**30**	**15**
0.075	202	20	10
0.05	156	60	30
0.05	187	50	25
0.05	237	40	20
0.05	315	30	15
0.05	472	20	10

Notes
[a] Source: Optimal Design software, developed by Raudenbush et al.

For dichotomous outcomes, we will be powered to detect a minimum difference between the S-STEM recipients' and the matched BPS respondents' outcomes of 2.3-3.8 percentage points (depending on the comparison group's mean proportion) as shown in Exhibit E3. For example, if 30 percent of the comparison group have successfully earned a Bachelor's degree, then the MDD the study will be powered to detect is 3.4 percentage points. If the true percentage of S-STEM recipients with a Bachelor's degree is 33.4 percent or higher, or 26.6 percent or lower, the study would be powered to detect this difference.

Exhibit E.3: Minimum detectable differences for dichotomous outcomes corresponding to an MDE of .075 for observed outcomes in the comparison group		
MDE	Proportion of the comparison group where one of two possible outcomes is observed (e.g., proportion of observed "yes" responses for a yes/no or other binary variable)[13]	MDD[14]
0.075	0.1	0.023
0.075	0.2	0.03
0.075	0.3	0.034
0.075	0.4	0.037
0.075	0.5	0.038

Having identified the necessary sample sizes for the desired MDE (and MDD), we next discuss the estimated initial sample sizes needed to result in these final sample sizes. Previous research with similar types of respondents suggests that locating participants in a past program, where contact information at the time of participation was known, typically succeeds for approximately 75 percent of the target sample; despite reasonable attempts to locate respondents, about 25 percent of such a sample cannot be located due to outdated contact information. [15] Of those respondents that can be successfully located, response rates for past program participants are approximately 80 percent (Exhibit E4).[16] Finally, during the propensity score matching phase, empirical evidence suggests that an appropriate match in the comparison group cannot be found for 20 percent of the respondents (Rubin & Thomas, 1996).

For the proposed evaluation of S-STEM program, contact information for former scholarship recipients will be up to seven years out of date (the first cohort of recipients were funded beginning in 2006-07 but the S-STEM Monitoring System was not operational until 2009; assuming data collection begins in the Spring of 2013, contact information could be as much as seven years out of date). We estimate that 75 percent of the S-STEM recipients listed in the S-STEM Monitoring System will be successfully located; this estimate is based on past experience attempting to locate program participants with a similar number of elapsed years between collection of the contact data and attempts to survey these populations using similar Monitoring System data (e.g., the Noyce Monitoring System; the IGERT Monitoring System). Based on response rates in similar studies of former college students (Exhibit E4), we estimate that the response rate for located S-STEM recipients will be approximately 80 percent (and potentially higher for recipients who are currently enrolled at their S-STEM IHE). Finally, we apply an estimated 20 percent loss of respondents during the PSM phase. Applying these assumptions requires initial sample sizes of individual respondents be increased as follows:

$$N_{Initial} = N_{Final} / (.75 \times .80 \times .80) = N_{Final} / (0.48)$$

[13] Proportion of success for the control group was based on the degree attainment and transfer rates from the BPS study.

[14] $MDD = MDE * \sqrt{q * (1-q)}$; where q is the proportion of success for the control group.

[15] Abt Associates Inc., Needs assessment of the NIGMS Research Supplements to Promote Diversity in Health Related Research: Final Report, April, 2009.

[16] Abt Associates Inc., CAREER, GK-12, and IGERT studies.

Where N_{Final} = the final analytic sample size (of treatment and comparison group combined) needed for the desired MDE and $N_{Initial}$ = the initial sample size needed to yield N_{Final} given the assumed losses due to non-location, non-response, and loss during the propensity score matching procedure.

Exhibit E.4: Survey response rates among past program participants based on lag between participation and data collection

Program	Response Rate	Length of Time Between Participation and Data Collection
NSF IGERT Trainees	74%	0-10 years
NSF GK-12 Fellows	MS-degree seeking: 83% PhD-degree seeking: 92%	0-5 years
NOYCE Teaching Fellows	64% to 82%	0-7 years

Exhibit E5 shows the resulting number of S-STEM recipients needed in the two-year and four-year IHE samples to ensure MDEs of .075. (Note that for the comparison group, extant data from the BPS:04/09 survey are not subject to sampling or response rate restrictions.) For example, to ensure that the final analytic sample includes a mean of 15 S-STEM recipients in each four-year S-STEM IHE, we will select an initial sample of 31 such recipients per IHE (31 x .75 x .80 x .80 = 14.9 respondents).

Exhibit E.5: Target number of S-STEM recipients

S-STEM recipient sample	Mean N per IHE (Treatment + Comparison group) needed in final sample size	Mean N of S-STEM recipients per IHE (Treatment group) needed in final sample size	N of S-STEM recipients per IHE needed in initial sample
Recipients in the four-year IHE sample	30	15	31
Recipients in the two-year IHE sample[17]	50	25	52

In addition to a larger initial sample size of recipients within each selected IHE, we will also select an initial sample of IHEs that is 20 percent larger than the number of IHEs required in the analytic sample: because the number of recipients per IHE in the S-STEM population ranges from approximately 10 to 60, it is likely that the response rate of recipients within an IHE will be zero for some proportion of sampled IHEs; we estimate that this will occur in approximately 20

[17] As discussed in Supporting Statement Part B, Section B.1, there are only 75 unique two-year S-STEM awardees with eligible recipients. At those two-year IHEs with more than the minimum number of recipients per IHE needed, we will select a sample of recipients and from the remaining two-year IHEs (i.e., those without at least the minimum number needed per IHE) we will include the census of recipients.

percent of the sampled IHEs (Exhibit 5 shows the number of S-STEM IHEs with at least 10, at least 15, at least 20, at least 25, or at least 30 S-STEM recipients). As a result the initial number of IHEs sampled is 20 percent larger than the number needed in the final (analytic) sample.

Exhibit E.6: Among four-year IHEs, N of recipients per IHE and N of IHEs needed in the final and initial samples for MDE of .075, compared to the number of four-year IHEs with at least the minimum number of recipients per IHE, and the percentage of the S-STEM population of IHEs and recipients represented

N of recipients per IHE needed in the final sample[a]	N of IHEs needed in the final sample[b]	N of recipients per IHE needed in the initial sample	N of IHEs needed in the initial sample	N of IHEs with the minimum number of recipients needed in the initial sample	Percentage of S-STEM IHE population	Percentage of S-STEM recipient population
10	202	21	245	234	77%	94%
15	**138**	**31**	**166**	**170**	**56%**	**83%**
20	103	42	124	135	45%	74%
25	84	52	101	107	35%	65%
30	69	63	83	79	26%	54%

Exhibit reads: To produce an MDE of .075, 15 recipients per IHE in each of 138 IHEs are needed in the final sample. To achieve this final sample an initial sample of 31 recipients per IHE in each of 166 IHEs would be needed. There are 170 S-STEM IHEs in the population that have at least 31 recipients; this population of 170 IHEs represents 56 percent of all S-STEM awardee IHEs (in the 2006-2010 award cohorts) and these 170 IHEs have funded 83 percent of the population of S-STEM scholarship recipients.

Notes:

[a] Ns in this column match those shown in Exhibit E2, column 4

[b] Ns in this column match those shown in Exhibit E2, column 2

As shown in Exhibit E6, to achieve MDE of .075 (for the analysis of the effect of S-STEM on recipients awarded scholarships by four-year IHEs) 15 recipients per IHE from each of 138 IHEs are needed in the final analytic sample; the corresponding initial sample sizes are 31 recipients per IHE at each of 166 IHEs. There are 170 four-year IHEs in the 2006-2010 S-STEM awardee population of IHEs that have at least 31 scholarship recipients. These 170 institutions represent 56 percent of all four-year S-STEM awardees and collectively have funded 83 percent of S-STEM scholarships awarded to students at four-year IHEs. Note that selecting a smaller number of recipients per IHE (21) would require selecting a larger number of IHEs (245) than exist in the population (just 234 four-year IHEs have at least 21 recipients).

Estimates of Differences

Analyses comparing S-STEM recipients' outcomes to those of a nationally-representative comparison group of BPS respondents will be conducted separately for two- and four- year IHEs. The first step of the analysis is to create a matched comparison group of students using the respondents of the BPS survey. Matching will be done within each IHE from which Recipient Survey data are collected. The next section details the propensity score matching process; this is followed by an explanation of procedures to estimate differences from this matched sample.

Propensity Score Matching

Propensity score matching deals with selection bias by explicitly balancing the observable differences between program participants and non-participants and constructing matched treatment and comparison groups that are then used to estimate the effects of the program. Propensity score estimators are valid under the "conditional independence" assumption, which states that the assignment status of a participant or a non-participant (to the treatment or comparison condition) is "ignorable" conditional on his/her propensity score. In other words, propensity score matching relies on the statistical equivalence of matched treatment and comparison groups conditional on their observable characteristics. The major threat to the validity of propensity score estimators, therefore, comes from the existence of unobservable characteristics that affect both outcomes of interest and an individual's assignment status. One way to deal with the threat of unobservable characteristics is using as many "relevant" observable characteristics as possible in the propensity score matching process, so that the effect of these factors is reduced. In this study, we will employ an extensive set of matching variables (see Exhibit E7). Assuming that the processes by which students were selected to receive S-STEM scholarship were also based on the same information, this approach should account for most of the inherent differences between recipients and BPS comparison and minimize the selection bias. However, one of the limitations of using an already administered survey is that they do not ask the information in a form that you require. For example, information on financial need and academic performance prior to the receipt of an S-STEM scholarship are two of the important types of matching characteristics and BPS has only a few measures of these matching variables. This could possibly lead to biased estimates. PSM analysis will be performed using the following four steps:

Exhibit E.7: List of matching characteristics	
Financial aid	
	Received Federal Stafford Loan
	Received Pell Grant
	Received school grant/scholarship
	Received State grant/scholarship
	Received any other financial aid for education
Academic information	
	SAT I math score
	SAT I verbal score
	ACT composite score
	Cumulative GPA (or an estimate of GPA) through the end of the first school year
Demographic characteristics	
	Gender
	Age
	Race and Ethnicity
	Citizenship
Other characteristics	
	Type of degree (Associates or Bachelor's degree)
	Major (Current field of degree)
	Full-time enrollment status

Step 1: We will identify a set of characteristics, measured prior to the treatment group's receipt of S-STEM scholarship funding (i.e., called pre-treatment characteristics) that will be used in the propensity score model to match S-STEM recipients to BPS respondents. These characteristics include variables that likely are associated both with the likelihood of receiving an S-STEM scholarship (e.g., financial aid received for the first year of enrollment; SAT or ACT college admissions test scores) and with the outcomes of interest (e.g., persistence to degree attainment). S-STEM scholarship recipients selected by the awardee institution must be US citizens or permanent residents who are enrolled full time in a program leading to an associate or baccalaureate degree in a STEM discipline;[18] selected students must demonstrate financial need and academic potential or ability. Therefore, pre-treatment characteristics such as SAT/ACT scores, types of financial aid received, college credit for high school coursework, and first year GPA (if prior to receipt of an S-STEM scholarship), will be used as matching variables. These data will be obtained from survey data (the Recipient Survey and BPS extant survey data).

Step 2: Using these pre-treatment characteristics, we will fit a logistic regression model that predicts the probability of being awarded a STEM scholarship . We will then use the coefficients from this model to estimate, for each individual-- including each BPS

[18] Students enrolled for a graduate degree in a STEM discipline are also eligible for an S-STEM scholarship but are not included in the proposed evaluation.

respondent--a "propensity score," which represents the probability of receiving an S-STEM scholarship. Next, **within each IHE**, we will identify and exclude from further analyses those individuals for which no credible match from the other group can be found (that is, any S-STEM recipient for whom there are no credible matches in the BPS respondent group within that IHE will be excluded from analysis; and vice versa, any BPS respondent for which there are no credible S-STEM recipient matches will be dropped).[19]

Step 3: **Within each IHE** we will use the estimated propensity scores to create matched sets of S-STEM recipients and BPS respondents. There are a variety of techniques available for using propensity scores to create such matched sets including matching, stratification, weighting, and regression adjustment.[20] We will use stratification (also called interval matching) as our primary method, which entails constructing a number of propensity score strata **for each IHE** by dividing all treatment and comparison group members who are in the common support into subgroups of equal size based on the propensity scores. We will use five subgroups or strata, which is considered the standard practice (Rosenbaum & Rubin, 1983). We have chosen this method as 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: Finally, we will test whether there are any differences between the S-STEM recipients and corresponding "matched" BPS respondents within each propensity score strata **for each IHE**. There are several ways of performing this analysis. One way is using a t-test for each pre-treatment characteristic used in the propensity score estimation.[21] Another is using an F-test to jointly test whether the S-STEM recipients are similar to the "matched" BPS respondents in each propensity score stratum for

---

[19] More technically, those individual who fall outside of the "area of common support," the range of common propensity scores across S-STEM recipients and BPS respondents within that IHE will be excluded from analysis. Enforcing the criterion of common support is important to ensure the similarity of the matched recipients to non-recipients (Rosenbaum and Rubin, 1983; Caliendo and Kopeinig, 2008).

[20] 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.

[21] 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.

each IHE which takes the correlation between the matching characteristics.[22] 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), we will supplement them using standardized differences.[23] The standardized difference of a matching characteristic between S-STEM recipients and corresponding "matched" BPS respondents in a given propensity score stratum is calculated using:

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

Where:

X	denotes the variable of interest;
S	denotes the stratum;
T	denotes the treatment group,
C	denotes the comparison group;
$\bar{X}_{T,S}$ and $\bar{X}_{C,S}$	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.

We will consider standardized differences larger than 0.15 as suggestive evidence of treatment-comparison group unbalance with respect to the corresponding variables. If we find that statistical balance is not achieved across treatment and comparison groups in each stratum for each IHE, we will modify the logistic model used in Step 2 by including interactions and higher-order terms of the unbalanced characteristics and repeat Steps 2 through 4 until satisfactory balance is achieved.

---

[22] 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.

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

We will present the results of analyses conducted in each step such as the estimated logistic regression coefficients (with standard errors and p-values) in Step 2; histograms of the estimated propensity scores (overall) in Step 3; and results of the balance tests (p-values from the t- and F-tests and the standardized differences) in Step 4. As mentioned before, these analyses will be conducted separately for S-STEM recipients in 2 year colleges and 4-year colleges.

Estimation of Differences

Following the matching, we will estimate the effect of the S-STEM program separately for recipients in 2 year colleges and 4 year colleges by comparing S-STEM recipients' outcomes to those of their comparison group to determine what S-STEM recipients' expected outcomes would have been had they not received the scholarship.

After creating the propensity score strata, we will use a multivariate regression model to estimate the effect of S-STEM program. This regression model will employ a number of matching characteristics and other control 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" estimate since they will have been used twice: both in the propensity score model and in the estimation of effect sizes.[24] We will use the following regression model to estimate the program effects:[25]

$$Y_{ij} = \beta_{0j} + \sum_{k=1}^{4} \beta_{(k)j} S_{ij}^{k} + \beta_{5j}(trt_{ij}) + \sum_{k=1}^{4} \beta_{(k+5)j} trt_{ij} S_{ij}^{k} + \sum_{n=1}^{N} \beta_{(n+9)j}(X_{ij}^{n} - \overline{X_{.}^{n}}) + \varepsilon_{ij} \qquad (1)$$

Where:

Y_{ij} is the outcome measure of the ith student in the jth IHE, at the end of the study;

trt_{ij} is the treatment indicator for the ith student in the jth IHE (1=treatment, 0= comparison group);

S_{ij}^{k} is the indicator (dummy) variable for the kth propensity score stratum in the jth IHE. The model includes the total number of strata (5) minus one strata indicators (k=1,2 ,…, 4). The last stratum is the reference stratum and a dummy for this stratum is not included in the model;

---

[24] 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.

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

$(X_{ij}^{n} - \overline{X^{n}})$ is the n^{th} (n=1,2,…,N) covariate measure for the ith student in the jth IHE that are grand mean centered;

ε_{ij} is the student -level residual of the ith student in the jth IHE. The assumed distribution of these residuals is normal, with mean = 0, and variance = σ^2 when the outcome is continuous.

Interpretation of the parameters in the model is as follows:

$\hat{\beta}_{0j}$ is the covariate-adjusted mean value of the outcome for the comparison group in the reference propensity score stratum in the jth IHE,

$\hat{\beta}_{0j} + \hat{\beta}_{(k+1)j}$ (k=1,2,…,4) is the covariate-adjusted mean value of the outcome for the comparison group in the k^{th} stratum in the jth IHE,

$\hat{\beta}_{5j}$ is the estimate of the effect of S-STEM program for the reference stratum in the jth IHE (i.e. the difference between the mean value of the outcomes of the S-STEM recipients and the BPS comparison group in the reference stratum),

$\hat{\beta}_{(k+5)j}$ (k=1,2,…,4) is the difference between the effect of S-STEM program for the kth stratum in the jth IHE and the effect of S-STEM program for the reference stratum in the jth IHE ,

$\hat{\beta}_{5j} + \hat{\beta}_{(k+5)j}$ (k=1,2,…,4) is the effect of S-STEM program (i.e., the covariate adjusted difference between the outcomes of the S-STEM recipients and the BPS comparison group) for the k^{th} stratum in the jth IHE, and

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

Overall treatment effect

As seen, the model in Equation 1 allows for the estimation of separate treatment effect estimates for each propensity score stratum. More specifically, $\hat{\beta}_{5j} + \hat{\beta}_{(k+5)j}$ (k=1,2,…,4) is the difference estimate for the k^{th} (k=1, 2,…, 4) stratum in the j^{th} IHE. In order to calculate an overall treatment effect estimate, the stratum-specific estimates are aggregated as follows[26]:

[26] Stratum-specific treatment effect estimates can be aggregated to yield an overall impact difference 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).

$$TE = \sum_{j=1}^{J} P_j \left(\sum_{k=1}^{4} P_{kj} \left(\hat{\beta}_{5\mathrm{j}} + \hat{\beta}_{(\mathrm{k}+5)\mathrm{j}} \right) + P_{5j} \hat{\beta}_{5\mathrm{j}} \right) \quad (2)$$

where

P_{kj} is the proportion of treatment group members in the kth stratum in the jth IHE (i.e., n_{kj}/N_j where N is the total number of treatment students in the jth IHE and nkj is the number of treatment student in the kth stratum in the jth IHE) , and

P_{5j} is the proportion of treatment group members in the reference stratum in the jth IHE.

P_j is the proportion of treatment group members in the jth IHE (i.e. n_j/N, where n_j is the total number of treatment students in the jth IHE and N is the number of treatment student.

The overall covariate-adjusted mean for the control group is:

$$\bar{Y}_{AdjControl} = \sum_{j=1}^{J} P_j \left(\sum_{k=1}^{b-1} P_{kj} \left(\hat{\beta}_{0j} + \hat{\beta}_{(k+1)j} \right) + P_{rj} \hat{\beta}_{0\mathrm{j}} \right) \quad (3)$$

The overall covariate-adjusted mean for treatment group is:

$$\bar{Y}_{AdjTreatment} = \bar{Y}_{AdjControl} + TE \quad (4)$$

And the standard error of the Treatment Effect is:

$$\text{Std Error}(TE) = \sqrt{P^T VCV(\hat{\beta})P} \quad (5)$$

Where

P is a 5x1 vector that holds $P_j * P_{kj}$ (j=1,2,…,5), and

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

Estimated coefficients from the regression model and the overall treatment effect estimates will be presented along with corresponding standard errors and p-values. Hence, for dichotomous outcomes, estimates will be presented in the form of percentage points, whereas for continuous outcomes, overall estimates in “effect size” units (e.g., Hedges’ g) will be presented. The effect size is calculated as:

$$ES = \frac{TE}{PooledSD} \quad (6)$$

Where

TE is calculated as shown in Equation 2, and

$$PooledSD = \sqrt{\frac{(N_{tj}-1)S_{tj}^2 + (N_{cj}-1)S_{cj}^2}{(N_{tj}-1)+(N_{cj}-1)}} \quad (7)$$

Where

N_{tj} = sample size of treatment group in the jth IHE,

Ncj= sample size of comparison group in the jth IHE,

S_{tj}^2 = variance of the outcome for treatment group (unadjusted) in the jth IHE; and

S_{cj}^2 = variance of the outcome for comparison group (unadjusted) in the jth IHE.