Appendix G: Non-Response Bias Analysis

Appendix G provides details on addressing any unit or item non-response to the recipient survey.

There are two types of missing data that can arise in a survey, even after repeated attempts to collect data: (1) unit non-response, and (2) item non-response. Our approach to dealing with each of these is described below.

Unit Non-Response

Unit non-response occurs when an entire data instrument is not received from a potential respondent. The expected non-response rates for this evaluation are greater than 10%, hence we will examine for bias because of unit non-response by first examining the response rates overall, as well as by year and by relevant subgroups (e.g., cohort year, or by gender and race/ethnicity). Large differences in the response rates by year and for subgroups could indicate that potential biases may exist[1]. For example, if the response rate from women was very low and women were less likely to belong to the treatment group (since scholarships are given to STEM majors), then any difference in the outcomes between the treatment and comparison groups could result in a biased estimate of the effect of the treatment (measured as the difference between the treatment and the comparison group).

To address unit non-response, we will estimate the probability of a person responding to the survey both for responding and non-responding individuals, as function of baseline characteristics that are available for both types of individuals (e.g. cohort year, gender, race/ethnicity), and create weighting classes for adjusting the weights of responding individuals to alleviate the bias due to non-response, which is a commonly utilized approach. The following steps will be taken to accomplish this task.

Step 1: Fit Models

To estimate the probability of a person responding to the survey we will use logistic regression models, where the response (dependent) variable is a dummy variable that takes the value “1” for responding individuals and take the value “0” for non-responding individuals. Independent variables will include the demographic variables available in the S-STEM Monitoring System. These variables include: gender; race/ethnicity; disability status; type of degree program in which the recipient was enrolled (Associate’s or Bachelor’s); class year (first year, sophomore,

[1] Note that a large non-response rate does not necessarily create bias. For example, if the non-respondents were similar across the treatment and comparison group, then the difference estimate would not be biased necessarily; rather, any effect of the program could not be generalized to the non-respondents (i.e. it would create an external validity problem but not necessarily an internal validity issue).

etc.); first year S-STEM scholarship was received; grade point average; intended major field of study; and scholarship amount.

Several models will be fit to identify the set of explanatory variables that have statistically significant associations with the dependent variable ($p<0.20$ criterion) after controlling for other statistically significant control variables. This will be accomplished by using backwards elimination with forward checking.[2] In this method, all of the explanatory variables are entered as predictors in the logistic regression model. The explanatory variable with the largest non-significant value is dropped from the subsequent model. This step is repeated until the only explanatory variables that remain in the model are those that meet the $p<0.20$ criterion. In the forward checking step, each of the previously eliminated control variables are checked by adding each one to the model with only the significant predictors. In this step, each variable has a chance to get back into the model. The final model that results from this process is used to calculate individual response propensities.

[2] Backwards elimination methods are attractive from the point of view that they are often used and familiar. But use of this method using the conventional $p<0.05$ criterion has been criticized from the point of view that the selection criteria tend to favor covariates with strong relationships to the outcome, but may omit important confounders (i.e., variables that have a weaker relationship to the outcome, but have a strong relationship to the predictor variable of interest). Maldonado and Greenland (1993) evaluated a backwards elimination strategy and a change-in-estimate strategy using simulated data from a Poisson regression model. They found that the p-value based method performed adequately when the alpha levels were higher than conventional levels (0.20 or more), and found that the change-in-estimate strategy performed adequately when the cut point was set to 10 percent. However, their data, generated from a Poisson model, and their analysis model, with only a single covariate in addition to the key exposure variable, are very different than the models anticipated for our current purpose. Budtz-Jorgensen et al. (2007) compared several covariate selection strategies including backwards elimination and change-in-estimate. They looked at the backwards elimination strategy with three p-value cut-off levels, 0.05, 0.10, and 0.20, and, following the recommendation of Maldonado and Greenland (1993) used a 10% criterion for the change-in-estimate method. They found that, although the change-in-estimate strategy did an adequate job of identifying confounders and keeping them in the model, it sometimes threw out variables that were correlated with the outcome, but were not confounders. Therefore, this method threw out variables that, if retained, would have reduced the residual error and reduced the standard error of the exposure coefficient (thus increasing the power to detect exposure effects – exposure effect is analogous to our key predictor of interest). Although they found that backwards elimination with a $p<0.05$ criterion was un-suited for confounder identification, they found that when the p-value criterion was set to $p<0.20$, backwards elimination strategy resulted in a reduction of residual error variance and did not throw out important confounders. They recommended the backwards elimination strategy with a $p<0.20$ criterion over the change-in-estimate strategy.

Budtz-Jorgensen, E., Keilding, N., Grandjean, P., Weihe, P.. (2007). Confounder selection in environmental epidemiology. Assessment of health effects of prenatal mercury exposure. Annals of Epidemiology 17(1); 27-35.

Maldonado, G., Greenland, S., (1993). Simulation study of confounder-selection strategies. American Journal of Epidemiology 138(11); 923-936

Step 2: Use Model Results to Calculate Response Propensities

In Step 2, parameter estimates obtained from the fitted model in step 1 will be used to calculate the predicted probability that an individual will respond to the survey. The logistic regression model is represented as:

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \sum_k \beta_{ki}\ , \quad \text{(Eq. 1)}$$

where π_i is the probability that person *i* is a responder, given the *k* explanatory variables in the final model. The predicted probabilities will be obtained by solving Equation 1 for π_i, and substituting the parameter estimates (i.e., the values of β_k) from the fitted model in place of the parameters. The solution for the predicted probability for person *i* is given by:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \sum_k \hat{\beta}_{ki})}{1+\exp(\hat{\beta}_0 + \sum_k \hat{\beta}_{ki})}$$

Each person's predicted probability of response ($\hat{\pi}_i$) is called their "response propensity". Individuals with similar response propensities have similar characteristics (explanatory variables). In particular, they are similar on the characteristics that are most related to the probability of response.

Step 3: Group Individuals with Similar Response Propensities into Weighting Classes

In this step, individuals with similar response propensities will be grouped into weighting classes. Weighting classes will be formed to ensure that all individuals within a class fall within a narrow range of response propensity scores. The boundaries for the weighting classes will be determined by creating approximately equal-interval response propensity score groupings.

Step 4: Within Weighting Class, Inflate Weights of Responding Individuals to Sum to Population Total

The weights of responding individuals within a class will be inflated so that the responding individuals within the class represent the population of responding and non-responding individuals within that class. That is, because the non-responders within a given class are statistically similar to the responders, the weighting procedure increases the degree to which the responders "represent" the class as a whole.

Within each weighting class, the weights of all individual (both responders and non-responders) will be summed (since we took a census of applicants each individual has an original weight of 1). Next, the weights of just the responding individuals will be summed. Then, within each weighting class, new, adjusted weights of responding individuals will be calculated by multiplying the initial weights by a factor equal to the ratio of the sum of the weights of all individuals to the sum of the weights of the responding individual. The adjusted weight for the i^{th} person in the j^{th} weighting class is represented symbolically by:

$$w_{ij}^{adj} = w_{ij} * \frac{\sum_{k \in responders\,\&\,nonresponders} w_{kj}}{\sum_{i \in responders} w_{ij}},$$

where w_{ij} is the initial sampling weight for the i^{th} person in the j^{th} weighting class, the summation in the numerator is over all *k* individuals in the set of responders and non-responders within weighting class *j*, the summation in the denominator is over all *i* individuals in the set of responders in weighting class *j*, and there are $j = 1,\ldots,$ n weighting classes. The new, adjusted sampling weights sum to the population total. This result can be written symbolically as:

$$\sum_j \sum_{i \in responders} w_{ij}^{adj} = \sum_j \sum_{k \in responders\,\&\,nonresponders} w_{kj}$$

Item non-response

Item non-response refers to one or more specific uncompleted items on an otherwise completed/returned questionnaire. When the amount of missing data on an individual item is modest (across all returned surveys), we will calculate statistics on only the non-missing items, which is equivalent to an assumption that missing data on an item are missing completely at random. The amount of missing data for each item will be presented in all tables/figures included in reports.

Where necessary for analyses of differences, we will take distinct approaches to imputing values depending on whether data are missing for an item used to construct a covariate or predictor variable, or an outcome variable. For analyses where missing data on covariate or predictor variables require imputation to prevent having to omit those respondents from the analysis, we will use a “dummy-variable” method. This method entails (i) creating a dummy variable that equals “1” if the value of the variable is missing and “0” otherwise, (ii) adding the dummy variable to the model as a covariate, and (iii) replacing the missing value of the original variable with any constant, such as zero or the mean for non-missing cases.

If the missing data occurs in an item used to construct an outcome—that is, one of the primary outcomes of interest that we have specified above (for example, the post-fellowship number of publications produced with a foreign co-author)—we will impute values if more than 20% of respondents have missing values. We will use the multiple stochastic regression imputation approach recommended by Puma et al 2009.[3] In this multiple imputation approach, instead of generating one set of values to replace the missing outcome, we generate multiple sets of

[3] Puma, M.J. Olsen, R.B., Bell, S.H., and Price, C. (2009). What to Do When Data Are Missing in Group Randomized Controlled Trials. U.S. Department of Education, Washington DC. NCEE 20090049 (available at http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20090049).

imputed values, for example 10 sets of values. In this procedure, first, predicted values (to replace the missing values) are generated from an OLS regression that is estimated with data that is available for all individuals (respondents and non-respondents). Then, to each predicted value, we add a randomly-selected residual from the OLS regression, to account for the inherent uncertainty in predicting missing data—this comprises the first set of imputed values. Ten sets of such predicted outcome values are generated, each by adding a randomly selected set of residuals from the OLS regression. Next, the difference estimate is calculated using each of the ten datasets in which missing data were replaced with regression-predicted values with the random residuals. That is, we calculate ten different estimates of the effect of the program on the specified outcome; each difference estimate has used one of the ten datasets in which missing data were replaced with the predicted value plus residual. The final difference estimate (that is, the estimate of the effect of the program on the outcome) is the mean of the ten individual estimates. The multiple imputation method is preferred over a single imputation method because the single imputation tends to understate the true variability in the imputed variable and leads to underestimated standard errors.