Supporting Statement for Paperwork Reduction Submission: Part B, Collection of Information Employing Statistical Methods

Program Evaluation of the Partnerships for International Research and Education (PIRE) Program

Introduction

The National Science Foundation (NSF) requests that the Office of Management and Budget (OMB) approve, under the Paperwork Reduction Act of 1995, a three-year clearance for original data collection to be used in the **Evaluation of the Partnerships for International Research and Education (PIRE) Program**. The new data collections include a Principal Investigator Survey (Appendix B); a Postdoctoral Survey (Appendix C); a Graduate Student Survey (Appendix D); an Undergraduate Survey (Appendix E); a Foreign Senior Investigator Survey (Appendix F); and an Institutional Representative Survey (Appendix G).

The PIRE Program

The PIRE program supports U.S. researchers who wish to pursue a research agenda in collaboration with one or more international research partners. The program funds projects across a broad array of scientific and engineering disciplines, in an effort to catalyze long-term, sustainable international partnerships for collaborative research and education that will prepare a cadre of students and early-career researchers for strong leadership and engagement in global science and engineering. PIRE awards support intellectually substantive collaborations between U.S. and foreign researchers in which the international partnership is essential to the research effort. PIRE grantees must provide educational and professional development opportunities for U.S. based researchers, especially those early in their career, including postdoctoral fellows and graduate and undergraduate students. A significant aspect of the opportunity for U.S. participants is on-site research experience at an international laboratory, institution or research site, whether university-, industry- or government-based. PIRE funding may be used only to support U.S. participants in the research and educational activities, including international travel; foreign partners in PIRE are expected to secure and contribute their own funding (for example from NSF's counterpart agencies in other countries). However, Principal Investigators who apply for PIRE awards must include letters of support from their proposed foreign partners and evidence that these partners have access to their own independent funding to support their contributions to the research agenda.[1] A logic model of the PIRE program is included in Appendix A.

PIRE awards last five years. Classes are funded in an alternating two-year, three-year cycle so that two classes are active at any given time. The first class of funded projects was 2005, so their funding ended in 2010 to 2011; the second class was funded two years later, and ended in 2012. However, there have been some no-cost extensions. The third class was funded in 2010 and the fourth in 2012.

[1] NSF uses "Principal Investigator" and "co-Principal Investigator" to refer to U.S.-based award recipients; these terms do not apply to foreign personnel who may play a role on a particular project commensurate to that of a PI.

As of June 2013, across the 2005, 2007, 2010 and 2012 cohorts, PIRE has made a total of 59 awards. These projects range in size from relatively small, bi-national consortia (e.g., two U.S. and two non-U.S. institutions in one foreign country) to large, multi-national, multi-institutional awards (e.g., a dozen U.S. institutions and eleven non-U.S. institutions representing eight foreign nations). Many are multi-disciplinary, combining, for example, the expertise of econometricians with researchers in fluid dynamics; and, notably, many feature partnerships between academic and industrial or non-profit institutions. Collectively, these 59 PIRE projects have provided research and educational opportunities for more than 100 postdoctoral fellows, more than 625 graduate students and approximately 600 undergraduates. More than 600 U.S.-based and over 400 foreign-based faculty and researchers at university and non-academic institutions have participated in one or more PIRE-funded collaborations.

The PIRE program has the following objectives: (1) to promote opportunities for U.S. scientists and engineers to engage in international collaborations that enhance research excellence; (2) to provide international research and educational experiences for U.S. students and faculty that will prepare the U.S. science and engineering workforce for global engagement; and (3) to strengthen the capacity for U.S. researchers and institutions to build and sustain international partnerships. Beginning with the 2012 cohort of PIRE awardees, the program tailored its first objective (to support international collaborations that enhance research excellence) to focus on international partnerships that would support excellence in science, engineering and education to "inform the societal actions needed for environmental and economic sustainability and sustainable human well-being" (NSF, 2010).

There have been several notable changes to the program since its inception. In the 2005 and 2007 PIRE cohorts, budgets were capped at $2.5 million, but beginning with the 2010 cohort, these budget limitations were removed (NSF, personal communication). In addition, starting with the 2010 cohort, grantees had to propose a project of sufficient scope that its effects would extend beyond an individual PI's research group to the participating U.S. institutions by strengthening their capacity for sustained international engagement. Most recently, the 2012 PIRE competition was limited to projects that proposed international research and education partnerships to address the NSF-wide priority in Science, Engineering, and Education for Sustainability (SEES; NSF, 2011). NSF also suggested a host of potential partner agencies, both domestic and foreign, that could provide additional funding. For example, NSF's PIRE and USAID's PEER programs are jointly funding collaborations between U.S. investigators and their counterparts in developing countries where science and engineering capacity is emerging (National Academy of Sciences, 2014).

Overview of the Evaluation of PIRE

The evaluation of PIRE will examine the quantity and quality of research produced by the PIRE program and its participants; measure the research and career outcomes for PIRE participants; document how PIRE is perceived as changing the way U.S. institutions support, manage, or help implement international research and educational collaborations; and explore how PIRE research has made both intended and unanticipated contributions to research and education in environmental and economic sustainability. The evaluation is also designed to capture any promising practices or lessons learned about the implementation of PIRE projects. In summary, the evaluation will address the seven research questions shown in Exhibit B.1.

The study will measure outcomes at both the project (RQs 1, 2, 5, and 7) and participant levels (RQs 3, 4, and 6) and will employ project- and participant-level comparison groups as follows:

1. A matched comparison group of non-PIRE, NSF-funded projects that *do not require* an international collaboration, but which are similar to PIRE awards along other key criteria (see Section B.1 for details of how this comparison group will be constructed). From this program-level comparison group, three participant-level comparison groups will be formed (details on construction of these participant-level groups are provided Section B.1):
   a. PIs and co-PIs of comparison projects, matched to corresponding PIRE participants;
   b. Postdoctoral researchers in comparison projects, matched to corresponding PIRE participants; and
   c. Graduate student participants in comparison projects, matched to corresponding PIRE participants.
2. A group of respondents to nationally fielded surveys of degree recipients in science, engineering and health fields (the Survey of Doctoral Recipients, SDR; and the National Survey of Recent College Graduates, NSRCG) matched to PIRE postdoctoral, graduate student, and undergraduate student participants.

The evaluation 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 the online surveys of project participants and U.S. institutional officials. Although approval is sought only for the new data collection, our description of the evaluation includes both the extant and original data sources that will be considered in the study. Below is a brief summary of these data sources include (see Section A.1 for more information about the circumstances requiring these data sources).

New Data Collections

New data will come from online surveys conducted with the following groups:

- PIRE Principal and co-Principal Investigators and Principal and co-Principal Investigators in a matched comparison group of projects funded under NSF programs that do not require an international collaboration;
- PIRE postdoctoral researchers and postdoctoral researchers in a matched comparison group of projects funded under NSF programs that do not require an international collaboration;
- PIRE graduate student participants and graduate student participants in a matched comparison group of projects funded under NSF programs that do not require an international collaboration;
- PIRE undergraduate student participants;

Exhibit B.1
Research Questions for the Evaluation of NSF's PIRE program

			Data Sources: Extant			Data Sources: Primary Survey Data						
		Program or Participant Level	**Bibliometric**	**NSF program Data**	**National data (SDR, NSRCG)**	**PI and co-PI**	**Postdoc**	**Graduate student**	**Undergraduate**	**Foreign participant**	**Institutional representative**	**Analyses**
RQ 1.	What is the quantity and quality of the publications that PIRE projects have produced (and that are projected to be produced)?	Program	X	X		X						Descriptive
RQ 2.	How does the quantity and quality of publications produced by PIRE projects—that are required, by definition, to include an international collaboration—compare to the quantity and quality of publications produced by similar NSF-funded projects that *do not require* an international collaboration?	Program	X	X								Relational
RQ 3.	What are the program experiences of PIRE Principal Investigators (PIs and co-PIs), postdoctoral researchers, and graduate and undergraduate student participants? What are the research and educational or career outcomes for these participants?	Participant	X	X	X	X	X	X	X			Descriptive Benchmarking for career outcomes
RQ 4.	How do the research and educational or career outcomes for PIRE PIs, postdoctoral researchers, and graduate student participants compare to those of similar participant groups in similar NSF-funded projects that *do not require* an international collaboration?	Participant	X	X		X	X	X				Quasi-experimental for research outcomes; Descriptive, relational, and benchmarking for career outcomes
RQ 5.	How do international affairs representatives at PIRE PIs' institutions perceive the effects of PIRE on their institutions' policies and practices for supporting international research and educational collaborations?	Program									X	Descriptive
RQ 6.	What are the program experiences of foreign senior investigative partners in PIRE and how do they perceive the effects of PIRE on their research and educational practices and those of their institutions?	Participant		X						X		Descriptive
RQ 7.	How have PIRE projects contributed to research that may inform global societal challenges?	Program				X				X		Descriptive

- PIRE foreign senior research partners (i.e., scientists and engineers at the level equivalent to a U.S. Principal or co-Principal Investigator); and
- Institutional representatives (administrators in international affairs offices) at the institutions of PIRE lead PIs (or other official who is familiar with the institution's policies and practices for international research and education).

Extant Sources

- Extant NSF data from PIRE and comparison projects' proposals, award letters, budgets, and annual and final reports to NSF;
- Extant data from the biennial Survey of Doctoral Recipients (SDR) and National Survey of Recent College Graduates (NSRCG); and
- Bibliometric data from the *Web of Science* databases maintained by Thomson Reuters.

Program-level questions about PIRE publication quantity and quality will draw on extant NSF program data and bibliometric data. Descriptive analyses (e.g., counts and citation indices of publications produced by the PIRE projects) will address RQ 1. To address RQ 2, a relational analysis will explore associations between these bibliometric indicators and the presence or absence of PIRE's signature characteristic—namely, the program's requirement that projects include an international collaboration as a required component. A comparison group of non-PIRE, NSF-funded projects in which an international collaboration is not required (but which may arise despite no formal program requirement) will be used in this relational analysis.

Survey data and descriptive analyses will address the program experiences of PIRE participants (RQ 3). The study will compare career outcomes of PIRE postdoctoral, graduate and undergraduate participants (as reported in these surveys) to benchmarks from nationally representative data from the SDR and the NSRCG. In addition, the study will compare the career outcomes of PIRE participants to those of similar participant types in a matched comparison group of non-PIRE, NSF-funded projects (RQ 4). This comparison will use relational analyses with controls for characteristics that could be correlated with career outcomes.

To measure research outcomes for PIRE participants (RQ 3 and RQ 4), the study will obtain extant data for publications that are indexed in bibliometric databases; primary survey data will provide information about publications that are in progress (e.g., in press, under review) and other research products such as conference presentations and software applications. In addition, the study will compare counts and citation indices for publications by PIRE participants to those of matched participants from a comparison group of projects using a quasi-experimental, comparative interrupted time series (CITS) design. For these analyses, participants in the selected comparison projects will be matched to corresponding PIRE participants using a greedy matching distance algorithm.[2] Matches will be based on a pre-participation measure of one of the key research outcomes for the study (e.g., the number of publications prior to project participation; the mean citation measure for publications prior to project participation). Details of the greedy matching distance algorithm and the CITS approach are provided in Section B.3, Estimation Procedures.

[2] Participants will be matched from within each PIRE-comparison project matched pair.

Finally, survey data and descriptive analyses will be used to understand: institutional engagement in implementing PIRE project activities and any changes in policies or practices as a result (RQ 5); foreign senior investigators' role in PIRE, their perception of the benefits and challenges of participation, and any effects of PIRE on their home institutions (RQ 6); and what role, if any, PIRE research has played in advancing knowledge or technologies to address contemporary societal challenges of a global scale (RQ 7).

Rationale for the Evaluation Design

This research design was selected after a determination that neither a randomized control trial nor a regression discontinuity (RD) design was feasible. Because the PIRE program makes awards on the basis of merit, it was not feasible to randomly assign some PIRE proposals to "award" status (i.e., a treatment group) and others to a "non-award" status (i.e., a control group). Moreover, award decisions for past cohorts of the program had already been made when the evaluation was planned. Although a regression discontinuity design was contemplated, PIRE proposals do not receive a continuous score that is compared to an exogenous "cutoff" to make funding decisions, a key requirement for an RD design. Rather, each proposal receives three or more categorical scores, from one (lower merit) to five (highest merit); subsequently, panelists discuss the rated proposals and place each into one of three categories (highly competitive, competitive or not competitive). From these categories, NSF program officers select which proposals to fund, often taking into account additional factors such as geographic and disciplinary variation across the portfolio of funded projects, the institution of the lead PI (e.g., institutions in EPSCOR states).

Another potential quasi-experimental design (other than an RD design) was also deemed inappropriate, namely a design where unfunded PIRE proposals were matched to funded PIRE projects using propensity score matching (PSM) techniques to control for selection bias. For the evaluation of the PIRE program, this method was not feasible. PSM techniques require data on a large number of pre-treatment (in this case, pre-PIRE award) characteristics of the members of the treatment and comparison groups to model selection and develop propensity scores. Data on pre-award characteristics for funded and unfunded PIRE proposals are limited. Moreover, for the PIRE program, it is far from clear what measures *at the project level* exist for project characteristics before the award decision is made—the project does not exist as a measurable entity until the award decision has been made. For example, the PIRE award itself could affect which individuals come together to form the research group that ultimately engages in research, produces publications, and collaborates with foreign partners. In the absence of the PIRE funding, those individuals did not form an already established group for which characteristics can be measured. At the participant level also, it is unclear what group of individuals would form the counterfactual for an unfunded PIRE proposal: without PIRE funding, what individual postdoctoral fellows, graduate students, or undergraduate students could be identified as those who "would have participated" in PIRE and could therefore comprise a valid comparison group? After considering the above evaluation designs, the current approach was adopted as the most feasible design.

B.1 Respondent Universe and Sampling Methods

Across the 2005, 2007, 2010 and 2012 cohorts, NSF has awarded a total of 59 PIRE projects. The study design requires the construction of a comparison group of non-PIRE, NSF projects that *do not require* an international collaboration, but which are similar to PIRE projects along other key criteria. There are an estimated 1,400 to 1,500 comparison projects in the population of possible candidates that match PIRE projects in terms of comparable years of active funding, award duration and award amount. From this population we will select a purposive sample by identifying one or two matches for each of the 59 PIRE awards by using the criteria listed below (see Matching PIRE projects to Comparison Projects).[3] From these PIRE and comparison projects, we will survey individual respondents from the following groups:

- Principal and co-Principal Investigators (PIRE and comparison projects);
- Postdoctoral researchers (PIRE and comparison projects);
- Graduate student participants (PIRE and Comparison projects);
- Undergraduate student participants (PIRE projects only);
- Foreign senior investigators (PIRE projects only); and
- Institutional administrators at lead PIs' sponsor institutions (PIRE projects only).

Before describing the universe and sampling methods for these respondent groups, we briefly describe the criteria for matching PIRE and comparison projects.

Matching PIRE Projects to Comparison Projects

To construct the matched comparison group of NSF-funded projects, candidate comparison projects funded by non-PIRE NSF programs will be identified and matched to each PIRE project included in the evaluation on the basis of the following criteria:

- The PI of the comparison project must not also be (or have been) PI of a PIRE project;
- The comparison project duration (award period) must be within 1 year of the duration of the PIRE project (e.g., if the PIRE award is 5 years in duration, the comparison award must be 4, 5 or 6 years in duration);
- The comparison project award amount must be within 20 percent of the PIRE project award amount;
- The comparison project must be funded by an NSF program that does not *require* an international component (projects in which international collaboration is not required but is encouraged, optional, or not mentioned explicitly meet this criterion);
- The comparison project must have at least one graduate student participant;
- The comparison project must have research as its primary focus (PIRE projects have research as a primary focus);

[3] A preliminary pilot test of this matching shows that for some PIRE awards there will be several NSF projects that match on the specified criteria and some that will have just one NSF project that meets these criteria. For a given PIRE award we do not know, a priori, how many matches will result. For purposes of determining respondent burden and developing sampling plans, we have estimated 1 match for 39 PIRE awards and 2 matches for 20 PIRE awards for an estimate of 79 comparison awards; note however, that the exact N of comparison awards that meet matching criteria is unknown. It is unlikely that a given PIRE award will have more than 3 eligible matches.

- The comparison project must engage expertise in scientific or engineering disciplines that overlap at least partially with those involved of the PIRE project (i.e., the NSF directorate in which the program funding the comparison award is housed must be commensurate with at least one of the NSF directorates that corresponds to the PIRE award and the scientific or engineering disciplines in which comparison project personnel have expertise must be the same as the expertise of at least one member of the PIRE project team);[4] and
- The comparison project must include at least two institutions (U.S. or foreign) total (by definition, a PIRE award includes at least one U.S. and at least one foreign institution).

If multiple matches for a PIRE project are identified, we will select one or two projects at random for inclusion in the comparison group.

Respondent Universe for PIRE and Comparison Project Participants

The target population for each group to be surveyed includes all lead PIs and co-PIs, postdoctoral researchers, graduate and undergraduate student participants from the U.S. and all foreign senior investigators (i.e., researchers at the equivalent level of a co-PI) who have participated in these PIRE projects. (Participant status is based on the time of participation in PIRE activities: the population of PIRE graduate students, for example, includes former graduate students who participated in PIRE while pursuing their graduate degree as well as current graduate students participating in an active PIRE project.) NSF estimates that these 59 PIRE awards include, on average, one lead PI and four co-PIs, for a total of five PIs/co-PIs per project; two postdoctoral researchers; twelve graduate students; eight undergraduate students; and five foreign senior personnel. In addition, for each PIRE project, there is a designated lead institution in the U.S. (i.e., the lead PI's institution) which has an estimated 52 senior institutional representatives (e.g., at the Assistant Dean or Vice Provost level), from which one or two oversee the institution's international research and educational endeavors will be sampled. These figures yield the following estimated populations (see Exhibit B.2):

[4] NSF awards are typically made from programs within divisions that sit within directorates; for programs like PIRE that can cut across programs, divisions or directorates, NSF has, *a priori*, identified which directorates each of the 59 PIRE awards corresponds to—this was done partly to facilitate identification of proposal reviewers with the necessary expertise to judge the merits of the proposal.

Exhibit B.2 Respondent type	PIRE Respondent Universe (estimated)	PIRE respondent census or sample	Comparison Group Respondent Universe (estimated)	Comparison Group respondent census or sample
Principal Investigators (includes lead PIs and co-PIs)	295	295	395	395
Postdoctoral researchers	118	118	158	158
Graduate students	708	484	948	767
Undergraduate students	472	472	n/a	n/a
Foreign senior investigators	295	295	n/a	n/a
Institutional administrators	3,068*	118	n/a	n/a

Notes:
* The population of institutional administrators at the 59 PIRE institutions is estimated from data from IPEDS of 956 public and private, not-for-profit doctoral, research/scholarship and professional practice or master's degree granting institutions of higher education in 2009-10 and a total of 49,640 executive, administrative or managerial full-time equivalent staff in doctoral-degree granting universities with very high research activity in Fall 2009; estimate is an average of 52 such staff at the types of PIRE sponsor institutions multiplied by 59 PIRE awardee institutions. There are no available estimates for the number of such administrators who oversee an institution's engagement in international research and education activities. See: http://nces.ed.gov/datalab/tableslibrary/viewtable.aspx?tableid=7084, retrieved May 9, 2014 and http://nces.ed.gov/datalab/tableslibrary/viewtable.aspx?tableid=7080, retrieved May 9, 2014.

Sample Sizes for PIRE and Comparison Project Participants

We will survey the census of PIRE PIs, PIRE postdoctoral participants, PIRE foreign senior investigators and PIRE undergraduates. We will select a sample of PIRE graduate students, and a convenience sample of institutional administrators from international affairs offices at PIRE lead PIs' institutions. From the comparison group of 79 non-PIRE NSF projects (see below for discussion of how these comparison projects will be identified), we will survey the census of PIs and postdoctoral participants and select a sample of graduate students.[5]

Based on the contractor's attained response rates in studies with similar populations, the target response rates for each respondent group are:

Exhibit B.3	Expected response rates	
Respondent type	PIRE Group	Comparison Group
Principal Investigators (includes lead PIs and co-PIs)	81%	55%
Postdoctoral researchers	81	55
Graduate students	73	46
Undergraduate students	65	n/a
Foreign senior investigators	81	n/a
Institutional administrators	81	n/a

For the target response rates in Exhibit B.3 that are below the threshold at which OMB guidance requires a non-response bias analysis, a plan for non-response bias analysis is included (see Section B.4 and Appendix L).[6]

[5] Although we will survey the census of PIs and postdocs in PIRE and comparison projects, certain analyses may not include all survey respondents: as discussed below, for some analyses, we will match PIRE PIs to comparison project PIs, and PIRE postdocs to comparison project postdocs; this matching process may result in an analysis sample that is smaller than the census for these participant groups.

[6] See Guideline 1.3.4 in the *Office of Management and Budget Standards and Guidelines for Statistical Surveys* (OMB, 2006).

PI Survey: PIRE and Comparison Group Census

The target population for the PI survey includes PIs and co-PIs of the 59 PIRE projects and the PIs and the PIs and co-PIs of approximately 79 comparison projects (i.e., non-PIRE projects matched to each PIRE project). The number of PIs and co-PIs from comparison projects will depend on the actual project(s) selected; for planning purposes, we project the number will be 395. We will survey the census of PIRE and comparison group PIs and co-PIs.

We plan to include four independent groups of other PIRE participants, classified by their status at the time they participated in PIRE activities:

- Postdoctoral researchers;
- Graduate students;
- Undergraduate students; and
- Senior researchers at foreign institutions.

For the postdoctoral and graduate student participant groups, we will also survey corresponding participants from projects in the comparison group.

Postdoctoral Survey: PIRE and Comparison Group Census

The study will survey the census of postdoctoral researchers in PIRE projects (n=118) and comparison projects (n=158).

Graduate Student Survey: PIRE and Comparison Group Samples

From the population of PIRE graduate students we will select a sample of 472 respondents, and from the population of comparison project graduate students we will select a sample of 767 graduate students.

Undergraduate Student Survey: Census (PIRE only)

We anticipate that the target population of undergraduate student participants in PIRE projects is 472 (59 PIRE awards x 8 undergraduates per award). We will survey the census of undergraduate participants in PIRE; if we discover that the population is higher than 472, we will select a sample of undergraduate students (not to exceed 484).

Foreign Senior Investigator Survey: Census (PIRE only)

We will survey the census of 295 foreign senior investigators participating in PIRE at foreign institutions.

Institutional Administrator Survey: Sample (PIRE only)

The study will include a convenience sample of institutional representatives at PIRE projects' lead PIs' institutions who oversee the institution's policies and practices for faculty and students traveling abroad for research or educational purposes. Each of the 59 lead PIs will provide names of up to two such campus administrators for a sample size of 118 (these names will be collected as part of the PI survey).

Selecting a Graduate Student Sample

For the graduate student participant groups (from PIRE and comparison projects) identified above, we will select a stratified systematic sample from strata developed to represent crossclassifications of key characteristics (e.g., award cohort and research field). To ensure that the sample includes variation across key characteristics of participants (e.g., gender, race/ethnicity), participants will be selected systematically using Lahiri's circular method from each stratum (Lahiri, 1951). Prior to sampling, participants within projects will be sorted by gender, race/ethnicity and other agreed upon characteristics. Systematic sampling after sorting by these variables increases the likelihood of having a wide distribution of participants across these characteristics in the selected sample.

The selected sample sizes were based on the need to be able to detect policy-relevant and meaningfully different results for outcomes, whether continuous or dichotomous, for PIRE participants and comparison project participants.[7] Results of power analyses are generally presented as a percentage of the standard deviation of the outcome measures, which are referred to as minimum detectable effect (MDE) sizes. For comparisons of continuous outcomes using the full PIRE analytic sample, these sample sizes give the evaluation the power to detect an MDE of 0.35 standard deviations between PIRE and comparison project participants' reported outcomes. For comparisons of binary outcomes using the full PIRE analytic sample, and varying comparison group means, the study will be powered to detect a minimum difference of 5 and 20 percentage points between PIRE and comparison project participants' reported outcomes. [8]

Sample size calculations were made under the following assumptions:

1. A two-sided statistical test at the standard 5 percent significance level was used ($\alpha = 0.05$);
2. Desire to detect standardized effect sizes of 0.20 in the difference between PIRE participants and similar participants in the comparison group for outcomes of interest;
3. Power to detect these standardized effect sizes set to 80 percent ($1-\beta = 0.80$);
4. Equal sample sizes in each group ;
5. The proportion of variance explained by covariates (including pre-participation measures of key outcomes) was estimated to be 10 percent;
6. Sample sizes were adjusted to account for potential response rates based on previously OMB-approved evaluations of the NSF IRFP (postdoctoral fellowship) and EAPSI (graduate student fellowship) programs.[9]

[7] The sample sizes required have been adjusted to account for potential response rates based on recent evaluations with similar populations: Abt's evaluation of NSF's IRFP program (an international postdoctoral fellowship program) achieved response rates of 81 percent for awardees and 55 percent for non-awardee IRPF applicants; Abt's evaluation of NSF's EAPSI program (for a graduate student international summer research program) achieved response rates of 73 percent for awardees and 46 percent for non-awardee applicants.

[8] See Appendix K for the assumptions behind these power analyses. To provide some context for these differences, note that our anticipated minimal percentage point differences are small enough to detect the largest differences detected in prior evaluations of NSF's IRFP and EAPSI programs (Martinez, Epstein et al., 2012; Martinez, Neishi et al., 2012). In this evaluation, many of the differences between fellows and the comparison group were less than 10 percentage points. For instance, among five different outcomes, the largest difference found between the fellows and the comparison group was 11 percentage points for IRFP. Among six different outcomes, the largest difference found between the fellows and the comparison group was 26 percentage points for EAPSI.

[9] Actual sample sizes for the PIRE evaluation may be higher because participants will have taken part in a PIRE or comparison project more recently than the earliest cohorts of IRFP (1992) and EAPSI (2000) fellows and because the comparison group members were not denied NSF funding for they had applied (in contrast to the comparison groups used in the IRFP and EAPSI evaluations).

See Appendix K for details on the sample selection and power analyses.

Matching PIRE and Comparison Project Participants

To estimate the effects of PIRE on research, career and educational outcomes of PIs, postdoctoral researchers and graduate students, we will match PIs, postdoctoral researchers and graduate student participants in the comparison projects to the census or samples selected from PIRE projects (i.e., all PIRE PIs and postdoctoral researchers are included; a sample of PIRE graduate students will be selected). For each of these three groups, participants will be matched on a pre-participation outcome (e.g. the number of publications prior to project participation or the mean citation measure for publications prior to project participation).

There are many multivariate matching techniques for identifying potential matches for each PIRE project. One frequently used technique, propensity score matching, would require a larger sample of PIRE projects to create reliable propensity scores; in addition, to create propensity scores the evaluation would require a larger number of variables from existing data than are available on participants prior to data collection. Thus, we will use another technique, *greedy matching* to match participants within each PIRE-comparison project pair. Unlike PSM techniques, greedy matching is not sensitive to sample size. Greedy matching begins by (1) randomly sorting the *N* PIRE participants and *M* comparison project participants; (2) then, the first PIRE case in the list is matched to the closest comparison participant—namely, the one most similar to the PIRE participant. Next, (3) the second PIRE participant is matched to its closest comparison participant among those remaining; and (4) this process repeats until all PIRE participants have been matched.

Here, "similarity" between PIRE and comparison participants is defined as the average *distance* per year between the PIRE and comparison participants' score on a pre-test measure of research outcomes.

For example, using the number of publications as the pre-test measure, the average distance per year between the ith PIRE participant and the jth comparison participant is defined as

$$D_{ij} = \frac{\sum_{n=1}^{N} X_n^1 - X_n^0}{N};$$

where X^1= { $X_1^1, X_2^1, X_{3,}^1 \ldots X_N^1$ } is the number of publications across N years for PIRE participants and

X^0= { $X_1^0, X_2^0, X_{3,}^0 \ldots X_N^0$ } is the number of publications across N years for Comparison participants.

For each of the three groups of PIRE and comparison project participants, matches will be strengthened using pre-participation characteristics collected from extant or survey data (e.g., demographic characteristics, year highest degree received, field of research, prior international research or education before participation in the PIRE or comparison project) that are likely to be associated with outcomes of interest (e.g., post-participation career status, professional accomplishments, extent of international engagement). Depending on the number of potential matches identified for each respondent group as well as the availability of data on pre-participation characteristics, we will explore which matching technique results in the strongest match across these characteristics (Bergstrahl & Kosanke, 1995; Bergstrahl, Kosanke, & Jacobsen, 1996; Ho et al., 2007; King et al., 2011).

For quasi-experimental and relational analyses that rely on this participant-level matching, the analytic sample sizes may be smaller than the census for these groups: the matching process may result in some loss of respondents for whom no sufficiently close match can be identified. Nevertheless, for descriptive analyses (Research Questions 3-7), we will use data from all survey respondents; thus a census of these populations is needed.

B.2 Information Collection Procedures

Primary data collection activities include web-based surveys of respondent groups shown in Exhibit B.4. The key steps in the information collection include:

- Obtaining contact information;
- Inviting respondents to complete the survey and follow-up to ensure a high response rate; and
- Protecting respondent privacy and securing collected data.

We describe the data collection procedures for each activity below.

Exhibit B.4

Respondent type	PIRE Group	Comparison Group
Principal Investigators	✓	✓
Postdoctoral researchers	✓	✓
Graduate student participants	✓	✓
Undergraduate participants	✓	
Foreign senior investigators	✓	
Institutional administrators	✓	

Obtaining Contact Information for Respondents

Principal and Co-Principal Investigators (PIRE and comparison group)

To locate survey respondents for the PI survey, researchers will construct a database of current contact information using information available through NSF's extant data, including annual reports in which contact information is typically included for all PIs and co-PIs. The evaluation contractor, Abt Associates (Abt), will also use other location methods such as general web searches to find the most up-to-date contact information for any PI or co-PI who has changed institutions since the most recent data available through NSF program records.

Postdoctoral researchers and graduate student participants (PIRE and comparison group)

To identify current and former postdoctoral and graduate student participants in PIRE and comparison projects, the contractor will extract participant information maintained in NSF's Enterprise Information System (EIS). Because contact information in this system will likely be out of date for former project participants, the evaluation contractor will also use other location methods such as general web searches and the fee-based electronic database, AccurInt. If necessary, the contractor will contact lead PIs to request contact information for individuals in the sample for whom this information is missing (see Appendix M for this invitation).

Undergraduate and foreign senior investigators (PIRE)

To identify current and former undergraduate participants in PIRE, the contractor will extract participant information maintained in NSF's Enterprise Information System (EIS). Because contact information in this system will likely be out of date for former project participants, the evaluation contractor will also use other location methods such as general web searches and the fee-based electronic database, AccurInt.

To identify current and former foreign senior personnel who participated in PIRE, the contractor will extract participant information from annual reports that PIs submit to NSF, which include names, institutional affiliations, and often, contact email addresses for these respondents. For missing contact information, the evaluation contractor will use general web searches and if necessary, the contractor will contact lead PIs to request contact information for individuals in the sample for whom this information is missing. (See Appendix M for the invitations.)

Institutional Administrators (PIRE)

As part of the PI survey, the lead PI of each PIRE project will be asked to identify the name, title, and institutional office of up to two administrative officials knowledgeable about PIRE who help coordinate international research and educational activities at the institution. In the event that a PI does not provide this information, a search of the institution's web page will be used to identify two candidate respondents in the appropriate offices.

Invitation and Follow-Up with Survey Respondents

Respondents included in the census or sample targeted for the evaluation will receive an email inviting them to participate in an online survey (see Appendix M for email invitation). This email will explain the purpose of the study and a link to the web survey. Each survey link will be unique to the selected respondent. This survey link will launch the respondent's web browser and open the survey to an introductory "landing page" that will describe the purpose of the survey, its expected length, and instructions for navigating through the survey. Links to a "frequently asked questions" page will be included along with information about how the respondent's privacy will be safeguarded. The landing page will clearly display the OMB control number and expiration date, along with confirmation that the study has received IRB approval and contact information for potential survey respondents to use if they have questions about the study. Respondents who consent to participate in the survey will be asked to click on a button to launch the survey.

During the two-month survey field period, up to three email reminders and three telephone reminders will be used to encourage survey completion. (See Appendix M for text of reminders.) If desired response rates have not been achieved, the contractor may extend the survey deadline by one to two weeks. Throughout the data collection cycle, a toll-free study telephone number and email address will be available to allow potential respondents to easily obtain answers to questions or concerns about the study. In their survey invitation, PIs will also be informed of the timeline for fielding of the other surveys so that they may encourage participants in the PIRE or comparison project to participate in the survey.

Once approval is received from OMB, the web-based surveys will be programmed for online data collection. The study team will test the programmed surveys to ensure functionality and accuracy of data capture. (See Appendices B through G for copies of the survey instruments.) Wherever possible, items in the proposed surveys are identical to, or adapted from survey items used in other approved studies. Appendix I lists the sources for all survey items.

Data Security and Privacy Protection

Abt Associates, the contractor that will conduct the proposed data collection activities, has conducted numerous studies involving sensitive and non-sensitive information. All project staff members employ both electronic and physical safeguards to protect data from unauthorized access. Electronic project directories, files, and databases are accessible to project staff only and are protected by discretionary access control lists (ACLs), group memberships, passwords, and locking workstations. Access to the data processing area and database servers is limited to authorized personnel. Building security staffs all sites 24 hours, 7 days per week. To protect against data loss, Abt also uses automated, redundant backup procedures and file management techniques to ensure that files are not inadvertently lost or damaged. All data, including the web-based survey data, will be maintained on a secure server with appropriate levels of password protection. Respondent names and other personal identifiers (e.g., email address) saved in raw data files will be removed from analysis files created. Data files compiled into the PIRE evaluation database will include a unique, random study identification number (study ID number) for each survey respondent; names and any other personal identifiers will be removed. (A separate data file linking this identifier to contact information for survey respondents will be created, encrypted-at-rest, and kept by the evaluation contractor until all analyses are completed. Once NSF has received the final report and database, the contractor will destroy the file linking respondent identifying information to the unique study ID number.)

B.3 Estimation Procedures

The research questions in Exhibit B.1 will be addressed using descriptive, comparative, and relational analyses of data collected using interviews, surveys, and extant NSF files. Each of the analytic approaches planned for the study is described below.

Descriptive and Benchmarking Analyses

The following research questions will be addressed using descriptive analyses:

RQ 1. What is the quantity and quality of the publications that PIRE projects have produced (and that are projected to be produced)?

RQ 3. What are the program experiences of PIRE principal investigators (PIs and co-PIs), postdoctoral researchers, and graduate and undergraduate student participants? What are the research and educational or career outcomes for these participants?

RQ 5. How do international affairs representatives at PIRE PIs' institutions perceive the effects of PIRE on their institutions' policies and practices for supporting international research and educational collaborations?

RQ 6. What are the program experiences of foreign senior investigative partners in PIRE and how do they perceive the effects of PIRE on their research and educational practices and those of their institutions?

RQ 7. How have PIRE projects contributed to research that may inform global societal challenges?

In addition, benchmarking will be used to assess the career outcomes of PIRE participants (RQ 3). The analytic approaches to addressing each question are described below.

RQ 1. What is the quantity and quality of the publications that PIRE projects have produced (and that are projected to be produced)?

To describe the *quantity* of publications by PIRE projects, extant data (PI annual reports submitted to NSF) will be used to calculate the mean number and standard deviations (or standard errors) of research reports, journal articles and other products published per PIRE project.[10] Survey data will be used to calculate the number of these products that are in press, under review, or in preparation per PIRE project (see Part A for a description of the surveys, copies of which are included in Part B Appendices B-F). Because publication rates are known to vary (or likely vary) by field, project size (i.e., awarded amount), duration, and year of onset (Durieux & Gevenois, 2010; Pendlebury, 2003), measures of projects' productivity will be calculated overall and by corresponding subgroups. Results will also be calculated for the mean percentage of publications with co-author(s) at foreign institutions, both overall and by subgroups (e.g., relevant fields of research, year PIRE project was awarded).[11]

To describe the *quality* of PIRE publications the evaluation will use bibliometric citation data maintained in the *Web of Science* by Thomson Reuters. Descriptive statistics (i.e., project-level averages and proportions as appropriate) will be presented at the project level for the following measures:

- Mean number of times a project's publications have been cited;
- The mean field-normalized citation impact for a project's publications;[12] and
- The mean percentage of observed to expected citations per publication per project.[13]

[10] For data on the census (i.e., a population), standard deviations will be presented; for data on a sample of a population, standard errors will be presented.

[11] The percentage of publications with a foreign co-author is one measure of international research collaboration for the U.S. authors. For publications prior to 2009, Thomson Reuters does not have indices indicating the institutional affiliations of each author, but only the institutional affiliations associated with each publication. For example, if a publication listed two authors at the same institution but within different departments, the record would list this institution once, not twice.

[12] Because publication and citation patterns vary across disciplines and publication types, normalization techniques account for these patterns to produce a standardized citation index.

[13] This bibliometric indicator likewise takes into account patterns of citations for publications of a particular type, within a particular journal, field, and year of publication to predict the "expected" number of citations for comparison to the actual number of citations received by a given publication.

Because larger (in terms of amount awarded) and older projects have more resources and time to accrue publications and citations, citation measures will be presented overall and by appropriate subgroups.

RQ 3. What are the program experiences of PIRE principal investigators (PIs and co-PIs), postdoctoral researchers, and graduate and undergraduate student participants? What are the research and educational or career outcomes for these participants (hereafter, "PIRE participants")?

Descriptive Analyses

Program experiences of PIRE participants will be summarized in descriptive analyses using means and standard deviations for continuous variables and proportions or percentages for categorical variables derived from the fielded survey items (these surveys are described in Part A and are included in Appendices B-G). Measures of program experiences include, for example:

- The proportion of participants at each level (PI, postdoctoral, graduate, undergraduate) who travelled outside the U.S. for the PIRE project;
- The purpose, frequency, and duration of such international travel;
- Characteristics of participants' interactions with foreign colleagues; and
- Benefits and challenges of participation in international research and education.

Research outcomes for PIRE participants will be examined using descriptive statistics on the quantity and quality of publications produced or in progress, using extant bibliometric data and survey data; these measures are similar to those that will be calculated at the project level (described above under RQ 1). Results reported for each PIRE participant group will include:

- Mean number (and standard deviation or standard error) of research reports, journal articles and other products published per participant;
- Mean number of such products that are in press, under review or in preparation per participant;
- Mean number of times a participant's publications have been cited;
- The mean field-normalized citation impact for a participant's publications;[14]
- The mean ratio of observed to expected citations per publication per participant;[15] and
- The mean percentage of publications with co-author(s) at foreign institutions, per participant.[16]

[14] Because publication and citation patterns vary across disciplines and publication types, normalization techniques account for these patterns to produce a standardized citation index.

[15] This bibliometric indicator likewise takes into account patterns of citations for publications of a particular type, within a particular journal, field, and year of publication to predict the "expected" number of citations for comparison to the actual number of citations received by a given publication.

[16] The percentage of publications with a foreign co-author is one measure of international research collaboration for the U.S. authors. For publications prior to 2009, Thomson Reuters does not have indices indicating the institutional affiliations of each author, but only the institutional affiliations associated with each publication. For example, if a publication listed two authors at the same institution but within different departments, the record would list this institution once, not twice.

As with analyses of project-level publication measures (RQ 1), these descriptive analyses will be conducted overall and by subgroups to take into account factors related to publication and citation accumulation.

The educational and career outcomes of PIRE participants will be examined and summarized using descriptive statistics from data collected via surveys administered to PIRE participant groups (described in Part A). The analyses will summarize study participants' educational and employment outcomes subsequent to their participation in PIRE projects, as well as the proportion of PIRE participants in completed PIRE projects, or former participants in active PIRE projects, who have pursued subsequent international collaborations.[17] For example, the study will report the proportion of PIRE graduate students who completed their graduate degree as well as characteristics of subsequent employment (e.g., academic or non-academic, working in a job that requires (or does not require) education in a STEM field). Together, these descriptions will provide a portrait of the educational and career trajectories of PIs, co-PIs, postdoctoral researchers, graduate students, and undergraduates who have participated in PIRE projects.

Benchmarking Participant Outcomes to Nationally Representative Reference Groups

In addition to the descriptive analyses described above to answer Research Question 3, career outcomes will be also be benchmarked against national estimates.

Reference groups, against which specific outcomes for PIRE participants can be compared, will be created from extant data from the Survey of Doctoral Recipients (SDR) and National Survey of Recent College Graduates (NSRCG) (see Section A.1, Extant Data). Samples from these extant data sources will be selected by identifying a subset of postdoctoral and graduate student SDR (or NSRCG) respondents matched to PIRE sample participants based on characteristics such as degree type (i.e., bachelor's, master's or doctorate), field and date received. Surveys administered to PIRE participants include a subset of items from these nationally administered surveys to allow these benchmarking comparisons.

For PIRE undergraduates who have completed their bachelor's degree, educational and employment outcomes will be compared to data from a nationally representative sample of respondents to the NSRCG. Postdoctoral and graduate student participants' outcomes will be compared to data from a nationally representative sample of respondents to the SDR. These comparisons will use regression models with controls for characteristics that could be correlated with outcomes.

Using the reference group, ordinary-least-squares regression models will be fit to data in which the career outcomes are dependent variables and the independent variables include a dichotomous PIRE indicator (1=PIRE participant, 0=reference group from the SDR or NSRCG), covariates that control for demographic differences (e.g., gender), and other covariates that could be correlated with employment outcomes. Parameter estimates will be presented in tables with accompanying explanations. The coefficient for the PIRE indicator will represent the unique relationship of PIRE to differences in employment outcomes between the PIRE and reference

[17] Former participants in active PIRE projects are those who are no longer participating in the project (e.g., a former PIRE postdoctoral fellow who is now an Assistant Professor and no longer affiliated with the PIRE project).

groups, controlling for other factors included in the model. Descriptive tables will show means and percentages for the outcomes, both overall and by field of discipline and other major subgroups of interest. It is important to note that at these analyses are non-experimental and any differences in outcomes will be viewed as exploratory, interpreted with caution, and presented with discussion of their limitations.

RQ 5. How do international affairs representatives at PIRE PIs' institutions perceive the effects of PIRE on their institutions' policies and practices for supporting international research and educational collaborations?

Research Question 5 will be addressed using descriptive summaries of open-ended, free-response items on surveys administered to administrators in the appropriate international affairs offices at the institutions of PIRE projects' lead PIs (or other official who is familiar with the institution's policies and practices for international research and education). Specifically, qualitative analyses will be used to identify similarities and differences in how institutions support international research and educational collaborations; what institutional changes in educational policies or practices for graduate or undergraduates have occurred as a result of PIRE; what role any pre-existing institutional partnerships with educational institutions in foreign countries have played in the PIRE project; how PIRE has affected these prior partnerships or fostered the creation of new inter-institutional, international partnerships; and any challenges that have resulted from the participation of graduate or undergraduate students in sciences or engineering in education or research abroad, and how the institution has responded to those challenges. Patterns of responses will be examined by institutional characteristics (e.g., Carnegie classification, public or private control) to identify potential relationships between these characteristics and support for international STEM collaborations. Such data may identify an important area for further exploration in future evaluations of PIRE.

RQ 6. What are the program experiences of foreign senior investigative partners in PIRE and how do they perceive the effects of PIRE on their research and educational practices and those of their institutions?

Descriptive statistics (e.g., means and standard deviations or proportions, as appropriate) will be used to summarize responses to surveys administered to foreign senior investigative partners. These analyses will describe the program experiences of foreign PIRE participants, their engagement in planning the PIRE project activities, involvement of foreign graduate or undergraduate-level students from foreign institutions, and visits to the U.S. by these foreign partners; the benefits and challenges of collaborating with U.S. researchers; and effects of PIRE on foreign institutions' support for research and educational collaborations with U.S. or other international colleagues, and foreign senior investigators' perception of the role of PIRE project in addressing global societal issues. For continuous variables, means and standard deviations will be reported; for categorical or dichotomous variables, proportions will be reported. Moreover, we will present these statistics both overall for foreign PIRE participants and by subgroups based on geographical region (e.g., foreign investigative partners from East Asian nations, European Union nations, Central/South America).

RQ 7. How have PIRE projects contributed to research that may inform global societal challenges?

Qualitative analyses of open-ended survey questions will be used to identify similarities and differences in how PIs and foreign senior investigators perceive their projects' contributions to addressing societal challenges. Societal challenges may include those described by investigators themselves and those that NSF has identified, such as research that may contribute to:

- Advances in the extraction and processing of natural resources; energy production, transmission and conservation;
- Mitigating and responding to the effects of climate change on human populations, biodiversity, geophysical and ecological phenomena;
- Research related to sustainable agricultural practices and access to safe food and water;;
- Improved capabilities for forecasting natural disasters and other hazards to human safety and wellbeing, mitigating their effects, and enhancing the capacity to respond and recover.

In its solicitation for the 2012 cohort of PIRE awards, NSF required all projects to address the theme of Science, Engineering and Education for Sustainability (SEES), part of a cross-program effort at the Foundation. Although earlier PIRE cohorts were not required to address this theme, NSF is interested in learning how the international collaborations that the PIRE program supports may facilitate advances in, or new perspectives on, challenges faced by societies worldwide, particularly challenges that may require large-scale, multi-national cooperation. These analyses will particularly help inform NSF's development of future PIRE solicitations.

Summary. Descriptive analyses conducted using extant NSF data, bibliometric data, and survey data will produce descriptive statistics including means and percentages. For items using continuous scales, we will calculate means and standard errors (or standard deviations for data from a census) to describe both central trends and variation across the samples/census. Frequency distributions and percentages will be used to summarize answers given on categorical scales.

In addition, cross-tabulations will be used to illustrate differences in responses between groups or the distribution of responses across subgroups of interest. For example, when examining data on the number of publications resulting from the PIRE award, our analyses will consider the field of study as well as number of years since the start of the PIRE award.

The descriptive analyses will provide NSF with a portrait of PIRE programs, participants, and affiliated institutions. The portrait will detail the average quantity and quality of publications produced by PIRE projects as well as participants' research outputs and educational and career trajectories. It will also present the variation across projects and participants. Causal connections between participation in PIRE and research and career outcomes cannot be made based on these analyses. However, they will be informative nonetheless because they will provide an overview and detailed description of an important NSF funding program and its participants.

Relational and Quasi-Experimental Comparative Analyses

Two research questions will be addressed using relational and/or quasi-experimental comparative analyses:

- RQ 2. How does the quantity and quality of publications produced by PIRE projects—that are required, by definition, to include an international collaboration—compare to the quantity and quality of publications produced by similar NSF-funded projects that *do not require* an international collaboration?
- RQ 4. How do the research and educational or career outcomes for PIRE PIs, postdoctoral researchers, and graduate student participants compare to those of similar participant groups in similar NSF-funded projects that *do not require* an international collaboration??

The specific approaches that will used to address each of these questions are described below.

RQ 2. How does the quantity and quality of publications produced by PIRE projects compare to the quantity and quality of publications produced by similar NSF-funded projects?

Relational analyses will be used to investigate how research outcomes for PIRE projects differ from those of similar NSF-funded projects (the process for identifying and matching these comparison projects to PIRE projects is described above in Section B.1). Ordinary least square (OLS) regression models will be fit to data in which the outcomes are dependent variables and the independent variables include a dichotomous PIRE indicator (1=PIRE project, 0=comparison project), covariates that control for differences in project-level characteristics (e.g., award cohort, duration, and funding amount), and variables that indicate matched pairs or matched groups (i.e., if many-to-one matching is used to match two comparison awards to a single PIRE award). Results will be presented in tables of parameter estimates with accompanying table notes. The coefficient for the PIRE treatment indicator will represent the unique relationship of PIRE to differences in number of publications (and other bibliometric outcomes) between the PIRE and comparison groups, controlling for other factors included in the model. Descriptive tables will also be presented that show means and percentages for the outcomes overall, by field of discipline, and for other major subgroups of interest.

It is important to note that at these analyses are non-experimental and any differences in outcomes will be viewed as exploratory, interpreted with caution, and presented with discussions of their limitations.[18]

RQ 4. How do the research and educational or career outcomes for PIRE PIs, postdoctoral researchers, and graduate student participants compare to those of similar participant groups in similar NSF-funded projects?

To examine how the research and career outcomes of PIRE participants compare to participants

[18] The chief obstacle to implementing a quasi-experimental design for this set of analyses is the lack of an appropriate representation of the counterfactual condition: it is difficult to determine, for example, what measures *at the project level* exist for project characteristics before the formation of the project. That is, the PIRE award itself could affect which individuals come together to form the research group that ultimately produces publications – in the absence of the PIRE funding, those individuals did not form an already established group for which characteristics can be measured. For example, in contrast to analyses of individual participants' publication records, there is no *project-level* publication history.

in other NSF-funded projects, the evaluation will match three groups of PIRE participants (PIs, postdoctoral researchers, and graduate students) to corresponding participants from a comparison group of non-PIRE, NSF-funded projects using a statistical technique called *greedy matching* (described below). For research outcomes, the evaluation will also incorporate a comparative interrupted time series (CITS) design. The reasons for adopting this analytical approach are discussed in the Introduction to Part B (see Rationale for the Evaluation Design).

Greedy matching and comparative interrupted time series (CITS) are quasi-experimental designs (QEDs) to approximate the counterfactual condition (i.e., what would have occurred in the absence of the PIRE program). A CITS design is considered to be one of the strongest non-experimental methods available to approximate a true estimate of the effect of a program on key outcomes (Hotz, Imbens, & Klerman, 2006; Dehejia & Wahba, 1999; Shadish, Cook, & Campbell, 2002). It is important to note that even a strong quasi-experiment cannot account for unmeasured variables that may affect the measured outcomes of interest. These omitted variables could introduce bias into the estimated effects. However, when only non-experimental estimates of the true effects of a program are feasible, QEDs reduce bias that can obscure the true effect of a program, and the CITS approach is a strong alternative to an experimental design. This evaluation uses quasi-experimental approaches that reduce bias to estimate the effect of the PIRE program on participants' research and career outcomes.

As described above in Section B.1., a comparison group of non-PIRE, NSF-funded projects will be identified. From this comparison group of projects, participant-level comparison groups will be constructed for PIRE PIs, postdoctoral researchers and graduate students using greedy matching. To compare the publication quantity and quality for these matched participant groups, a CITS design will also be incorporated. The steps that will be taken to create each of these three matched comparison groups of participants and estimate program impacts are explained below.

Constructing Participant-Level Comparison Groups: Greedy Matching

Participants in PIRE and comparison projects within each PIRE-comparison project pair will be matched using a greedy matching distance technique. Matches will be based on a pre-project-participation measure of one of the key research outcomes for the study (e.g., the number of publications prior to project participation or the mean citation measure for publications prior to project participation). One frequently used multivariate matching technique, propensity score matching, would require a larger sample of PIRE participants to create reliable propensity scores; in addition, to create propensity scores the evaluation would require a larger number of variables describing participants than are available from existing data prior to data collection. Thus, we will use another technique, greedy matching to match participants. Unlike PSM techniques, greedy matching is not sensitive to sample size. Greedy matching proceeds by:

1. Randomly sorting the *N* PIRE participants and *M* comparison project participants;
2. Next, the first PIRE case in the list is matched to the closest comparison participant—namely, the one most similar to the PIRE participant;
3. Next, the second PIRE participant is matched to its closest comparison participant among those remaining; and

4. This process repeats until all PIRE participants have been matched.

Here, “similarity” between a PIRE and comparison participants is defined as the average “*distance”* per year between the PIRE and comparison participants’ score on a pre-test measure of research outcomes. For example, using the number of publications as the pre-test measure, the average distance per year between the ith PIRE participant and the jth comparison participant is defined as

$$D_{ij} = \frac{\sum_{n=1}^{N} X_n^1 - X_n^0}{N} ;$$

where X^1= { $X_1^1, X_2^1, X_{3,}^1 \ldots X_N^1$ } is the number of publications across N years for PIRE participants and

X^0= { $X_1^0, X_2^0, X_{3,}^0 \ldots X_N^0$ } is the number of publications across N years for comparison participants.

Estimation of Differences

Following the greedy matching of PIRE and comparison participant groups, we will estimate the effect of PIRE program participation on research productivity and quality using a CITS approach. The CITS analyses for research productivity and quality will include five years of bibliometric data prior to project onset and multiple years of post-onset data to estimate the impact of PIRE on the following outcomes. (Appendix J provides technical details on the interrupted time series models we will use.)

- Mean number (and standard deviation or standard error) of research reports, journal articles and other products published per participant;
- Mean number of times a participant’s publications have been cited;
- The mean field-normalized citation impact for a participant’s publications;[19]
- The mean ratio of observed to expected citations per publication per participant;[20] and
- The mean percentage of publications with co-author(s) at foreign institutions, per participant.

By including data on participants in comparison projects with the opportunities to publish research occurring at the same time as their matched PIRE participants, the CITS approach controls for the effects of any factors external to PIRE that might occur at the same time as participation in PIRE and that could also affect the outcomes of interest (e.g., global economic trends in research and development).

[19] Because publication and citation patterns vary across disciplines and publication types, normalization techniques account for these patterns to produce a standardized citation index.

[20] This bibliometric indicator likewise takes into account patterns of citations for publications of a particular type, within a particular journal, field, and year of publication to predict the “expected” number of citations for comparison to the actual number of citations received by a given publication.

By including data on the participants prior to the onset of their participation in the project (PIRE or comparison), the CITS design controls for persistent (time-invariant) characteristics of projects that might differentially affect outcomes (i.e., persistent, time-invariant differences between participants). For example, the fields of research of participants likely affect both pre- and post-participation outcomes, and it is reasonable to assume that this research field effect on these outcomes is stable over time. Thus, our impact analysis models will include data on research fields to control for differential effects of different research disciplines on PIRE outcomes. Using data for PIRE participants and matched comparison individuals both prior to and after project onset allows us to control for persistent, project-specific factors that could explain any observed differences, thus reducing the number of plausible alternative explanations for observed effects.[21]

To assess the differences in career outcomes for PIRE and comparison group participants, a CITS design is not feasible, since the changes in career outcomes over time are not readily measured on a scale that is consistent both pre- and post-participation (e.g., for a graduate student participating in PIRE, a career outcome such as attaining a tenure-track assistant professorship has no clear pre-participation analog). Instead, the evaluation will use regression models to compare educational and career outcomes of PIRE and comparison group PIs, postdoctoral researchers and graduate students. Ordinary least square (OLS) regression models will be fit to data in which selected career outcomes (e.g., engagement in international collaboration after the conclusion of project participation; employment in a job requiring a degree in engineering, science, or mathematics) are dependent variables and the independent variables include a dichotomous PIRE indicator (1=PIRE project participant, 0=comparison project participant), covariates that control for differences in participant characteristics (e.g., duration of project participation, demographics), and other covariates that could be correlated with career outcomes. Parameter estimates will be presented in tables with accompanying table notes. The coefficient for the PIRE treatment indicator will represent the unique relationship of participation in PIRE to differences in the career outcome of interest between the PIRE and comparison groups, controlling for other factors included in the model. Descriptive tables will also be presented that show means and percentages for the outcomes overall, by field of discipline, and for other major subgroups of interest. It is important to note that at these analyses are non-experimental and any differences in outcomes will be viewed as exploratory, interpreted with caution, and presented with discussions of their limitations.

Limitations

This section describes potential limitations of the proposed evaluation design and methods used to minimize the effect of these limitations.

Selection Bias

A major threat to the validity of evaluations using quasi-experimental methods comes from the existence of unobservable characteristics that affect both outcomes of interest and participation

[21] The CITS design used here assumes that the effect of these (time-invariant) characteristics on study outcomes for a given participant is the same both before and after the onset of participation in the project (whether the project is a PIRE or comparison project).

in (or “selection” into) the treatment or comparison program. The optimal way to address the threat of unobservable characteristics is to utilize as many relevant observable characteristics as possible in the matching process, so that the effect of these factors is reduced. It is especially important to use characteristics that existed before the treatment period began.

However, when matching PIRE projects to comparison projects, there is a limit to the number of such “pre-award” characteristics at the project level on which to match projects. Before a PIRE or comparison award was made (or before the proposal for the award was prepared) the group of participants in the award did not form an established group for which characteristics can be measured. For example, there is no project-level publication history prior to the NSF decision to award a particular proposed PIRE project, since the project did not exist, per se, pre-award. Although ratings of proposals during the NSF proposal review process were considered as a proxy indicator for the expected outcomes of the project, PIRE and comparison awards were rated by different reviewers based on different award criteria and thus these ratings were determined to be not comparable. Consequently, no program-level pre-program characteristics can be used in analyses.

Nevertheless, we will be able to address selection bias more successfully when comparing the outcomes of participants in PIRE and comparison projects. Because individuals do have measurable characteristics that existed before they participated in the project, we can use more rigorous matching methods (such as greedy matching) to mitigate the threat of selection bias for analyses of individual research and career outcomes. Where appropriate in reports, we will explicitly identify where causal inferences about the effects of PIRE must be made with caution.

Citation Indices and Impact Factors

One primary measure of research outcomes is citation indices. The assumption that undergirds these types of measures is that the more highly cited a research publication, the greater its influence on those citing it. In other words, citations are used to index how important a publication is considered in a given field. However, this assumption is not wholly accepted. Critics have argued that the reasons for citing a publication can be varied and may not always signify the quality of a publication (Hanney et al., 2005; Moed, 2005), and some highly influential publications may not be acknowledged as such for many years or decades until they are “re-discovered” as transformative contributions to a field. Moreover, in some fields, such as computer science, mathematics, and engineering sciences, research is not disseminated primarily through peer-reviewed journal publications. Nevertheless, there is evidence for a correlation between peer reviews of quality and citation indices (e.g., Rinia et al., 1998), and bibliometric analyses are one of the few quantifiable proxies for research quality that are available. We will make every effort to restrict comparisons to within similar fields of research, and we will interpret results with caution.

Benchmarking

There are four primary limitations of data that will be used for benchmarking. First, the administrations of the SDR and NSRCG are out of phase with our planned survey data collection. For example, the most recent administrations of the SDR for which it is feasible to obtain data within our timeline are the 2010 and 2012 SDR waves; in contrast, it is expected that surveys for the PIRE evaluation will be fielded (pending OMB approval) between October or November 2014 and February 2015.[22] As a result, if we use the year of doctoral degree attainment as a variable on which to compare our study population of PIRE postdoctoral fellows and graduate students who have completed their doctorate to SDR populations, the groups would be out of phase by two to four years, depending on the SDR cycle. For example, 2010 SDR respondents who earned their PhDs in 2004 would have had six years to achieve professional and career outcomes by the time of the data collection. In contrast, PIRE respondents who received their PhD in 2004 would have had 10 years to achieve outcomes, possibly giving them an advantage in terms of career outcomes. In order to address this problem, we plan to compare 2010 SDR respondents who earned PhDs in 2000 to PIRE respondents who completed their PhDs in 2004, to allow for the same duration of time to elapse between the receipt of the degree and the reporting of outcomes.

Similarly, our surveys of former undergraduate and master's level PIRE participants will be out of phase with the most recently administered NSRCG data. Moreover, the NSRCG was discontinued after the NSRCG 2010, which was administered between February and September 2012. As a result, benchmarking data for PIRE graduate student participants (i.e., those who received a master's degree, but not a PhD) will be limited to a sample of respondents who received a bachelor's or master's degree between July 2007 and June 2009.

The second limitation of the SDR and NSRCG data is that these surveys were not designed to measure many of the outcomes that are pertinent to the PIRE evaluation. Research productivity and international collaboration, for example, are particularly notable omissions. The use of the SDR and NSRCG for a nationally representative comparison will be limited to a subset of items such as characteristics of current employment and the match between education and job requirements.

Third, the SDR sample has the potential to include some PIRE respondents. Unless these individuals could be removed from the SDR sample, we might expect some level of contamination. Because personal identifying information for SDR respondents is not available (even under restricted use data licenses), we would not be able to remove PIRE participants from the SDR sample. Likewise, past experience has shown that the Division of Science Resources Statistics at NSF is unable to identify and thus remove these individuals from the SDR sample prior to transferring the data to the evaluation contractor. Asking PIRE survey respondents to indicate whether they have completed the SDR (or NSRCG) in the past could provide a rough estimate of the level of cross-contamination, but would likely be subject to recall bias; this method would likely underestimate the contamination effect. However, given the likely large size of the sample drawn from SDR data relative to the PIRE participant groups surveyed, we predict that cross-contamination will be minimal.[23]

[22] SDR 2014 data are scheduled for release in April 2014. We will attempt to use the most recently available SDR data.

[23] The SDR includes approximately 40,000 respondents per wave.

Finally, it is important to note that these benchmark comparisons are subject to selection bias. The samples represent populations that are likely different from PIRE participants on a number of characteristics associated with the outcomes and the likelihood of being a PIRE participant. Consequently, all benchmark comparisons will be treated with caution and their limitations will be clearly articulated in reports.

Recall Bias

PIRE and comparison program participants will be surveyed regarding their activities from a period that ranges from 2005 to 2012 (the specific years will vary according to cohort). The width of the range raises the concern that recall bias could be an issue in survey responses. Recall bias refers to the problems that respondents might have accurately recalling the events asked about in the surveys. Recall bias has been extensively studied in the medical literature. For example, Litwin and McGuigan (1999) documented that patients could not accurately recall their health status as little as three years after undergoing a surgical procedure (patients tended to report pre-surgery quality of life as better than it actually was). To address the possibility of recall biases, we will explore whether there are marked differences in the responses of individuals who participated in earlier PIRE cohorts when compared to later cohorts.

Limitations of Greedy Matching

Greedy 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. 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. For example motivation, an unobservable characteristic, is often an important factor that affects an individual’s participation in a program as well as his/her outcomes. Program participants may be more motivated than non-participants and thus have better outcomes. In this case, using propensity score matching may not fully remove the inherent difference between the treatment and the comparison groups.

The best way to deal with the threat of unobservable characteristics is using as many “relevant” observable characteristics as possible in the matching process, so that the effect of these factors is reduced. This study will employ all available matching variables to minimize selection bias. Assuming that PIRE award decisions were also based on the same information, this approach should account for some of the inherent differences between PIRE and comparison groups members and minimize selection bias.

B.4 Methods for Maximizing the Response Rate and Addressing Issues of Non-Response

Maximizing Response Rates

The evaluation will draw upon the expertise of the contractor, Abt Associates, in collecting and analyzing similar types of data for other large, federally funded, institutionally based programs.

Achieving strong response rates on a survey begins with a well-designed, user-friendly instrument and includes a clear and convincing rationale for the survey and the importance of respondents' participation. Once these are in place, the ability to achieve high response rates will depend on to the survey fielding team's ability to accomplish three tasks:

1. *Locate* as many respondents as possible (known as the "find rate");
2. *Reach* as many respondents as possible; and
3. *Convince* as many respondents as possible to complete the survey.

Below is a discussion of the proposed approaches to successfully accomplishing these steps, illustrated with examples from past program evaluations with similar populations.
Locating respondents depends on their role in the projects included in the evaluation as well as the timing of their participation. PIRE and comparison project PIs, postdoctoral fellows, and foreign senior investigators, and representatives from PIRE lead PIs' institutions will likely be easier to locate than graduate student and undergraduate student participants because the former groups typically remain in academia during and subsequent to project participation. Individuals in all respondent groups who have left academia likely will be more difficult to find than individuals who remain affiliated with universities. Foreign participants can be more difficult to find than domestic participants, especially if these individuals participated in the earliest PIRE cohort (2005) and have changed institutions since the period of these awards (2005-2012).

The evaluation will utilize several strategies to locate respondents. For all respondent groups, contact information available from project information files and annual and final reports in NSF program data will serve as a starting point for location efforts. Web-based search engines (e.g., Google) on each participant name in combination with the participant's institution at the time of participation will be used to verify and update email addresses, campus mailing addresses, and telephone numbers. Other sources, such as academic article databases (*e.g*., PubMed) and individuals' web pages (most common for those in academia), will be used to supplement these searches to identify the most current contact information. Email addresses may also appear in published journal articles and in presentations and other materials posted on the web. Only those individuals for whom the information available on the Internet appears current (e.g., a recent university web site) will be classified as "found."

For the remaining ambiguous or missing cases, fee-based electronic databases such as AccurInt (a database linked to LexisNexis) will be used to verify or update the mailing addresses and phone numbers for participants listed in PI-submitted annual reports.[24] If a mailing address is found, the survey team will mail an invitation to participate in the survey with information on how to access the survey web site. For returned invitations, the team will use the US Postal Service's mail forwarding service to obtain change of address information. Postal invitations will be mailed up to three times using the forwarding address provided. Furthermore, if an AccurInt search reveals a telephone number, the team will dial that phone number, confirm the identity of the call recipient, and invite that individual to participate in the survey. Abt Associates' past

[24] If necessary, for respondents for whom we cannot locate a valid email address, we will mail a letter describing the survey and providing a URL link for that individual to login and complete the survey.

experience conducting survey data collection with similar populations indicates that AccurInt searches will be needed in particular for graduate students in terminal master's programs who have completed their degree and for participants in the earliest cohort of PIRE (and corresponding comparison projects).

An additional strategy to boost find rates for graduate and undergraduate students is to ask PIs to provide any updated information about their former students' whereabouts. Likewise, PIs may also be able to provide updated information about foreign participants' current home institutions and contact information. This approach has proven reasonably effective in Abt's evaluation of an NIH research training program, in which the evaluation team needed to locate individuals who had participated in the program up to eight years prior to the data collection. In this study, 48 percent of trainee advisors provided valid contact information and an additional 26 percent had some knowledge about their trainees (Abt Associates, 2009). This approach will likely be most effective for the more recent PIRE and comparison projects (2007, 2010 and 2012 PIRE cohorts or comparison projects of similar duration that began within the past 5 to 7 years). The survey team for the evaluation of the PIRE program is considering a staged, sequential roll-out of surveys based on initial find rates for graduate and undergraduate project participants (e.g., fielding the PI surveys some number of weeks prior to fielding surveys for harder-to-locate respondent groups). Finally, as fielding proceeds, the survey team will track all returned-as-undeliverable emails and attempt further searches to identify correct contact information.

Even if a potential respondent's email address is correct, some emails may be blocked from delivery or diverted into spam blockers and will not reach their targets. To increase email deliverability, the survey team follows standard guidelines to ensure that invitations are CAN-SPAM compliant; such suggestions include using the respondent's name, minimizing the number of links, images and overall message size relative to the amount of text, using a subject line that communicates clearly the content of the email, and including a physical mailing address and an opt-out link to avoid having survey invitations or reminders be marked as a spam.[25]

The final requirement for obtaining strong response rates is to persuade respondents to spend the time it takes, however much this duration has been minimized, to complete a survey. The survey team will adopt several strategies to accomplish this. First, the team will request that NSF send an initial email to all survey respondents (in the census or sample, depending on the respondent group) to explain the goals of the study and ask for their participation. (See Appendix M for email text.) In the contractor's experience, such initial requests for cooperation from funders significantly improve study participation. Second, the team has taken steps to prepare surveys of minimal length that are well-organized and that contain clearly worded survey items and a small number of response options per item (e.g., using yes/no questions when possible to route respondents past unnecessary items). Online implementation will be designed for easy navigation; for example, items will be formatted to fit within a standard window to prevent respondents from having to scroll left-to-right or top-to-bottom within a page, and definitions of key terms will be presented with mouse-over hover boxes to prevent navigating away from the

[25] See http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf

survey item for clarification. Techniques to motivate respondents to proceed to the end of the survey, such as a scale illustrating the degree of survey completeness, will be implemented if feasible.[26] The online surveys will be pre-loaded with the information available from extant data (e.g., institutional affiliation) so that respondents will need only to verify its accuracy. Skip patterns and pull-down menus will be implemented where possible to minimize respondent effort. All surveys will include navigation instructions, a Frequently Asked Questions section, and a toll-free telephone number and email address for technical assistance or questions about the survey.

Finally, during the survey field period, follow-up with non-respondents will include three to five emails and one or two telephone calls to further increase response rates. The contractor has successfully used these types of strategies to achieve good response rates. Exhibit B.5 illustrates response rates for various evaluations of NSF programs conducted by Abt Associates; these can be used to predict expected response rates for this evaluation.

[26] Complex skip patterns, in which the number of remaining items depends on respondent answers, can impede the clarity of this type of progress bar, undermining the purpose of showing respondents their progress. This tool will be tested before launch for clarity.

Exhibit B.5: Response Rates in Previous Studies with Similar Populations

Program and Respondent Group from Previous Evaluation	Length of Time Between Program Participation and Data Collection	Response Rate
Respondents similar to Principal Investigators (Faculty)		
NSF CAREER Awardees (Faculty)	0-10 years	84%
NSF CAREER Fellows’ Department Chairs (Faculty in a CAREER awardee’s home department)	0-10 years	84%
NSF EAPSI Fellows’ U.S. Advisors (Faculty advisors to EAPSI graduate students)	0-11 years	71%
Postdoctoral Respondents		
NSF IRFP (Postdoctoral Fellows)	11-19 years 6-10 years 0-5 years	78% 77% 88%
NSF IRFP Unfunded Applicants (declined applicants for an IRFP postdoctoral fellowship)	11-19 years 6-10 years 0-5 years	48% 50% 65%
Graduate Student Respondents		
NSF EAPSI Fellows (Graduate students)	6-11 years 0-5 years	68% 78%
NSF GK-12 Fellows (Graduate students)	5-10 years	MS 45% PhD 57%
NSF GK-12 Fellows (Graduate students)	0-5 years	MS 83% PhD 92%
Undergraduate Student Respondents		
NSF Noyce Scholarship Recipients (undergraduate and post-baccalaureate STEM majors)	0-8 years	65%
Foreign participants in NSF programs		
NSF IRFP Foreign Host Scientists (hosted a postdoctoral fellow from the U.S.)	0-19 years	61%
NSF EAPSI Foreign Host Scientists (hosted one or more EAPSI graduate students from the U.S.)	0-11 years	61%

Notes:
IRFP data from Abt’s Evaluation of NSF’s International Research Fellowship Program (IRFP). See Martinez, Epstein et al. (2012).
EAPSI data from Ab’s Evaluation of the East Asia and Pacific Summer Institute (EAPSI) Program. See Martinez, Neishi et al. (2012).
CAREER data from Abt Associates’ Evaluation of NSF’s Faculty Early Career Development (CAREER) Program. See Carney et al. (2008).
GK-12 data from Abt Associates’ Evaluation of NSF’s Graduate STEM Fellows in K-12 Education Program. See Gamse et al. (2010).
Noyce data from Abt Associates’ Evaluation of NSF’s Robert Noyce Teacher Scholarship Program. See Bobronnikov et al. (forthcoming).

Addressing Issues of Non-Response

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. The evaluation 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. Because expected non-response rates for this evaluation are greater than 10 percent, unit non-response bias analyses will be conducted by 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.[27] For example, if the response rate from women was very low and women were less likely to belong to a PIRE participant group then any difference in the outcomes between the PIRE and comparison participant groups could result in a biased estimate of the impact of the treatment. 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), 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.

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 impact analyses, 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 impact analyses where missing data on covariate or predictor variables requires 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 impact 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). In this multiple imputation approach, instead of generating one set of values to replace the missing outcome, we generate multiple sets of 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 are available for all individuals (respondents and non-respondents). Then, to each predicted

[27] 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 impact 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).

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 impact 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 impact of the program on the specified outcome; each impact estimate has used one of the ten datasets in which missing data were replaced with the predicted value plus residual. The final impact estimate (that is, the estimate of the effect of the program on the outcome) is the mean of the ten individual impact 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.

B.5 Tests of Procedures or Methods

The evaluation has a convened an external Subject Matter Expert Working Group (SMEWG) representing expertise in STEM policy and program evaluation and survey and bibliometric analysis methodologies (these individuals are named in Section B.6). This group has reviewed the program logic model, study design, and data collection instruments (the logic model appears in Appendix A; survey instruments are in Appendices B-G). The six proposed surveys have been pilot tested with PIs, postdoctoral researchers, and graduate and undergraduate students, foreign senior investigators and institutional administrators. Feedback from these pilot testers has been incorporated to clarify items and to determine estimates of the time required to complete each survey (these estimates appear in Part A, Section A.12). Pending any requested revisions to the surveys from OMB, surveys will be programmed for web-based implementation; these online versions will be tested for functionality, skip patterns, response options, ease of responding, and formatting by the contractor prior to survey launch.

B.6 Individuals Consulted

Key personnel who have been involved in the study design are presented in the table below. These individuals include the Abt Associates team. Abt staff members have deep knowledge on statistical methods, experience in evaluation of large scale programs, expertise in scientific research, and content knowledge of STEM higher education programs.

Members of the SMEWG were also consulted in the design, and may also be consulted in the analysis of data. Finally, NSF program staff members familiar with the programs have been included in the design of the evaluation. The NSF point of contact for this study is **Suzanne Plympton,** OMB Desk Officer. The Contracting Officer’s Representative overseeing the implementation of the evaluation is **John Tsapogas,** Office of Integrative and International Affairs (OIIA).

Exhibit B.6 Individuals Consulted

Name	Role
Alina Martinez, Principal Associate, Abt Associates Inc.	Principal Investigator
Carter Epstein, Scientist, Abt Associates Inc.	Project Director
Amanda Parsad, Senior Scientist, Abt Associates Inc	Director of Analysis
Laurie Bozzi, Associate, Abt Associates Inc.	Task leader, data collection
Jonathan Adams, Chief Scientist, Digital Science, Macmillan Publishers Ltd	Subject Matter Expert Working Group
Rajika Bhandari, Institute of International Education	Subject Matter Expert Working Group
Susan Cozzens, School of Public Policy, Georgia Institute of Technology	Subject Matter Expert Working Group
Irwin Feller, Emeritus, Pennsylvania State University	Subject Matter Expert Working Group
Diana Hicks, School of Public Policy, Georgia Institute of Technology	Subject Matter Expert Working Group

References

Abt Associates. (2009). *Needs assessment of the NIGMS Research Supplements to Promote Diversity in Health-Related Research: Final report.* Cambridge, MA: Author.

Bobronnikov, E., Gamse, B., Price, C., Roy, R. & Velez, M. (forthcoming). *Implementation and impact findings from the evaluation of the Robert Noyce Teacher Scholarship Program.* Report prepared for the National Science Foundation under contract GS-10F-0086K**,** Order No. NSFDACS09D1625

Carney, J., Smith, W.C., Parsad, A., Johnston, K. and Millsap, M. (2008) *Evaluation of the Faculty Early Career Development (CAREER) Program.* Report prepared for the National Science Foundation under contract GS-10F-0086K, Order No. D050540.

Cominole, M., Siegel, P., Dudley, K., Roe, D., and Gilligan, T. *2004 National Postsecondary Student Aid Study (NPSAS:04) Full Scale Methodology Report* (NCES 2006–180). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved October 19, 2011 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006180

Crisp, G., Nora, A., and Taggart, A. (2009). Student Characteristics, Pre-College, College, and Environmental Factors as Predictors of Majoring in and Earning a STEM Degree: An Analysis of Students Attending a Hispanic Serving Institution. *American Educational Research Journal, 46*(4) 924-942.

Dehejia, R. and Wahba, S. 1999. Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association 94*(448), 1053–1062.

Durieux, V. and Gevenois, P. A. (2010). Bibliometric indicators: Quality measurements of scientific publication. *Radiology, 255*(2), 342-351. Retrieved July 2, 2014 from http://radiology.rsna.org/content/255/2/342.full.pdf.

Gamse, B., Smith, W.C., Parsad, A., Dreier, J., Neishi, K., Carney, J., Caswell, L., Breaux, E., McCall, T. and Spader, J. (2010). *Evaluation of the National Science Foundation's GK-12 Program: Final report, volume I: Technical report.* Prepared under contract GS10F-0086K, Order # NSFDACS06D1412.

Hanney, S., Frame, I., Grant, J., Buxton, M., Young, T. and Lewison, G., (2005). Using categorisations of citations when assessing the outcomes from health research. *Scientometrics, 65*(3), 357- 379.

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.

Hotz, V.J., Imbens, G. W., and Klerman, J.A. 2006. Evaluating the differential effects of alternative welfare-to-work training components: A re-analysis of the California GAIN Program. *Journal of Labor Economics, 24*:2, 521-566.

Lahiri, D.B. 1951. A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute 33(*2): 133-146.

Litwin, M.S. & McGuigan, K.A. (1999). Accuracy of recall in health-related quality-of-life assessment among men treated for prostate cancer. *Journal of Clinical Oncology 17*(9), 2882-2888.

Martinez, A., Epstein, C., Parsad, A. and Whittaker, K. (2012). *Evaluation of NSF's International Research Fellowship Program: Final report.* Prepared for the National Science Foundation under Contract GS-10F-0086K, Order No. NSFDAC S09T1516

Martinez, A., Neishi, K., Parsad, A., Whittaker, K. and Epstein, C (2012). *Evaluation of the East Asia and Pacific Summer Institutes Program: Final report.* Prepared for the National Science Foundation under Contract GS-10F-0086K, Order No. NSF DAC S09T1516.

Moed, H. (2005). *Citation analysis in research evaluation.* Dordrecht: Springer.

National Academy of Sciences (2014). Partnerships for Enhanced Engagement in Research (PEER) Science. See: http://sites.nationalacademies.org/pga/dsc/peerscience/index.htm

National Science Foundation (2011). National Science Foundation: FY 2012 Budget Request to Congress. Retrieved May 1, 2011 from http://www.nsf.gov/about/budget/fy2012/pdf/fy2012_rollup.pdf

National Science Foundation (2010). Science, Engineering, and Education for Sustainability (SEES) home page. Retrieved February 27, 2014 from http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504707

Office of Management and Budget. (2006). Office of Management and Budget standards and guidelines for statistical surveys. Retrieved July 2, 2014 from http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf

Pendlebury, D. A. 2003. *White paper: Using bibliometrics in evaluating research.* Philadelphia: Thompson Reuters.

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* (NCEE 20090049). Retrieved July 2, 2014 from http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20090049).

Rinia, E.J., van Leeuwen, T.N., van Vuren, H.G. and van Raan, A.F.J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands. *Research Policy, 27,* 95-107.

Shadish, W.R., Cook, T.D., & Campbell, D.T. 2002. *Experimental and quasi-experimental*

Wine, J., Janson, N., and Wheeless, S. (2011). *2004/09 Beginning Postsecondary Students Longitudinal Study (BPS:04/09) Full-scale Methodology Report* (NCES 2012-246). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved June 20, 2012 from http://nces.ed.gov/pubs2012/2012246_1.pdf

Appendices

A. PIRE Program Logic Model
B. Principal Investigator Survey
C. Postdoctoral Survey
D. Graduate Student Survey
E. Undergraduate Student Survey
F. Foreign Senior Investigator Survey
G. Institutional Administrator Survey
H. Survey Supplementary Materials:
    1. Research Fields for surveys
    2. List of Countries for surveys
    3. Survey Items: Mapping to Research Questions and Sources
I. Approach to Matching
J. Estimation of Impacts
K. Power Calculations
L. Non-response bias analysis
M. Recruitment and Reminder Materials
N. 60-day Federal Register Notice