NORC at the UNIVERSITY of CHICAGO

DOWNTOWN CHICAGO
55 East Monroe Street
30th Floor
Chicago IL 60603
office (312) 759-4000
fax (312) 759-4004
www.norc.org

November 23, 2011

This memo provides information on NSF's and NORC's responses to the OMB feedback received on November 7, 2011. The OMB feedback and NSF/NORC response are detailed below and organized around topics in hopes of facilitating easy identification in the full, revised OMB draft. The original set of comments is also included at the end of this memo.

1. DESIGN AND METHODOLOGY

a. OMB

We would encourage that for future cohorts efforts be taken now to ensure that the applicant and selection process data are scrutinized to be sure that any limitations (e.g., the use of subjective selection criteria) be considered and addressed as feasible to provide even stronger control groups in the future (e.g., thinking about how to "objectify" some of the more subjective selection elements). To that end, documenting lessons learned during the current evaluation may be useful.

NSF/NORC RESPONSE

NSF will consider the result of the evaluation and seek ways to implement lessons learned into the GRFP review process, in particular the applicant and selection process. NORC will keep track of lessons learned during the evaluation and will provide a set of suggestions to NSF for future efforts.

b. OMB

There is no differentiation in the discussion of response rates and locating strategies for older and more recent cohorts. Past NSF evaluations have suffered from an inability to locate older cohorts, potentially introducing significant bias into the results. Therefore, we would like to see NSF go through the exercise of estimating response rates by treatment and control group by cohort. We'd also like to different aspects of locating strategies emphasized for the older and newer and treatment and control groups.

NSF/NORC RESPONSE

The following explanation has been added to Section B.3 of the OMB package, elaborating on locating strategies and how these may be tailored to the quality of locating data within each cohort:

~~~~~~

Locating

Accurate address and telephone contact information are essential for notifying sample members of their selection in the study and further prompting for survey completion. Because we will be using the GRFP applicant records as the data source for sample member contact information captured at the time of applying to the program, the locating strategy has been designed to
~~~~~~

handle varying degrees of outdated information. Past NORC studies have found that 80% of located cases ultimately go on to complete the survey when following such a prompting strategy. Therefore, to target a 65% response, NORC will need to locate 75-85% of sample members within each cohort. To accomplish this, NORC will use a multi-stage strategy for locating sample members that will be responsive to varying amounts of locating information within any given cohort.

As noted, NSF will provide contact information from GRFP applicant records, including names and birthdates for all cases. For many cases, available information will also include social security numbers (SSN), address information, phone numbers, email addresses, and educational institutions (i.e., intended graduate school, current or previous college or university). Cases with incomplete contact information and cases with outdated contact information will be submitted for locating.

Our primary locating tools will include Accurint and LinkedIn searches. Accurint is a locating service that maintains a database of national information. When there is a match, the Accurint search yields address, phone, and/or email information. Previous NORC studies with populations similar to GRFP applicants have been able to locate 60% of their sample using Accurint searches. Because Accurint searches rely on SSN and birthdate information, LinkedIn will additionally be used for cases that lack these critical information fields. LinkedIn searches have been found to be highly successful in locating profiles of professionals when using academic institution information. NORC estimates that an additional 15-20% of the sample can be located through LinkedIn searches. LinkedIn is a professional networking web site where individuals create profiles listing their academic and professional credentials. Often these profiles list current employers, and/or contact emails and phone numbers. NORC has developed methods of searching LinkedIn profile pages using educational institution information listed on profile pages. These searches will be used to identify cases where sample members profiles list a current employer, current educational institution, or where there is contact information listed. Locators will then enter this information into our cases management system. This information will also be used to guide academic directory and employer directory searches.

These locating strategies will be employed during two phases of the study: prior to data collection (i.e., Pre-field), and during data collection.

Pre-field Locating

The locating strategy is based on the assumption that current address information may be incomplete for a portion of the sample. To account for this, pre-field locating will be conducted using Accurint. In addition, with NSF's support, NORC may contact coordinating officials at GRFP-sponsoring institutions to request updated contact information for more recent cohorts who may still be currently-enrolled graduate students.

Locating during Data Collection

Once the survey is in the field, we will mail letters to addresses obtained through Accurint searches or from applicant data and send out an email prompt. Cases having mail returned as undeliverable or invalid email addresses or phone numbers will be designated for more intensive locating treatments. These cases will be forwarded to NORC's locating department, where staff will conduct additional Accurint individual searches.

In addition, NORC will use LinkedIn searches using educational institution data contained in the applicant files (including BA institution, current institution, and intended graduate institution at the time they submitted the GRFP application) to locate sample members on LinkedIn using automated search techniques. Locators will manually review LinkedIn profiles for new contact information, including: email, phone, current employer, and current location. For older cohorts, LinkedIn searches will be necessary to determine updated location information. For example,

where LinkedIn searches produce an affiliation with an educational institution, the information will be used to conduct academic directory searches.

NORC expects that more recent cohorts will have more up-to-date information and will require less intensive locating efforts in comparison to older cohorts. For older cohorts, employing LinkedIn searches will likely be necessary to locate 75% to 85% of sample members. Location rates within cohorts, as well within awardee status group will be monitored throughout the data collection period, and the locating strategies applied as necessary.

~~~~~~

c. OMB

In addition, NSF needs to design in a nonresponse bias analysis, especially for the older cohorts.

NSF/NORC RESPONSE

Additional text has been added to Section B.4. "GRFP Follow-Up Survey: Tests of Procedures" containing a description of the how NORC will address and correct for nonresponse bias. The following has been added to the OMB package, Section B.4:

~~~~~~

Nonresponse Bias Analysis

High nonresponse overall or differential response rates between the treatment and control groups, and/or between older and newer cohorts can jeopardize the integrity of a study. The contractor plans to conduct a series of comparisons to assess the extent to which nonresponse has resulted in the respondent sample being different from the original baseline sample.

NORC will employ different approaches to examining non-response bias and accounting for it in the final analysis. Two potential options for our non-response bias tests are the Confidence Interval Bias Test and Cochran's Bias Test.[1] For the Confidence Interval Bias Test, variables such as demographic characteristics will be compared for the original baseline sample and the respondent sample. A confidence interval (CI) around the mean value among the responders will be calculated and compared to the mean value of the baseline sample—if the mean of the baseline sample falls within the CI, this suggests that the responders are sufficiently similar to the baseline sample and that it is not necessary to correct for nonresponse bias.

A more rigorous test to detect bias is the Cochran Bias Test, which is calculated by taking the difference of the mean of the responders and the mean of the baseline sample and dividing it by the standard error of the responders. A resulting bias of greater than 0.10 is considered problematic. It is important to note, however, that the Cochran test is extremely sensitive and leads to the conclusion of bias on most factors being tested. Thus, it is important to consider more than one type of bias test to determine if bias exists.

If non-response bias appears to be an issue, NORC plans to re-weight the sample data according to each respondent's likelihood or propensity of being a respondent. A logistic regression using baseline characteristics is used to predict the probability of being a respondent. Respondents in the sample with characteristics that most often are associated with nonresponse would effectively receive a higher weight to make up for their low incidence in the sample. These steps are essential to ensuring that our final estimates are not biased by the under-representation of any important subgroups, particularly the older cohorts. Where such methods are used, they will, of course, be carefully noted in the final report.

[1] Cochran, William G. (1977). *Sampling techniques* (Third ed.). NY: Wiley.

~~~~~~

d. OMB

In terms of study power, we appreciate the MDE estimates but would also like to know what effect size you would reasonably expect to see overall and for key analytical subgroups based on the literature. It also would be helpful to know what effect sizes NSF would define as policy or programmatically significant.

NSF/NORC RESPONSE

Additional text and tables have been inserted to Section B.1. "GRFP Follow-Up Survey: Sampling Methodology." The revisions better situate the power analysis in context by examining the literature for effect sizes, and the sample required to achieve a range of anticipated effect sizes. These revisions provide more complete information on the planned sample size, in total and specific sub-group populations, in relation to detecting estimated effects.

The following text in Section B.1 has been revised to address these concerns. Note that Tables B.1 – B.4 accompany the text in the OMB package but are not shown below:

~~~~~~

The sampling data file will contain unit-record identifiers, application information and QG rankings for all eligible GRFP applicants who received the fellowship or honorable mention award from four cohorts based on program application year: 1994-1998 (Cohort 1), 1999-2004 (Cohort 2), 2005-2008 (Cohort 3), and 2009-2011 (Cohort 4).

The main sampling frame is the list of GRFP Fellows and Honorable Mentions for 1994 – 2011. As shown in Table B.1.1, we propose to select a random sample of 13,188 cases from Cohorts 1 through 4 (1,099 QG1 Fellows, 1,099 QG2 Fellows, and 1,099 QG2 Honorable Mentions per cohort). Assuming a 65% overall response rate for GRFP Fellows and Honorable Mentions, we will obtain 8,568 completed questionnaires (714 QG1 Fellows, 714 QG2 Fellows, and 714 QG2 Honorable Mentions per cohort). Table B.1.2 additionally shows expected completes for specific sub-group populations without oversampling of minorities and disabled, pooling the four cohorts. The fewest expected completes are from disabled individuals, with expected 574 disabled and 7,994 others across the entire sample.

The sampling plan was designed to select a sample large enough to make statistically valid estimates of program outcomes in answering RQ1 (*What is the impact of the GRFP fellowship on the graduate school experience?*), RQ2 (*What is the impact of the GRFP fellowship on career outcomes?*), and RQ4 (*Is the program design effective in meeting program goals?*). A variety of analytic techniques will be used to address the research questions. While the size of the analytic sample, minimum detectable effects, and statistical power vary with the specifications of a given comparison, it is important to broadly assess statistical power and minimum detectable effects for a given sample to determine if each is sufficient for answering the research questions.

In some cases, comparisons will be based on the full sample pooled across all four cohorts, such as when examining the overall impact of the GRFP fellowship on the graduate school experience (RQ1). Here comparisons will be made among all current and former graduate students, spanning all four cohorts. In other cases, comparisons will be between sub-samples defined according to specific cohorts or population characteristics. For example, when examining the impacts of the GRFP fellowship on career outcomes (RQ2), comparisons will be made within a sub-sample comprised of former graduate students (Cohort 1 and Cohort 2 applicants). When examining GRFP program goals (RQ4), comparisons will be made within a given cohort, or between sub-

populations defined according to population characteristics (e.g., minority vs. other, female vs. male, disabled vs. other).[2]

Table B.1.3 provides information on the effect sizes that can be detected, based on the expected number of respondents for pooled samples and sub-samples, following conventional standards of an 80 percent level of statistical power and a 95 percent confidence level (alpha=0.05) for different comparisons. For comparisons based on the pooled sample of 2,856 completed questionnaires in each comparison group (Cohorts 1 – 4 QG1 Fellows, QG2 Fellows, and QG2 Honorable Mentions) and an expected estimate of 50 percent for a particular outcome across the full sample of cases, we would be able to detect a 3.7 percentage point difference between two groups. If the expected estimate for a particular variable of interest is 90 percent, we could detect a 2.1 percentage point difference.

The fewest expected completes from a key comparison group is among disabled students (n=574) compared with nondisabled respondents (n=7,812). Based on these estimates, we could detect a 6.0 percentage point difference if the expected outcome estimate is 50 percent and a 3.2 percentage point difference if the expected outcome estimate is 90 percent. Table B.1.3 provides equivalent information on other comparison groups.

To provide context for the minimum detectable effects in our study, we examined past studies of GRFP applicants to see what differences have been reported in the literature. This enables us to determine if our sample size will be sufficient to support the planned analyses. While the following review focuses on Ph.D. completion rates as a useful point of reference to past studies, it is important to note that this evaluation will examine a sizable number of different outcomes related to graduate education, careers, and professional productivity.

The most recent comprehensive evaluation of the GRFP (Goldsmith, et al., 2002) provided evidence of mean differences between QG1 Fellows, QG2 Fellows, and QG2 Honorable Mention recipients. Results indicated that, for example, Ph.D. completion rates by 1999 among 1979-1988 GRFP applicants were 75.3 percent for QG1 Fellows, 69.4 percent among for QG2 Fellows, and 66.0 percent for QG2 Honorable Mentions, suggesting a 3.4 percentage point program effect among comparable QG2 applicants.[3] This difference would not be significant at the 95% confidence level, given the power of the proposed sample to detect differences between quality groups within a single cohort. However, if differences of similar size were found in two or more of the 5-year cohorts, the pooled cohort comparisons would have sufficient power for the difference to be significant at the 95% confidence level. The GRFP evaluation will examine Ph.D. completion rates among all but the most recent cohort of applicants.

The Goldsmith, et al. (2002) evaluation also reported gender differences in Ph.D. completion rates by 1999 among the 1979-1988 GRFP applicants. The results showed 70.5 percent of male QG1 Fellows completed their Ph.D. by 1999, while 70.2 percent of QG2 Fellows and 67.5 percent of QG2 Honorable Mentions did so, indicating a 2.7 percentage point program effect. A difference of this magnitude among QG2 males would not be significant at the 95% confidence level, if all cohorts were pooled. The differences in Ph.D. completion rates were larger among females, with a 9.8 percentage point difference between QG2 Fellows and Honorable Mentions (68.3 vs. 58.5 percent); female completion rates were 73.3 percent for QG1 Fellows, 68.3 percent for QG2 Fellows, and 58.5 percent for QG Honorable Mentions.[4] In contrast to males, differences of this size would be detected with 95% confidence even within cohorts.

---

[2] See Table C.1 for additional details on data sources per specific analysis.

[3] See Goldsmith, et al. (2002), Table G14, p.141. http://www.nsf.gov/pubs/2002/nsf02080/nsf02080.pdf.

[4] See Goldsmith, et al. (2002), Table G9, p.136. http://www.nsf.gov/pubs/2002/nsf02080/nsf02080.pdf

Baker (1998) also examined gender and race differences in Ph.D. completion by 1988 among 1972 – 1981 GRFP applicants. Baker's findings indicate that rates of Ph.D. completion favored male GRFP applicants over their female counterparts by 4.5 percentage points among QG1 Fellows (77.4 vs. 72.9 percent), 9.1 percentage points among QG2 Fellows (73.6 vs. 64.5 percent), and 9.8 percentage points among QG2 Honorable Mentions (67.0 vs. 57.2 percent). Among QG2 applicants, male Fellows differed from male Honorable Mentions by 6.6 percentage points (73.6 vs. 67.0 percent), while female Fellows and Honorable Mentions differed by 7.3 percentage points (64.5 vs. 57.2 percent), again indicating a larger program effect among female applicants. Differences of this magnitude would be detected with the pooled sample of cohort 1 and cohort 2 in the current design.

Because the evaluation will include several different outcome measures and the analyses will compare a variety of different groups, Table B.1.4 presents the sample sizes required to detect a range of anticipated effect sizes based on mean differences between groups. Cohen's *d* metric for effect sizes (calculated as the mean difference between groups divided by the overall sample standard deviation) is frequently used to estimate the sample sizes needed to detect such an effect. Note that a smaller value necessitates a larger sample size. Following Cohen's conventions, a mean difference between two groups of 0.20 is considered small and would be detectable for samples of 393 or more cases within each comparison group.

In combination with the previous tables, it is evident that the proposed sample will be sufficient to detect effect sizes (mean differences between two groups) as small as 0.10 for pooled sample comparisons between QG1 Fellows, QG2 Fellows, and QG2 Honorable Mentions, and for comparisons between males and females. For pooled sample comparisons between disabled versus others, minorities versus other cases, or other sub-sample comparisons, the proposed sample will be sufficient to detect effect sizes between 0.10 and 0.20.

~~~~~~

e. OMB

We suggest replacing the word "impact" in SS B.4.II, where the institutional data collection is discussed, since we do not think that this study measures impacts on institutions.

NSF/NORC RESPONSE

Appropriate revisions have been made throughout the OMB package.
~~~~~~

2. SURVEY INSTRUMENT

a. OMB – survey length

It is rather long, and therefore has a high risk of dropouts and item non-response.

NSF/NORC RESPONSE

An initial pretest focused on the administration time of the instrument, resulting in a range of 30-40 minutes. The survey revisions discussed below under section 2b document questions that have been eliminated or consolidated, to address OMB concerns. Revisions to the survey following cognitive testing will also get administration time down to 30 minutes or less which is a typical threshold for participants.

Cognitive testing also will be used as a tool to explore the respondent's understanding of the survey questions and the cognitive processing to formulate an answer. Feedback collected during cognitive testing will include suggestions for improving the survey and reduce respondent burden. Additional explanation of the approach to be taken during cognitive testing is provided below in section 2d.

b. OMB – control groups

[The survey] does not ask the same questions of both the test and the control groups, calling into question its analytic power.

Survey instrument should become completely parallel between the test and control groups. This could be done by removing questions that explicitly ask about only the experiences of the fellows or by re-writing these questions to allow the control group to answer them as well. For instance and returning to question A5.3, you could consider asking "[Agree or Disagree] I had a variety of research projects from which to choose." Since you are considering the fellow and honorable mention groups to be matched samples, you could then analyze these data under the assumption that any difference was due to the fellowship program.

For those questions that you do not want to make parallel between the test and control groups, we suggest you add them to a student interview protocol for the implementation/qualitative portion of the research project. We believe that qualitative data derived from a purposeful sample may be the best way to get the information needed for RQ 4 that you had planned on deriving from the survey. A purposeful sample of both fellows and non-applicants (who are in the same academic department as a fellow) will allow a focus on how fellows contribute passively to their academic departments and programs.

NSF/NORC RESPONSE

The Follow-up Survey has been revised and NSF/NORC will look to the analysis of the data to identify program effects (differences due to the fellowship program). The only remaining survey questions to be asked *only* of Fellows are required to identify respondents who may have been awarded the Fellowship but chose to decline the award. Because the sampling file is based solely on program application data, this question is necessary to define the appropriate survey skip patterns.

In most cases, those questions from Section I, Part A that were to be asked only of Fellowship recipients, have been moved Section II of the survey which collects information on graduate student experiences and will be administered to the full set of Fellows and Honorable Mentions. However, in some cases, it became apparent that the information being asked by the Section I question was already captured elsewhere; such items were removed from the survey.

For tracking purposes, Table 1 at the end of this memo lists those survey questions that have been eliminated. Table 2 lists the questions that have been moved, along with the original and new

location. Table 3 identifies those questions that have been revised, along with the location of the item. Finally, Table 4 presents the items that have been moved *and* revised. A rationale for the revisions is provided in all the tables.

c. OMB - reliability

The survey asks a number of speculative questions that do not lend themselves to any sort of reliability. On this last point, we are referring to questions such as A5.3, which ask respondents to indicate whether receiving the GRFP award *improved* their opportunities to choose their research project. Since the rules of the GRFP prohibit previous graduate experience before the fellow applies for the grant, questions such as this ask the respondents to compare their lives to an un-lived hypothetical. A question such as this will not produce externally reliable data.

NSF/NORC RESPONSE

NORC has revised the survey accordingly (see response given to the previous item).

As a point of clarification, please note that the GRFP *does* allow applicants to have up to, but not more than, twelve months of full-time graduate study.

d. OMB – cognitive testing

We believe that the remainder of the survey (after applying the above alterations) could benefit from cognitive pre-testing, and suggest that you write such a procedure into your PRA request. Cognitive lab testing will allow you to find areas of redundancy and increase component validity across the questionnaire. We specifically suggest that you focus cognitive testing on Part II (Graduate School Experiences) Sections C and D, as well as Part III questions 28 and 29. While we understand that you took many of these categories from previous surveys, a greater effort should be made to tie the possible response categories to your research questions while reducing the overall burden placed on the respondents.

NSF/NORC RESPONSE

NORC has provided additional explanation in the OMB package, Section B.4, regarding the approach to be taken during cognitive testing, as follows:

~~~~~~

Pilot and Cognitive Testing

The survey was time tested with five individuals. Complete pilot testing with up to nine respondents will occur in December and will gather respondent comments on directions, clarity of items and overall logic of the programmed Web survey. Results from this pilot test will be used to refine the survey.

Cognitive testing also will be used as a tool to explore the respondent's understanding of the survey questions and the cognitive processing to formulate an answer. The scripted and unscripted cognitive probing during the interview will be directed towards understanding these issues. NORC will conduct five cognitive interviews. After each interview, respondents will be asked to provide feedback on the interview including respondent's overall interview experience, suggestions for improving the survey, and an open question and answer period for the respondent and interviewer.

The factors that will be examined during the cognitive interviews include: respondent understanding of the task/questions, respondent burden, interview timings, incorporating feedback from interviewers/respondents on problems with the instruments. Some specific questions that will guide the cognitive testing include:

- Do respondents have any difficulty comprehending the survey questions?
~~~~~~

- Are there any survey questions that can be improved or clarified?
- Are there any additional survey questions that should be included?
- How burdensome is the survey and has burden been reduced as much as possible?
- Can respondents provide accurate responses to survey questions that ask about events that may be more than a few years in the past?
- Has all relevant feedback from respondents and cognitive interviewers been incorporated?
- Is the timing of the instruments within the appropriate parameters?

The cognitive interviews provide valuable information on issues with the GRFP survey. Following the set of five interviews, data will be examined and materials will be revised in order to address the issues that emerge from testing.

~~~~~~

3. INTERVIEW PROTOCOLS

OMB

We believe further pretesting of the interview protocols is necessary, particularly for flow. Most of the protocols jump from one major topic (say program administration) to another (i.e. the experiences of the fellows) and back; semi structured interviews work best if they feel like a natural discussion. The faculty protocol is the only one that seems to flow well, and should be used as a model.

NSF/NORC RESPONSE

The following explanation is included in the OMB package, Section II, B.4, as follows:

~~~~~~

Interview protocols will be tested via cognitive interviews with faculty and administrators at graduate institutions similar to those selected for the site visit. This iterative cognitive interviewing process will allow NORC's qualitative research experts to quickly identify which questions yield answers relevant to the identified research questions, and which need to be revised or replaced to improve clarity and flow. NORC will pilot test the three interview protocols with at least two participants per protocol (i.e., at least six total participants) prior to the site visits. NORC will consult with NSF before selecting appropriate local institutions for pilot testing to make sure that NORC does not select an institution that would more appropriately be included in the full study.

~~~~~~
~~~~~~

SURVEY REVISIONS

Table 1. Questions that have been deleted

Original placement	Original Question(s)	Rationale
Section A #4, item 2&3	"Having the GRFP award on my CV helped/will help me in my job search"; "Having the GRFP award on my CV helped/will help in getting additional financial support"	Speculative question
Section A #5, item 1 &2	"Without receiving the GRFP award I would not have attended graduate school"; "Without receiving the GRFP award I would not have pursued graduate work in a STEM field"	Speculative question
Section A #5, item 3	"Receiving the GRFP award improved my opportunities to choose my research projects"	Similar question asked in Sec. C, 3b, item 2.
Section A #5, item 4	"Receiving the GRFP award improved my opportunities to work with faculty on their research projects"	Similar question asked in Sec. C, 3b, item 3; and Sec. C, 3c, all items.
Section A #6	"Did receiving the GRFP award influence your choice of which graduate institution to attend?	Cannot be applied to test and control groups
Section A #6a, items 1&2	"The award enabled me to enroll in a more selective institution"; "The award enabled me to enroll in an institution I would have not been able to afford"	We can assess selectivity and cost differences between Fellows and honorable mentions based on institutional data.
Section A #6a, item 3	" The award enabled me to enroll in an institution with better opportunities to research"	This is sufficiently captured elsewhere.
Section A #7a	"To what extent did the GRFP award enable you to transfer or change institutions?"	Cannot be applied to test and control groups
Section A #8, items 2	"The GRFP award supported my living expenses	Cannot be applied to test and control groups
Section A #8, item 3	"The GRFP award would be better if it lasted more than 3 years"	Cannot be applied to test and control groups
Section A #8, item 4	"The GRFP award would be better if it did not have a service requirement"	Cannot be applied to test and control groups

Table 1. *Continued*

Original placement	Original Question(s)	Rationale
Section A #8, item 5	"The GRFP award would be better if it allowed me to enroll at a foreign institution"	Cannot be applied to test and control groups
Section A #8, item 6	"The GFFP award would be better if it allowed me to concurrently accept other federal fellowship money"	Cannot be applied to test and control groups
Section A #9	"To which three of the potential five years during your graduate study did you (do you intend to) apply your GRFP funding?	Cannot be applied to test and control groups
Section A #10	"Please select from the dropdown list and identify which other awards you received during your 5 years of eligibility"	Cannot be applied to test and control groups
Section A #11	"For which years did you receive another award?"	Cannot be applied to test and control groups

Table 2. Questions that have been moved

Original Placement	Original Question	New placement	Rationale
Section A #7	"Did you change or transfer institutions at any time during your graduate study?"	Sec. B, 1e	Section A will now only pertain to fellowship recipients
Section A #10	"Did you receive another fellowship or sponsored program award at any time during this five year period?"	Section A follows after question 3	
Section A #12	"In your field, are there other fellowships or other sources of student support that are more desirable than the NSF Graduate Research Fellowship?"	Section A follows after question 3	

Table 3. Questions that have been revised

Placement	Original Question	New wording
Section C #3a item 1	My department offered **sufficient** enrichment activities (seminars, colloquia, social events, etc.) in addition to regular classes	My department offered a **variety of** enrichment activities (seminars, colloquia, social events, etc.) in addition to regular classes

Table 4. Questions that have been moved *and* revised

Original Placement	Original Question	New placement & question	Rationale
Section A #4, item 1	"Peers considered me a good student"	Section C #3a item #12 "My peers considered me a good student"	Revised question applies to test and control groups.
Section A #5 item 5	"The GRFP award enabled me to change departments during my graduate studies"	Sec. C, 3b, item 14 "The program made it easy to change departments"	Revised question applies to test and control groups
Section A #5 item 6	"The GRFP award enhanced my opportunity to work with a variety of different faculty on their research"	Sec. C, 3b, item 15. "The program made it easy to change advisors"	Revised question applies to test and control groups
Section A #6a items 4-11	" Select the ways in which receiving the GRFP award influenced your choice of which graduate institution to attend" Ex item: " The award enabled me to enroll in an institution in order to work with a specific faculty member"	Sec. C, 7. "Reflecting on your graduate school enrollment decision, to what extent do you agree or disagree with the following statements? I decided to enroll in my particular graduate institution… Ex item: " To work with a specific faculty member"	Revised questions apply to test and control groups
Section A #8 item 1	"I was an asset to faculty projects since I had my own GRFP funding."	Sec. C, 3c, item 12. "Faculty considered me an asset to their projects"	Revised question applies to test and control groups

OMB FEEDBACK RECEIVED ON NOVEMBER 7, 2011.

From: Mar, Sharon [mailto:Sharon_Mar@omb.eop.gov]
Sent: Thursday, November 03, 2011 1:55 PM
To: Plimpton, Suzanne H.
Cc: Earle, Janice M.
Subject: OMB Comments on GRFP Evaluation

Janice and Suzanne,

First I want to note that we were very happy with the overall improvements on design and methodology for the GRFP evaluation since OMB's last review of this collection. We particularly appreciated the efforts to identify plausible control groups. Since in many respects, we want to view GRFP as a model evaluation, we are providing you with the below comments for consideration that we believe will strengthen the GRFP evaluation. Once NSF has had the opportunity to go over our comments, we are more than happy to get on a conference call (if needed) and go over any questions that NSF may still have.

Design and Methodology:

We would encourage that for future cohorts efforts be taken now to ensure that the applicant and selection process data are scrutinized to be sure that any limitations (e.g., the use of subjective selection criteria) be considered and addressed as feasible to provide even stronger control groups in the future (e.g., thinking about how to "objectify" some of the more subjective selection elements). To that end, documenting lessons learned during the current evaluation may be useful.

We have several comments on the proposed design. There is no differentiation in the discussion of response rates and locating strategies for older and more recent cohorts. Past NSF evaluations have suffered from an inability to locate older cohorts, potentially introducing significant bias into the results. Therefore, we would like to see NSF go through the exercise of estimating response rates by treatment and control group by cohort. We'd also like to different aspects of locating strategies emphasized for the older and newer and treatment and control groups.

In addition, NSF needs to design in a nonresponse bias analysis, especially for the older cohorts.

In terms of study power, we appreciate the MDE estimates but would also like to know what effect size you would reasonably expect to see overall and for key analytical subgroups based on the literature. It also would be helpful to know what effect sizes NSF would define as policy or programmatically significant.

We suggest replacing the word "impact" in SS B.4.II, where the institutional data collection is discussed, since we do not think that this study measures impacts on institutions.

Survey Instrument:

We see three major issues with the survey instrument:

1. It is rather long, and therefore has a high risk of dropouts and item non-response.
2. It does not ask the same questions of both the test and the control groups, calling into question its analytic power.
3. The survey asks a number of speculative questions that do not lend themselves to any sort of reliability. On this last point, we are referring to questions such as A5.3, which ask respondents to indicate whether receiving the GRFP award *improved* their opportunities to choose their research project. Since the rules of the GRFP prohibit previous graduate

experience before the fellow applies for the grant, questions such as this ask the respondents to compare their lives to an un-lived hypothetical. A question such as this will not produce externally reliable data.

To address these issues we recommend NSF consider the below three options, which, taken together should greatly improve not only the survey instrument but overall design of the research project.

1. Survey instrument should become completely parallel between the test and control groups. This could be done by removing questions that explicitly ask about only the experiences of the fellows or by re-writing these questions to allow the control group to answer them as well. For instance and returning to question A5.3, you could consider asking "[Agree or Disagree] I had a variety of research projects from which to choose." Since you are considering the fellow and honorable mention groups to be matched samples, you could then analyze these data under the assumption that any difference was due to the fellowship program.
2. For those questions that you do not want to make parallel between the test and control groups, we suggest you add them to a student interview protocol for the implementation/qualitative portion of the research project. We believe that qualitative data derived from a purposeful sample may be the best way to get the information needed for RQ 4 that you had planned on deriving from the survey. A purposeful sample of both fellows and non-applicants (who are in the same academic department as a fellow) will allow a focus on how fellows contribute passively to their academic departments and programs.
3. We believe that the remainder of the survey (after applying the above alterations) could benefit from cognitive pre-testing, and suggest that you write such a procedure into your PRA request. Cognitive lab testing will allow you to find areas of redundancy and increase component validity across the questionnaire. We specifically suggest that you focus cognitive testing on Part II (Graduate School Experiences) Sections C and D, as well as Part III questions 28 and 29. While we understand that you took many of these categories from previous surveys, a greater effort should be made to tie the possible response categories to your research questions while reducing the overall burden placed on the respondents.

Interview Protocols:

We believe further pretesting of the interview protocols is necessary, particularly for flow. Most of the protocols jump from one major topic (say program administration) to another (i.e. the experiences of the fellows) and back; semi structured interviews work best if they feel like a natural discussion. The faculty protocol is the only one that seems to flow well, and should be used as a model.