Attachment 6:
Designing an Incentive Experiment for the NSFG

By James Wagner, PhD, & Mick Couper, PhD, University of Michigan,
William Mosher, PhD, and Van Parsons, PhD, NCHS
April 1, 2013

Summary

This attachment describes an experiment to test the use of a $60 incentive in the NSFG compared with the current $40 incentive. (The current $80 incentive, which is used only in weeks 11 and 12 of each 12-week data collection period and is given to only 8% of respondents in the sample, is not affected by this experiment.) We propose that the experiment be randomized at the segment (neighborhood) level. Based on previous NSFG experiments summarized below, and the literature on incentives, we expect that the response rate will be about 6 percentage points higher in the $60 group than in the $40 group. If that happens, we should have sufficient power to analyze the experiment. We will analyze the results of the experiment for 3 types of outcomes: response rates; costs and cost-related indicators of interviewer effort; and characteristics of respondents as indicators of bias and bias reduction.

Testing a smaller increase in the incentive amount (e.g., from $40 to $50) would likely produce such small differences that we would not be able to obtain a significant difference, or a difference that would make a sufficient long-term improvement in the results for the groups with the lowest response rates, including white women and most groups of men—in response rates, in bias reduction, or in cost reduction.

Background

Response rates for NSFG have declined since the end of the 2006-2010 data collection period. Screener response rates have been steady, so the response rate for the main interview is the main factor in these declines.

The following data show that interviewers in 2011-12 are

- working more hours per interview ,
- visiting households more, and
- getting lower response rates than they were in 2006-2010.

	2006-10	2011-2012
Hours of labor per interview:	9.1 hours	9.5 hours
Average no. of visits to HH per interview:	7.1 visits	8.1 visits
Response rate:	76.6%	72.7%

Table 1: Response rates for the National Survey of Family Growth in 2006-10 and in the first year of interviewing in 2011-2012, by sex and race/ethnicity

	2006-10	2011-12	2006-10	2011-12
	Phase 1 response rate		**Final Response rate**	
Total	**58%**	**57%**	**77%**	**73%**
Male	**57%**	**56%**	**75%**	**72%**
Black	61%	61%	**77%**	**73%**
Hispanic	***55%***	***53%***	74%	73%
White & other	56%	55%	**75%**	**72%**
Female	**59%**	**57%**	**78%**	**73%**
Black	63%	62%	81%	77%
Hispanic	59%	60%	80%	79%
White & Other	***59%***	***54%***	***76%***	***70%***

In 2002, the *Phase 1 response rate* (not shown in table 1, but noted in the 2006 NSFG OMB Submission) was 64%. That rate has dropped to 57% in 2011-12. The phase 1 response rates are now low enough to cause concern--only 57% for both males and females after 10 weeks of effort and an average of 8 in-person visits. Of most concern, however, are the response rates for "white and other" females, which are the largest group of females in the population, and whose phase 1 response rates have dropped to just 54%. Similarly, the phase 1 response rates for Hispanic males have dropped to 53%.

Note that the *final response rates* for white and other females dropped from 76% in 2006-10 to 70% in 2011-12. For white and other males, response rates dropped from 75% to 72%. Response rates for black and Hispanic women are still in the upper 70's, but for all other groups, they are now in the lower 70's. If nothing is done, response rates will probably continue to decline, despite high levels of effort and cost per case. We believe that we need to act now before this problem becomes more serious, and perhaps more difficult to solve.

The phase 1 incentive in the NSFG has been $40 for more than 10 years--since 2002. Under an agreement with the NCHS IRB, this incentive is paid in cash, at the time the interview begins, and is described as a "token of appreciation." It is not described or seen as payment for the respondent's time. Nevertheless, inflation and the passage of time since 2002 have reduced the impact of the incentive, in gaining the attention of respondents, as indicated by the data shown above.

The NSFG has a history of providing incentive payments to respondents, and of testing alternative levels of payment. Note that in each past experiment (described below), where the difference in the incentive was $20 or more (in one case, $30), we found significant differences

in response rates. Incentives in the NSFG are in the form of cash payments at the time the interview begins.

Four major experiments with incentives have occurred in the NSFG:

1) **1993 (Cycle 5) Pretest:** In a field experiment in the 1993 pretest for NSFG Cycle 5, a $20 cash incentive was found to produce a significantly higher response rate (67.4%) than when no incentive was offered (58.9%)—a difference of 8.5 percentage points. For respondents who were offered $20, response rates were higher, and field costs per case were lower than for those who received no incentive.

2) **2001 (Cycle 6) Pretest:** In a field experiment in the 2001 Pretest for Cycle 6, a $20 incentive was contrasted with a $40 incentive. The response rate for those offered $20 was 62%, and for those offered $40, it was 72%—a difference of 10 percentage points. Those receiving the higher amount were also less likely to express objections or reluctance to the interview than those receiving $20.

3) **2002-3 Main Study:** In the 2002-3 Cycle 6 Main Study, a $40 incentive was used, but response rates were still lagging in key groups after seven months of interviewing. NSFG staff requested and received from OMB permission to use an $80 incentive in a half-sample of the cases remaining in the final four weeks of data collection during February, 2003. The $80 incentive raised the weighted response rate from 64% to 79%. The sample in the last 4 weeks had a higher proportion of married women, Hispanic men and women, and full-time workers of both sexes.

4) **2006-7: Phase 2 $50 vs $80:** We tested a $50 Phase 2 incentive vs. an $80 phase 2 incentive to see if it was really necessary to offer $80 to obtain higher response rates and reduce bias, but the results showed that it was necessary to offer the higher amount to reduce bias by bringing respondents with different characteristics into the sample.

Hypothesis:

Based on this extensive set of experiments, and the literature on incentives in surveys (e.g., Singer, 163-177, in Groves et al, **Survey Nonresponse**, 2002), we think it is likely that a $20 increase in the incentive will produce about a 6 percentage point increase in the response rate to the survey, and a reduction in interviewer labor per case, sufficient to pay for most of the cost of the incentive increase. That is why we are asking for a $20 increase in the incentive in Phase 1. We are not asking for any increase in the $80 incentive in Phase 2.

Below, we have described the parameters of the incentive experiment.

The Experiment

Our standard incentive amount is $40. If the experiment runs for 4 quarters (one year), $40 would be tested in about **228 control segments** (or neighborhoods), with about **3,020 eligible persons.**

Our experimental amount is $60. If the experiment runs for 4 quarters (one year), $60 would be tested in about **228 segments** (or neighborhoods), with about **3,020 eligible persons**.

The table below shows the expected number of respondents in 2, 3, and 4 quarters of data collection, and the probable significance of differences that we are likely to find. We considered testing more than one amount, but that approach has serious disadvantages.

First, it decreases the likelihood that we can demonstrate any significant result in any reasonable period of time. This means that the experiment will need to run longer to achieve acceptable levels of significance. This increases the cost of the experiment, and it also has a public relations risk.

In all of our previous experiments in 1993, 2001, 2002, and 2006-7, we know of no case of respondents finding out about the different amounts being offered in different neighborhoods. But with social media now widespread, that risk is probably still low but no longer zero. In any case, as soon as we have clear results, we want to end the experiment to avoid this risk. Having 3 different amounts rather than 2 would increase this risk still further.

Second, there are operational difficulties with multiple amounts. Interviewers need to carefully track which amount is being offered to which unit. Multiple amounts make this more complex and may lead to mistakes—that is, interviewers giving respondents the wrong amounts.

Effect Sizes (Response Rate Increase) and Power

We have done some preliminary power calculations. As a first step, we looked at the design effects for the first year of data collection. We estimated the response rate by dividing the sample of segments in half. We then estimated the design effect on this calculation for each half sample. The estimates were between 2.5 and 3.0, and we used the lower bound (2.5) in the power calculations given below. We incorporate these design effects into our power calculations in **Table 2.**

The specified power is 0.80 and alpha is set at 0.05. We assume that each quarter we identify 1,511 main lines (cases in the sample).

Table 2. Samples Sizes Required for Specified Type I and II Error Rates

Cooperation Rate Increase	Final Response Rate (Screener Rate=0.935)	Simple Random Sample (SRS) n	SRS n * Deff (2.5)	Number of Quarters (2-arm)
0.02	75%	11,508	28,770	19.0
0.04	77%	2,960	7,400	4.9
0.06	79%	1,350	3,375	2.2
0.09	81%	778	1,945	1.3

If we test a $60 incentive, we could expect about a 6% increase in the response rate, and we could expect to finish the experiment with significant results in 2 quarters for the total. (We will also want to be able to show results separately for men and women, because men and women have had different response rates in the past to the same incentive amounts.)

Alternatives

If we tested a smaller amount, such as $50 rather than $60, the literature suggests that the response rate would increase only 3-4%, and it **would take 4.9 quarters (more than a year) to detect a significant effect of a $10 increase in the incentive**. After two quarters, we will have an effective sample size (assuming a design effect of 2.5) of about 1209. **We are underpowered for small effect sizes, which are likely with a $10 increase in the incentive**.

Historical Controls

Another possibility is to use information from the current data collection as "historical controls." In order to do so, we need to make some fairly strong assumptions about the comparability of the response rate and the impact of $40 over time. We can say that the results for the first year of interviewing in the current survey (Sept 2011-Sept 2012) have been quite consistent. **Table 3** shows the screening and main interview response rates for the first four quarters of interviewing in Sept 2011-Sept 2012.

Table 3. Screener and Main Interview Response Rates for the first 4 Quarters of 2011-2012 interviewing in the NSFG

	Q 1	Q 2	Q 3	Q 4
Screener	0.94	0.93	0.93	0.94
Main	0.79	0.77	0.79	0.77

These historical controls could be combined with new data from a $40 arm of the experiment to increase the power of the experiment. (By September of 2013, we expect to have completed 10,000 interviews, so there would be 10,000 controls and about 1-2,000 experimental cases.) There are a number of methods for combining this information. We also note that our analysis should control for effort (number of calls). We think this strategy is less than ideal, but we will

assess its use as a device for looking at the results of the experiment for small sub-groups once the results are in.

Costs

In addition to comparing response rates, a second objective for an experiment is to determine if there are cost differences. The increase in incentive may produce cost savings by reducing the number of calls (interviewer visits to households) required to complete interviews. We have experienced a higher average number of calls per completed interview in 2011-2012.

Assuming that randomization occurs at the segment level (see "note" below), we will use a simple model to compare costs. There are three components to the cost model: *trip costs*, *call costs*, and *incentives*. Simple averages will be used for trip and call costs. The incentive costs for each case are known.

Bias

A third objective for the experiment is to determine if the experimental treatment leads to different estimates by bringing in respondents with different characteristics. Although not a direct measure of bias, such differences indicate that the experimental treatment is giving us more information than the control treatment. We will compare estimates from the two incentive groups in at least 3 ways.

First, we will compare *means and proportions* for the key statistics defined for the study between the two groups. Comparing unweighted means and proportions is simple, but since unweighted estimates are rarely used in practice, it is not the most relevant comparison. We can compare weighted means and proportions using nonresponse-adjusted weights. This approach reflects the prevailing practice and, therefore, compares the estimates that would likely be achieved under either treatment.

Second, in addition to comparisons of means and proportions, we will examine subgroup differences on key statistics of interest between the two incentive groups. This will allow us to evaluate whether the respondents we are bringing in with the higher incentive differ on key measures *in ways that weighting or post-stratification may not fix*. While these analyses will be underpowered relative to the marginal analyses, they will nonetheless be informative of nonresponse bias differences between the two incentive groups.

Third, we note that many analysts are less concerned with means and more interested in regression coefficients from statistical models (e.g. Axinn, et al., Demography 48 (3), August, 2011). We will test for differences between the treatment groups in estimated coefficients from models similar to those found in Axinn, et al.

With respect to bias, we hypothesize that the experimental treatment will produce changes in estimates as described above. However, even if it does not lead to changed estimates, if it leads to a similar sample size at a lower cost, then the experimental treatment will still be preferred. Similarly, the higher incentive may produce a higher phase 1 response rate, which

would produce statistics that have smaller design effects and smaller sampling errors because they would be based on more interviews. This would also be a desirable result.

We also note that the change in incentive may interact with other treatments in our design. In particular, the impact of the second phase design may be affected by the incentive change. If, for example, the increased incentive reduces the effectiveness of the second phase in bringing in new types of respondents (see Axinn, et al. 2011; Peytchev, Peytcheva, and Groves, Public Opinion Quarterly, 74 (2), 2010) and does so at a lower cost, then we may consider modifying the second phase of the design.

Finally, we note that any experiment creates operational costs above the normal operating costs. These costs include:

- programming the contractor's sample management system to manage multiple amounts,
- costs associated with preparing and mailing multiple versions of informational materials and consent forms;
- costs associated with training interviewers to handle multiple versions of materials, and
- contractor costs for design and analysis of the experiment.

We expect that the experiment will show increased response rates from a $60 Phase 1 incentive, cost savings that will at least partially offset the cost of the increased incentives, and bias reduction by bringing in otherwise under-represented groups. If that occurs, better results at modest cost with less bias will be obtained for the next several years.

A Note on the Level of Randomization

In relation to the operational issues mentioned above, any experimental design would need to consider at what level to randomize the units. The available units are: PSU, Interviewer, Segment, and Housing Unit. There is a great deal of correspondence between PSU and Interviewer. There are a few PSUs with more than one interviewer, but in most PSU's there is just one interviewer. Since interviewers within the same PSU share sample, it might be simpler to consider randomizing PSUs.

Randomizing PSUs to treatments greatly reduces the chance of making mistakes with the incentive amount. It also prevents interviewers from prioritizing sample based on differences in incentive amounts. And it virtually eliminates any chance that respondents will discover that others in other neighborhoods are being offered other amounts. *However, it greatly reduces the statistical power of the design*.

Randomizing housing units is the most powerful design. However, it makes the interviewers' task very complex as they need to know the incentive amount for each housing unit. This increases the chances that mistakes will be made. It also increases the latitude that interviewers would have for prioritizing cases based on incentive amounts, which could confound the interpretation of the data.

Randomizing segments is a compromise design that has more power than randomizing PSUs, but less than randomizing housing units. Interviewers will have cases in more than one treatment. They could prioritize cases based on incentive amount, but this would require more frequent trips to segments with the higher amount, and it can be prevented with good management practices. This seems less likely than if the assignment of treatment were at the housing unit level.