APPENDIX I

Power Analysis for the 2013 NSCG Methodological Studies

Power Analysis for the
2013 National Survey of College Graduates (NSCG) Methodological Studies

I. Background

This memo provides minimum detectable differences based on proposed sample sizes and provides clarification on the methodology sampling plan. .

New Cohort Experiments:

- **Priority Mail Study**: One representative sample of 10,000 cases in the experimental group [*Note*: If we use two experimental groups, we may put 5,000 in each]
- **Mode Switching Study**: One representative sample of 2,500 cases in the experimental group
- **Incentive Timing Study**: Four representative samples each of which include 3,000 hard to enumerate cases [*Note*: This totals 52,423 cases, based on selecting representative samples]

Old Cohort Experiments:

- **Incentive Conditioning Study** – All 2010 NSCG Incentivized cases and all 2010 NSRCG cases will be eligible and divided equally among the three treatment groups. [*Note*: Approximately 12,000 cases should be eligible, so each group will have approximately 4,000 cases.]

II. Minimum Detectable Differences Equation and Definitions

To calculate the minimum detectable difference between two response rates with fixed sample sizes, we used the formula from Snedecor and Cochran (1989) for determining the sample size when comparing two proportions.

$$\delta \geq \left(\left(Z_{\alpha^*/2} + Z_{\beta} \right)^2 \left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right) D \right)^{1/2}$$

where:

δ	=	minimum detectible difference
α^*	=	alpha level adjusted for multiple comparisons
$Z_{\alpha^*/2}$	=	critical value for set alpha level assuming a two-sided test
Z_{β}	=	critical value for set beta level
p_1	=	proportion for group 1
p_2	=	proportion for group 2
D	=	design effect due to unequal weighting
n_1	=	sample size for a single treatment group or control
n_2	=	sample size for a second treatment group or control

The Census Bureau standard alpha level of 0.10 was used in the calculations. The beta level was included in the formula to inflate the sample size in order to decrease the probability of committing a type II error. Committing a type II error, claiming there was no difference in response rates across mode groups when a difference was present, would be detrimental to the purpose of this study. With this in mind, the beta level was set to 0.10.

The estimated proportion for the groups was set to 0.50 for the sample size calculations. Setting the estimated proportion at this value was a conservative approach that minimized the ability to detect statistically significant differences when using a given sample size.

Design effects represent a variance inflation factor due to sample design when compared to a simple random sample, and design effects were calculated by examining the weight variation present in all cases in the 2013 NSCG new cohort (4.8293 for new cohort experiments), and the old cohort (4.9577 for the old cohort experiment)[1]. However, the design effect may not be necessary when calculating the minimum detectable difference for the methodology experiments, particularly in the new cohort. The full sample is the frame from which the experiment samples are selected, and each sample is designed to be a representative systematic random sample. Because all experiment samples and the control will be representative, the weight distributions

[1] Census Bureau staff provided these design effects based on NSCG data.

and weighted response influence distributions[2] should be similar throughout all samples, negating the need to include a design effect. To explain it another way, because we are including cases of all types in all experiment samples, we do not expect to see a weight-based or sampling-based effect on response in any of the samples, and thus, we do not need to account for it across samples. However, for the sake of completeness, minimum detectible differences were calculated two ways, both including and ignoring the design effect.

III. Pairwise Comparisons and the Bonferroni Adjustment

The number of pairwise comparisons included in the initial evaluation is one (treatment vs. control). However, for some of the experiments (the incentive timing and possibly the priority mail), the number of pairwise comparisons increases. In addition, if all pairs were compared, the number of pairwise comparisons increases to 21 or 36. Rather than only comparing a treatment to the control to detect statistically significant differences, treatment groups can be compared to each other because there is more than one treatment group. In these instances, the α^* is adjusted to account for the multiple comparisons.

The Bonferroni adjustment reduces the overall α by the number of pairwise comparisons so that, when multiple pairwise comparisons are conducted, the overall α will not suffer. The formula is as follows:

$$\alpha^* = \frac{\alpha}{n_c}$$

The adjusted alpha α^* is calculated by dividing the overall target α by the number of pairwise comparisons, n_c. It is worth noting that, despite being commonly used, the Bonferroni adjustment is very conservative, actually reducing the overall α to below initial targets. Below is an example showing how the overall α is calculated using the Census standard alpha level of 0.10, the Bonferroni adjustment, and 25 pairwise comparisons.

$$\alpha_{overall} = 1-(1-\alpha^*)^{n_c}$$

$$\alpha_{overall} = 1-(1-0.004)^{25} = 0.095 < 0.100$$

$\alpha_{overall}$ is the resulting overall α after the Bonferroni correction is applied;
$\alpha_{target} = 0.100$, and is the original target α level;
$n_c = 25$, and is the number of comparisons
$\alpha^* = \alpha_{overall} / n_c = 0.004$, and is the Bonferroni-adjusted α

[2] See section B for an explanation of weighted response influence being used within the 2013 NSCG methodological studies.

So the Bonferroni adjustment actually overcompensates for multiple comparisons, making it more likely that a truly significant effect will be overlooked.

Initial sample sizes were provided by NSF in Section I, and are used in the formula. All minimum detectable differences using the Bonferroni adjustment were calculated for the new cohort and are summarized on pages 7-8 in table form.

IV. A Model-Based Alternative to Multiple Comparisons

Rather than relying on the Bonferroni adjustment for multiple comparisons, effects on response, cost per case or other outcome variables could be modeled simultaneously to determine which treatments have a significant effect on response.

All sample cases, auxiliary sample data, and treatments are included in the model below, which predicts a given treatment's effect on response rate (or other outcome variable).

$$y = \beta_0 + \vec{\beta}_1 \vec{\mathrm{I}} + \vec{\alpha}\vec{X} + \varepsilon$$

Assuming response rate was the outcome variable of interest:

- y is the average response propensity (response rate) for the entire sample;
- β_0 is the intercept for the model;
- $\vec{\beta}_1$ is a vector of effects, one for each treatment
- $\vec{\mathrm{I}}$ is a vector of indicators to identify a treatment in $\vec{\beta}_1$
- $\vec{\alpha}$ is a scalar vector
- $\vec{X}$ is a matrix of auxiliary frame or sample data
- ε is an error term

After data collection is complete, the average response propensity is equal to the response rate. In the simplest case, no treatment has any effect (the 2nd term would drop out), and no auxiliary variables explain any of the variation in response propensities (the 3rd term would drop out). In that case, the average of the response propensities, and thus the response rate would just equal:

$$y = \beta_0 + \varepsilon$$

However, a more complicated model gives information about each treatment's effect (2nd term) while taking into account sample characteristics (3rd term) that might augment or reduce the effect of a given treatment.

As a simple example, ignore the error term, and assume the overall mean response propensity was 72%. Also, assume the mean response propensity for a given treatment group was 83%. If only terms 1, and 2 were included in the model (no sample characteristics accounted for), the given treatment appears to have increased the response propensity by 11%. However, if the sample was poorly designed, or if a variable not included in the sample design turned out to be a good predictor of response, there is value in adding the 3rd term. Say auxiliary information added by the 3rd term shows that the cases in this particular sample group are 5% more likely to respond than the average sample case (because of income, internet penetration, age, etc). This would suggest that while the treatment group had a response propensity 11% higher than the average, 5% came from sample person characteristics, and only 6% of that increase was really due to the treatment.

The benefits of this method over multiple comparisons are numerous. First, the number of degrees of freedom taken up by the model is the number of treatment groups plus one for the intercept. This is far fewer than the number of pairwise comparisons that might be conducted. Second, because confidence intervals are calculated around the $\vec{\beta}_1$ values, it is more intuitive to see each individual treatment's affect on outcome measures. Third, as mentioned in the above paragraph, if a variable was forgotten, left out, or was part of response or paradata (something that wasn't known during sample design), it can be controlled for in the model, making significant results more meaningful. While we are striving to ensure the experimental samples are as representative (and as similar) as possible, this ability to add other variables after the fact would lend extra credence to any significant effects.

A possible downside to this method is that it uses response propensities, not the actual response rate. While the mean response propensity after the last day of data collection equals the overall response rate, it is important to take note of how the propensity models are built, i.e., if they are weighted models, weighted response propensities should be used in this model. Again, though, the weights could simply be one of the auxiliary variables included in the $\vec{X}$ matrix.

V. Comments

In general, modeling the main effects is preferable to calculating many pairwise comparisons for reasons that include: a reduced loss of degrees of freedom, an ability to see all main effects with confidence intervals with one model, the ability to control for variables after the sampling and experimenting is complete, and the ability to avoid the over-conservative Bonferroni adjustment which might reject some significant differences as not significant.

However, it is worth noting (from pages 7-8) that even using the Bonferroni adjustment, and conducting all pairwise comparisons, a difference of 5% - 6% in outcome measures should be large enough to appear significant, when the design effect is excluded from the calculations.

Because the experimental samples are all systematic random samples, and should have similar sample characteristics and weight distributions, excluding the design effect seems appropriate.

Minimum Detectable Difference Equation for Response Rates

$$\delta \geq \left((Z_{\alpha^*/2} + Z_{\beta})^2 \left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right) \times deff \right)^{1/2}$$

δ	=	minimum detectible difference
$*\delta$	=	minimum detectible difference without using design effect
α^*	=	alpha level adjusted for multiple comparisons (Bonferroni)
$Z_{\alpha^*/2}$	=	critical value for set alpha level assuming a two-sided test
Z_{β}	=	critical value for set beta level
p_1	=	proportion for group 1
p_2	=	proportion for group 2
$deff$	=	design effect due to unequal weighting
n_1	=	sample size for group 1
n_2	=	sample size for group 2

Methodology Study 1: Priority Mailing/First Class Mailing Test

Option 1: 10,000 Cases in One Treatment Group (First Class Mail)

α^*	=	0.100
$Z_{\alpha^*/2}$	=	1.645
Z_{β}	=	1.282
p_1	=	0.5
p_2	=	0.5
$deff$	=	4.8293
n_1	=	10,000
n_2	=	20,277

δ = 0.0393
$*\delta$ = 0.0179

Option 2: 3,333 Cases in Each of Three Treatment Groups, Each Compared Individually to the Control Group (Multiple Comparisons Ignored)

α^*	=	0.100
$Z_{\alpha^*/2}$	=	1.645
Z_{β}	=	1.282
p_1	=	0.5
p_2	=	0.5
$deff$	=	4.8293
n_1	=	3,333
n_2	=	20,277

δ = 0.0601
$*\delta$ = 0.0274

Option 3: 3,333 Cases in Each of Three Treatment Groups, Each Compared To the Control Using Multiple Comparisons [(3!/2!1!) = 3]

α^*	=	0.033
$Z_{\alpha^*/2}$	=	2.133
Z_{β}	=	1.282
p_1	=	0.5
p_2	=	0.5
$deff$	=	4.8293
n_1	=	3,333
n_2	=	20,277

δ = 0.0701
$*\delta$ = 0.0319

Option 4: 3,333 Cases in Each of Three Treatment Groups, Each Compared To the Control or Each Other Using Multiple Comparisons [(4!/2!2!) = 6] (Smallest Pair Sample Sizes Used)

α^*	=	0.017
$Z_{\alpha^*/2}$	=	2.387
Z_{β}	=	1.282
p_1	=	0.5
p_2	=	0.5
$deff$	=	4.8293
n_1	=	3,333
n_2	=	3,333

δ = 0.0988
$*\delta$ = 0.0449

Methodology Study 2: Mode Switching Test

Option 1: 2,500 Cases in Experimental Group

α^*	=	0.100
$Z_{\alpha^*/2}$	=	1.645
Z_{β}	=	1.282
p_1	=	0.5
p_2	=	0.5
$deff$	=	4.8293
n_1	=	2,500
n_2	=	20,277

δ = 0.0682
$*\delta$ = 0.0310

Methodology Study 3: Incentive Timing

No Option for Just One Treatment Group

Option 1: 13,105 Cases in Each of Four Treatment Group, Each Compared Individually to the Control Group (Multiple Comparisons Ignored)

α^*	=	0.100
$Z_{\alpha^*/2}$	=	1.645
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	4.8293
n_1	=	13,105
n_2	=	20,277

$\delta =$	**0.0360**
$*\delta =$	**0.0164**

Option 2: 13,105 Cases in Each of Four Treatment Group, Each Compared to the Control Using Multiple Comparisons [(4!/2!2!) = 6]

α^*	=	0.017
$Z_{\alpha^*/2}$	=	2.387
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	4.8293
n_1	=	13,105
n_2	=	20,277

$\delta =$	**0.0498**
$*\delta =$	**0.0206**

Option 3: 13,105 Cases in Each of Four Treatment Group, Each Compared to the Control or Each Other Using Multiple Comparisons [(5!/2!3!) = 10] (Smallest Pair of Sample Sizes Used)

α^*	=	0.010
$Z_{\alpha^*/2}$	=	2.575
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	4.8293
n_1	=	13,105
n_2	=	13,105

$\delta =$	**0.0524**
$*\delta =$	**0.0238**

All New Cohort Studies Combined

Option 1: One Priority Mail Treatment Group, All Treatment Groups Compared Using Multiple Comparisons [(7!/2!5!) = 21] (Smallest Pair of Sample Sizes Used)

α^*	=	0.005
$Z_{\alpha^*/2}$	=	2.810
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	4.8293
n_1	=	2,500
n_2	=	10000

$\delta =$	**0.1272**
$*\delta =$	**0.0457**

Option 2: Three Priority Mail Treatment Group, All Treatment Groups Compared Using Multiple Comparisons [(9!/2!7! = 36)] (Smallest Pair of Sample Sizes Used)

α^*	=	0.003
$Z_{\alpha^*/2}$	=	2.965
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	4.8293
n_1	=	2,500
n_2	=	3333

$\delta =$	**0.1320**
$*\delta =$	**0.0562**