APPENDIX I

2015 NSCG Methodological Experiments – Minimum Detectible Differences

Minimum Detectable Differences for the
2015 NSCG Methodological Experiments

I. Background

This appendix provides minimum detectable differences for the proposed sample sizes in each of the 2015 NSCG methodological experiments.

New Sample Experiment:

- Adaptive Design Experiment – 34,000 cases will be selected for the control group and 8,000 cases will be selected for the treatment group.

Returning Sample Experiments:

- Adaptive Design Experiment – 10,000 cases will be selected for the control group and 10,000 cases will be selected for the treatment group.
- Questionnaire Impact Experiment – 60,000 cases will be selected for the control group and 3,500 cases will be selected for each of the three treatment groups.
- Email Reminder Experiment – 45,000 cases will be selected for the control group and 3,500 cases will be selected for each of the three treatment groups.

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 alpha level of 0.10 was used in the calculations. The beta level was included in the formula to inflate the sample size to decrease the probability of committing a type II error. The beta level was set to 0.10.

The estimated proportion for the groups was set to 0.50 for the sample size calculations. This conservative approach minimizes the ability to detect statistically significant differences.

Design effects represent a variance inflation factor due to sample design when compared to a simple random sample. Because all experiment samples and the control will be representative, the weight distributions should be similar throughout all samples, negating the need to include a design effect. We do not expect to see a weight-based or sampling-based effect on response in any of the samples. However, for the sake of completeness, minimum detectible differences were calculated both ways, including and ignoring the design effect[22].

III. Pairwise Comparisons and the Bonferroni Adjustment

The number of pairwise comparisons included in the adaptive design experiment evaluation is one (treatment vs. control). For the other experiments (the questionnaire impact and email reminder experiments), the number of pairwise comparisons increases because treatment groups can be compared. In these instances, α^* is adjusted to account for the multiple comparisons.

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

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

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 below initial targets. An example showing how the overall α is calculated using an alpha level of 0.10, the Bonferroni adjustment, and 25 pairwise comparisons follows:

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

[22] Design effects were calculated by examining the weight variation present in all cases in the 2015 NSCG new sample (5.983 for new sample experiment), and the returning sample (6.174 for the returning sample experiments).

$\alpha_{overall}$ 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

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

Sample sizes were provided by NCSES in Section I of this appendix and are used in the formula. All minimum detectable differences using the Bonferroni adjustment were calculated and are summarized at the end of this appendix 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 is 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

Once 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. If auxiliary information added by the 3rd term shows that the cases in a 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.

This method has several benefits over the multiple comparisons method. First, the number of degrees of freedom taken up by the model is the number of treatment groups plus one for the intercept, which 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 easier to observe a treatment's effect on the outcome measures. Third, variables 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, the ability to add other variables to the model helps control for unintended effects.

The method 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 note how the propensity models are built. If they are weighted models, weighted response propensities should be used in this model. The weights could be added as one of the auxiliary variables included in the $\vec{X}$ matrix.

V. Comments

It is worth noting from the calculations below that even using the Bonferroni adjustment, and conducting all pairwise comparisons, a difference of 3% - 4% 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 Differences for the 2015 NSCG Methodological Experiments

Minimum Detectable Difference Equation for Response Rates

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

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

Adaptive Design Experiment (new sample)

8,000 Cases in Experimental Group

α^*	=	0.100	
$Z_{\alpha^*/2}$	=	1.645	
Z_{β}	=	1.282	$\delta =$ **0.0445**
p_1	=	0.5	$*\delta =$ **0.0182**
p_2	=	0.5	
$deff$	=	5.983	
n_1	=	8,000	
n_2	=	34,000	

Adaptive Design Experiment (returning sample)

10,000 Cases in Experimental Group

α^*	=	0.100	
$Z_{\alpha^*/2}$	=	1.645	
Z_{β}	=	1.282	$\delta =$ **0.0514**
p_1	=	0.5	$*\delta =$ **0.0207**
p_2	=	0.5	
$deff$	=	6.174	
n_1	=	10,000	
n_2	=	10,000	

Questionnaire Impact Experiment

Option 1: 3,500 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	=	6.174
n_1	=	3,500
n_2	=	60,000

$\delta =$ 0.0632

$*\delta =$ 0.0254

Option 2: 3,500 Cases in Each of Three Treatment Groups, Each Compared to the Control Using Multiple Comparisons [= 3]

α^*	=	0.033
$Z_{\alpha^*/2}$	=	2.128
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	6.174
n_1	=	3,500
n_2	=	60,000

$\delta =$ 0.0737

$*\delta =$ 0.0296

Option 3: 3,500 Cases in Each of Three Treatment Groups, Each Compared to the Control or Each Other Using Multiple Comparisons [(4!/2!2!) = 6] (Smallest Pair of Sample Sizes Used)

α^*	=	0.017
$Z_{\alpha^*/2}$	=	2.394
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	6.174
n_1	=	3,500
n_2	=	3,500

$\delta =$ 0.1092

$*\delta =$ 0.0439

Email Reminder Experiment

Option 1: 3,500 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	=	6.174
n_1	=	3,500
n_2	=	45,000

$\delta =$ 0.0638

$*\delta =$ 0.0257

Option 2: 3,500 Cases in Each of Three Treatment Groups, Each Compared to the Control Using Multiple Comparisons [= 3]

α^*	=	0.033
$Z_{\alpha^*/2}$	=	2.128
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	6.174
n_1	=	3,500
n_2	=	45,000

$\delta =$ 0.0743

$*\delta =$ 0.0299

Option 3: 3,500 Cases in Each of Three Treatment Groups, Each Compared to the Control or Each Other Using Multiple Comparisons [(4!/2!2!) = 6] (Smallest Pair of Sample Sizes Used)

α^*	=	0.017
$Z_{\alpha^*/2}$	=	2.394
Z_β	=	1.282
p_1	=	0.5
p_2	=	0.5
deff	=	6.174
n_1	=	3,500
n_2	=	3,599

$\delta =$ 0.1084

$*\delta =$ 0.0436