Monitoring Methods for Adaptive Design in the National Survey of College Graduates (NSCG)

Stephanie M. Coffey[1], Benjamin Reist[1], Michael White[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746

Abstract
One goal of adaptive design is to allocate data collection resources efficiently, rather than exhausting money and time simply to increase response rate. Data monitoring is vital to this effort as it provides updated views of response information and data quality throughout the collection period. More up-to-date information can lead to interventions including: reducing contact attempts for low-impact cases unlikely to respond, switching contact mode to maximize response probability, or even stopping data collection. In the 2013 NSCG, data monitoring will help show the evolving state of data collection and inform interventions in a mode switching experiment.
This paper discusses several data monitoring methods explored for the NSCG using 2010 survey data including: R-indicators, response propensity by data collection type, and benchmarking. Some of these methods employ propensity models, and others rely on daily processing of survey data including nonresponse adjustments and weighting. In addition to discussing practical benefits and shortcomings of the methods, interventions are simulated to show their effect on the final state of data collection.

Key Words: adaptive design, r-indicators, propensity modeling

1. Introduction[1]

The National Survey of College Graduates (NSCG) is a longitudinal survey that collects information on employment, educational attainment, and demographic characteristics of the college-educated science and engineering (S&E) workforce in the United States. The U.S. Census Bureau conducts the NSCG on behalf of the National Science Foundation (NSF). The 2013 NSCG selected its sample using a dual frame design. One frame included respondents to the 2010 NSCG and 2010 National Survey of Recent College Graduates (NSRCG) and is referred to as the “old” cohort, and the other frame included respondents to the 2011 American Community Survey (ACS) and is referred to as the “new” cohort. Cases were eligible for the new cohort sampling frame if they had responded to the 2011 ACS, reported obtaining at least a Bachelor’s degree, were less than 76 years of age, and were non-institutionalized.

The 2013 new cohort includes an adaptive design experimental group, a representative subsample of 4,000 cases that we are using to test adaptive design capabilities. Because data collection for the 2013 NSCG has not ended, this paper will use a subset of 2010 new cohort cases to examine quality measures and simulated effects of adaptive design interventions being used in the 2013 adaptive design experimental group. The 2010 new cohort was sampled out of respondents to the 2009 ACS, making cases in the 2010 new cohort comparable to those in the 2013 new cohort for simulation purposes.

[1] Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau or The National Science Foundation

2. Adaptive Design Background & Motivation for NSCG

Adaptive design provides a framework for dynamically altering data collection strategies during the data collection process, in response to conditions “on the ground” (1). To determine how and when to intervene in data collection, auxiliary data and paradata should be monitored and used in a predictive manner so that resources can be efficiently allocated. Because the basis for adaptive design is the reallocation of limited resources like time and money, surveys that implement adaptive design need to be aware that there is a trade-off between cost, quality, and overall response rate. Making a data collection intervention requires deciding, on a practical level, whether it is “worth it” to go after a given respondent. Monitoring of data quality measures provides information on when to intervene, how to intervene, and if past interventions have had the desired effect. In addition, these measures provide other ways to evaluate data quality in addition to response rate. This paper will discuss quality measures used in the 2013 NSCG, how interventions could be implemented, and in a limited way, their resulting effect on cost and data quality.

One commonality to all surveys implementing adaptive design is that there is a driving force behind the implementation. Shrinking budgets for survey operations as well as an increase in reluctance in the general population to participate in surveys have resulted in declining response rates. This is happening simultaneously with the desire on the part of data users to be able to make reliable estimates for smaller subpopulations. Differential response rates between subgroups increase concerns about non-response bias and variance inflation of estimates for subpopulations with low response rates. Adaptive design seeks to improve survey representation even in the face of falling response rates as a way to maintain data quality while controlling costs.

While NSCG certainly suffers from reluctant respondents, and the desire to make reliable estimates for small subpopulations, another reason NSCG is implementing adaptive design is to reduce the length of time between the start of data collection and the delivery of the final data products. NSF is looking to reduce this time from 28 months in the 2010 NSCG to just twelve months in the future, providing more timely information to data users. The 2013 NSCG has a six-month data collection period. If interventions are made to convert high-impact cases to respondents earlier during data collection, adaptive design can be extended to determine stopping rules, which would allow the data collection time to be shortened even further, while maintaining data quality.

3. Adaptive Design in the 2013 NSCG

3.1 Interventions:

We employed three types of interventions in the 2013 NSCG mode switching experiment. However, data collection was not completed until August 25th, so we could not evaluate the effect of mode switching interventions on cost and quality in production for this paper. To see the effect of mode switching on NSCG data, we simulated these interventions using 2010 NSCG sample cases. The 5,000 cases used as the basis for the simulation were all part of a “Web First” data collection pathway, which was most similar to that used in the 2013 NSCG, and included web invites and reminders for the first seven weeks. In week eight, questionnaires were then sent to individuals to provide a second mode of response. In week twelve, cases were moved to Computer Assisted Telephone Interviewing (“CATI”) to convert reluctant respondents. Finally, in week 23, a subset of 320 cases was mailed a “last chance” incentive in order to increase response rate. The three interventions varied the intensity of data collection targeting from the pre-determined data collection pathway.

First, a case could be moved from web, an inexpensive, passive self-response mode, to CATI, a more expensive, active mode earlier than the pre-determine data collection pathway dictates. This would allow us to increase targeting for cases earlier in the data collection process. Practically speaking, moving cases to CATI could increase cost, but the cost may be worth the possible increase in data quality. If the case is high-impact (e.g., under-represented, or high weight/low propensity to respond) or has different characteristics than the current respondent population, converting a particular non-respondent to a respondent may reduce non-response bias.

Conversely, a case that was in CATI, an active mode, could be put “on hold”, so the case received either only passive mode invitations (web invites or paper questionnaires), or no invitations at all. While this might result in lower response, if the types of cases put on hold are already well represented, or are otherwise low-impact, there are cost and resource savings to be gained. This savings would help offset cases who are moved to more active modes of data collection.

Finally, a case could simply be held in web, a passive mode, and not mailed a questionnaire or moved to CATI. While the case could still respond via web, it is not provided with a second passive mode. This makes response less likely, but cost savings are created initially by not mailing the questionnaire, and additionally, by avoiding keying and other processing inherent in paper response.

3.2 Monitoring Methods to Determine Interventions:

3.2.1 R-Indicators

R-indicators were the main quality measures that determined interventions in the 2013 mode switching experiment. Conceptually, the R-indicator is a distance measure that represents how different the respondent population is from the full sample population, for a given set of variables. Specifically, the R-indicator compares modeled outcome propensities for individual cases against population and subpopulation mean modeled outcome propensities in order to quantify variations in those propensities. If all cases in the sample are equally likely to appear in the respondent population, given a set of model variables, the sample is considered balanced on those variables. "In fact, we view the R-indicator as a lack-of-association measure. The weaker the association the better, as this implies that there is no evidence that non-response has affected the composition of the observed data." (3) This suggests that “selective forces…are absent in the selection of respondents” out of the sample population (2), and non-respondents are missing at random, which reduces the risk of non-response bias. Shouten, et al.(3) provide a basic framework for using R-indicators for data monitoring, and our mode switching interventions grew naturally out of that framework.

R-indicators are based on propensity scores from logistic regression models, so the first implementation step determines what variables to include in the model. The NSCG is fortunate to sample from ACS respondents, and so has detailed frame variables about sample cases (4). We included NSCG sampling stratification variables, as these variables create the subgroups at which NSF and data users most commonly make estimates. Variables in our model included: Age Group, Demographic Group, Highest Degree

Reported, S&E Status[2], Occupation categories, and Gender. This list incorporates variables used in stratification and sampling and those most important for reporting estimates that meet NSF's mission, which is to provide educational, employment, and demographic information on the science and engineering workforce as a whole, and also by gender, disability status, and race/ethnicity (5). If we have a high R-indicator by the end of data collection, then there is no (or little) relationship between those variables and non-response. If these variables did not drive variation in response, estimates can be made at these subgroup levels with little concern for nonresponse bias, even at lower response rates. Once balancing propensities are calculated, four categories of R-indicators can then be calculated and tracked over time.

Full-Sample R-Indicators are based on the weighted standard deviation of the balancing propensities (2), and are calculated as follows:

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}\left(\hat{\rho}_i - \hat{\bar{\rho}}\right)^2}, \tag{1}$$

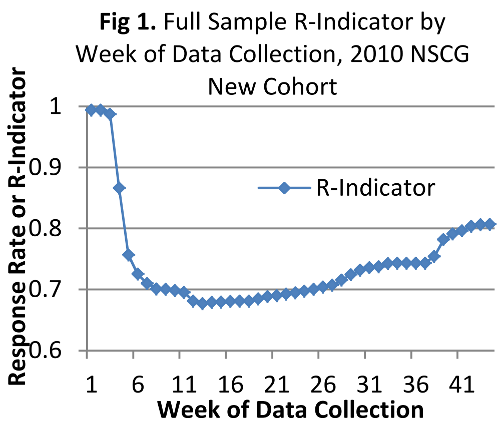
where N is the population size, s_i is an inclusion indicator and π_i is the probability of selection. The individual balancing propensity is $\hat{\rho}_i$, and the overall mean balancing propensity is $\hat{\bar{\rho}}$. As sample cases' balancing propensities diverge, the R-indicator decreases. Full sample R-indicators can take on values [0, 1], with a value of one meaning there is no relationship between variables included in the balancing model and non-response. These indicators can be tracked throughout data collection and compared with other full sample R-indicators, provided the same variables are used in the balancing propensity model.

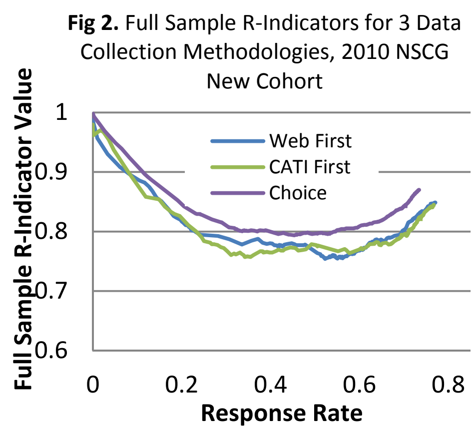
In the NSCG, we used the full sample R-indicator in two ways. First, plotted against time, as in **Figure 1**, we used the R-indicator to evaluate the effect of data collection interventions on representativeness over time. For example, in Week 38, something occurred that caused a larger increase in the R-indicator than occurred at any other time during data collection. If that was a planned intervention (e.g., a reminder letter, sending an incentive, etc.) this graph shows that intervention was successful in increasing representativeness. Second, when the full sample R-indicator is plotted against response rate, as in **Figure 2**, for several comparable survey groups, we could compare R-indicators and response rates, and determine which survey group provided either the highest response rate, the highest R-indicator, or a balance of both. Comparing these groups makes it clear that having a higher response rate does not always result in a more balanced or representative sample. **Figure 2** shows that while both the web first and CATI first data collection methodologies had higher response rates than the choice[3] data collection methodology, the choice group had a higher representativeness.

Practically, these plots can also suggest a time to make an intervention. If the response rate is increasing, but the R-indicator is not changing, that means that while more of the sample is responding, the new respondents are not fundamentally different than previous respondents, based on the variables included in the model. An intervention may be necessary to convert non-respondents that would increase representativeness.

---

[2] S&E Status is defined as either having a degree in, or having an occupation, in an S&E field.
[3] The cases in the choice sample received both a web invite and a paper questionnaire in Week 1, and were allowed to choose which method to respond in.

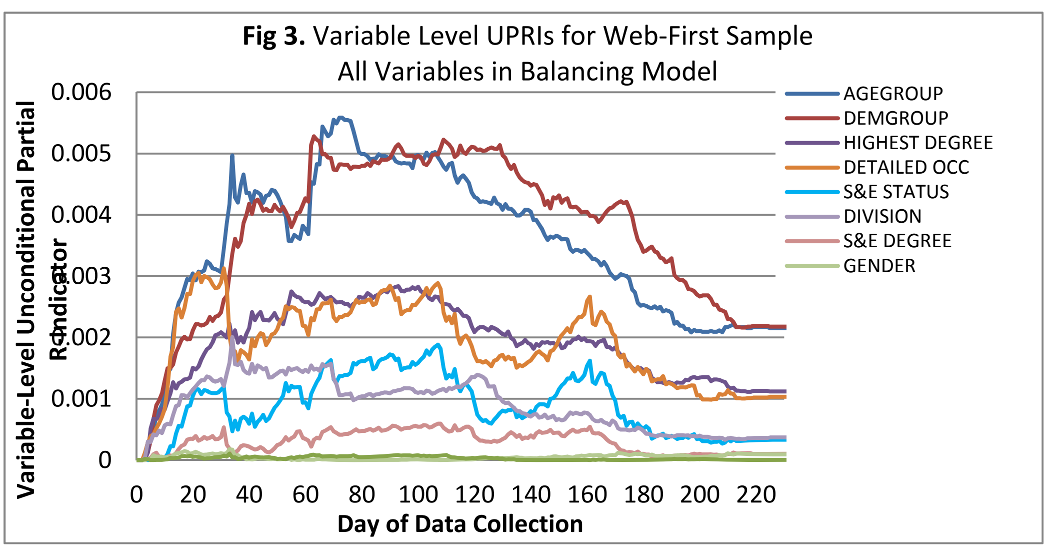




Variable Level Unconditional Partial R-Indicators ("Variable Level UPRIs") measure the between variance of balancing propensities for categories of a variable, and provide information on which variables have the greatest variation in balancing propensities by category, as shown in **Equation 2**. The differences between each value of $\overline{\rho}_{X,k}$, the mean balancing propensity for each category k of variable Z, and $\overline{\rho}_X$, the overall mean balancing propensity are weighted by the proportion of the population in each category, $\frac{N_k}{N}$, and summed to provide a total effect for each variable Z. Variable-level UPRIs, when graphed over time show the contribution of a variable to the lack of representativeness in response, given model X and range in value from [0.00,0.50], with a value of 0.00 meaning that the variable does not contribute at all to the lack of representativeness. If a single variable were to have a value of 0.50, that variable accounts entirely for the lack of representativeness in the model X.

$$P_u(\rho_X|Z) = \sum_{k=1}^{K} \frac{N_k}{N}\left(\hat{\overline{\rho}}_{X,k} - \hat{\overline{\rho}}_X\right)^2 \qquad (2)$$

Figure 3 shows the variable-level UPRIs for variables in the 2010 balancing propensity model for illustration. This figure is revisited in the simulation section. Variables with the largest values should be examined using category-level partial R-indicators, both unconditional and conditional, as these finer metrics can suggest interventions.

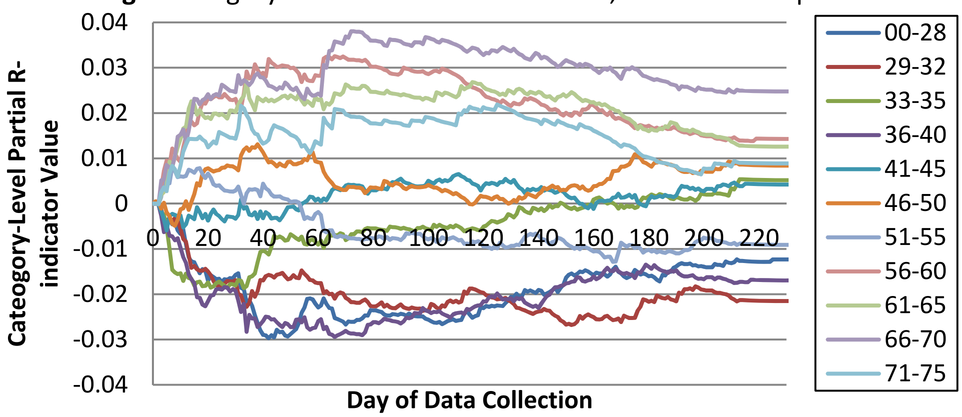


Category Level Unconditional Partial R-Indicators ("Category-Level UPRIs") measure over-or under-representativeness of individual subgroups of variable Z, and so the value is simply each category's contribution to **Equation 2**, as shown in **Equation 3** below.

$$P_u(Z = k, \rho_X) = \sqrt{\frac{N_k}{N}}(\hat{\bar{\rho}}_{X,k} - \hat{\bar{\rho}}_X) \tag{3}$$

Category-level UPRIs can range in value from [-0.50, 0.50], with negative numbers signifying under-representation and positive numbers indicating over-representation. More extreme values are more over- or under- represented, and a value of 0.00 means that the subgroup is not over- or under-represented. Importantly, over-and under-representation is relative, meaning that anytime there is an over-represented group, one or more other groups must be under-represented. **Figure 4** shows category-level UPRIs for AGEGROUP, one of the two variables most responsible for the variation in balancing propensities in the balancing model, shown in **Figure 3**. As an example, by Day 60, the 66-70 year-olds are the most over-represented group, and remain that way throughout the remainder of data collection. Adaptive design interventions seek to work against this entrenching of over- and under-representation to achieve a more balanced respondent sample.

Fig 4. Category-Level UPRIs for AGEGROUP, Web First Sample



Category Level Conditional Partial R-Indicators ("Category-Level CPRIs") were the third type of R-indicator we used. While Variable- and Category-Level UPRIs are based on the variance between categories of a variable, Category-Level CPRIs focus on the within-variance of variable categories. **Equation 4** defines Category-Level CPRIs.

$$P_c(Z = k, \rho_X) = \sqrt{\frac{1}{N-1}\sum_{l=1}^{L}\sum_{U_l}\delta_{k,i}[\rho_X(x_i) - \bar{\rho}_{X,l}]^2}, \tag{4}$$

where, N is the population size, l is the subgroup of variable L, $\delta_{k,i}$ is an inclusion indicator in the subgroup l, $\rho_X(x_i)$ is the individual case-level balancing propensity, and $\bar{\rho}_{X,l}$ is the mean balancing propensity for the same subgroup level. Conceptually, the category-level UPRI calculates the variability of subgroup mean propensities, while the category-level CPRI calculates the variability of propensities within a subgroup. Category-level CPRIs range in value from [0.00, 0.50], with higher numbers signaling more within-subgroup variation of balancing propensities. A value of 0.00 means the subgroup does contribute to the variation in balancing propensities of the variable as a whole.

"…We may interpret $P_c(Z = k, \rho_X)$ as measuring the contribution of variable Z to the R-indicator after controlling for the contribution all remaining variables…" in model X (3). By controlling for the other variables in the balancing model, it is possible to determine whether the variables identified using **Figure 3** and the subgroups identified using **Figure 4** are independently the most significant variables selected. If the category-level CPRI remains high, the variable has an independent effect from the rest of the model variables on representativeness. If the category-level CPRI is close to zero while the unconditional partial R-indicator is high, there is multicollinearity between the variable in question and the rest of the model. Other variables and subgroups should be examined for possible interventions (3).

4. Retrospective Work: Simulating Interventions with 2010 Data

As previously mentioned, the 5,000 cases used as the basis for the simulation all were in a "Web First" data collection pathway. For the simulation, we executed straightforward interventions on this 5,000 case sample at three main points in time. Table 1, below, shows the timing, action, and rationale behind each intervention. These interventions corresponded closely with the interventions made in the 2013 NSCG.

Table 1.

Date	**Simulated Intervention**	**Reasoning**
Pre-Week 7 (Day 41)	Move under-represented non-respondent cases to CATI early	Under-represented cases (may be different than in Week 7) have had 7 weeks to respond in web and have not responded. Move to CATI to actively target these cases.
Pre-Week 8 (Day 48)	Do not mail over-represented non-respondent cases a questionnaire	Over-represented cases do not need additional modes to encourage response. Cost can be controlled by reducing mailings, thereby reducing keying costs that result from paper questionnaires. Respondents still have the option to respond by web.
Pre-Week 13 (Day 81)	Do not put over-represented non-respondent cases in production CATI	Over-represented cases do not need additional modes to encourage response. Cost can be controlled by reducing telephone calls. Respondents still have the option to respond by web. Also reduces respondent burden for over-represented cases.

In the simulation, we only recalculated response rates and R-indicators every 20 days after the first intervention (Day 41, 61, 81, 101, etc.). Further, the simulation ended at Day 181, as after that point, the effect of the incentive interfered with the simulation.

We also made several assumptions in the simulation. First, we made an assumption to account for the increase in response rate of the under-represented groups moved to CATI early in Week 7. We assumed that the response rate would increase to that of the same population in the "CATI First" data collection pathway. Cases in the under-represented groups were selected to be changed from non-response to response to closely match both the increase in unweighted response rate, and the increase in weighted response rate. We also made the assumption that these cases would require an amount of CATI effort (calls, cost) equal to the average CATI effort required for cases in the "CATI First" data collection who were also in the under-represented groups. As an example, if we determined that a case in the Web First population should be changed from a non-respondent to a respondent on Day 101, the number of calls associated with that response would be equal to the average number of calls required in the CATI First group to convert a case in the under-represented population between Days 81 and 101.

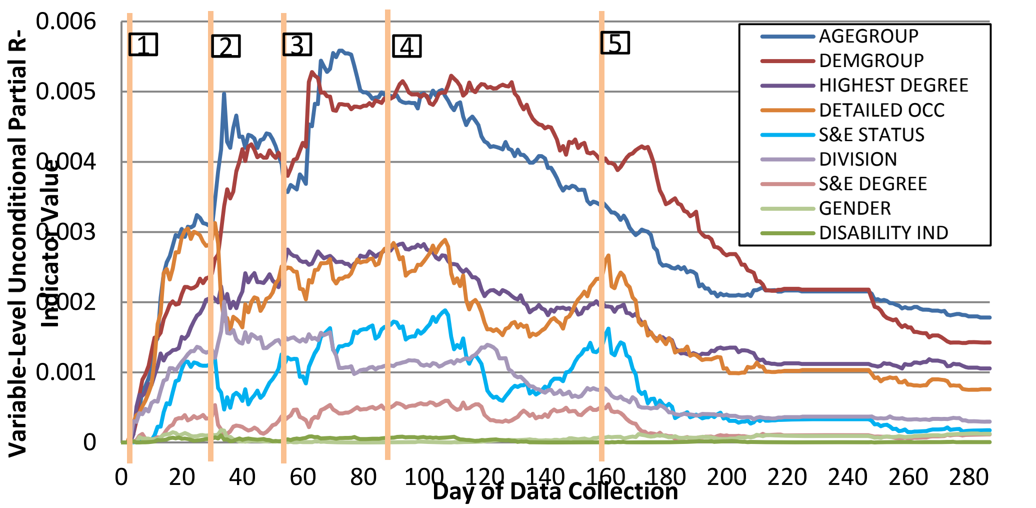
The second key assumption accounted for over-represented cases that were not mailed a paper questionnaire in Week 8. These cases could still respond in web, but were not given a second mode option to encourage response. We assumed that none of these cases would have responded. These sample persons had 8 weeks to respond via web, and did not. This assumption is extreme, as some individuals would have responded after Week 8 to the initial mode of data collection, but it was useful for this simulation. Assuming no response from this group let us evaluate representativeness given the worst-case effect on response rate and the best-case effect on cost.

The last key assumption accounted for over-represented cases not moved to CATI in Week 12. We assumed that none of these cases would have responded. These sample persons had 11 weeks to respond via web, and were the same population that did not receive a paper questionnaire in Week 7. Assuming no response from this group let us evaluate representativeness given the worst-case effect on response rate and the best-case effect on cost.

4.1 Intervention Procedure

Figure 5 shows the variable-level UPRIs for the Web-First sample throughout all of data collection for the 2010 NSCG, with markers corresponding to the following fixed, pre-determined data collection events. Event 1 was the initial web invite; 2 was the reminder web invite; 3 was the paper questionnaire mailing; 4 was the beginning of production CATI, and 5 was the last chance financial incentive mailing.

Fig 5. Variable-Level UPRIs for Web-First Sample, Balancing Model

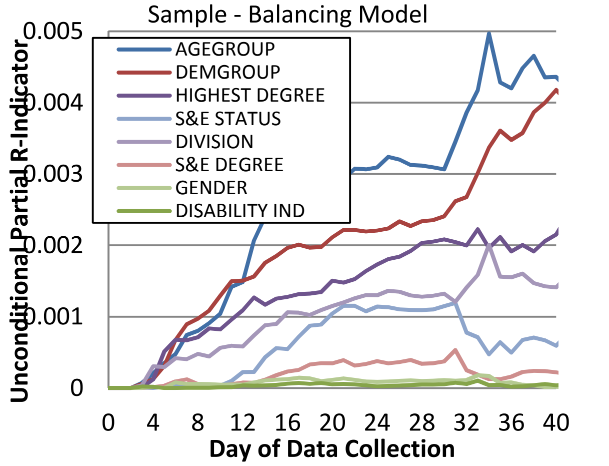


It is clear, from **Figure 5**, that up until Day 84, AGEGROUP and DEMGROUP were the most significant variables driving the variation in propensities, evidenced by higher values of their UPRIs. After Day 84, all of the variables experienced a decrease, and variables that were driving variation in propensities became less strong drivers. This is partially because more modes were introduced, and partially due to the late incentive increasing response among under-represented groups. So eventually, the drivers became weaker as CATI and the incentive converted reluctant respondents. The data collection period was over 9 months long, and CATI, which is extremely expensive, lasted 6 months. Ideally, rather than follow a predetermined data collection flow that lasts nearly 9 months, interventions would be executed to achieve a balanced sample in a shorter period of time, retaining data quality while reducing data collection time, and possibly cost.

4.1.1 Day 41 Intervention:

Looking at the variable-level UPRIs through Day 48 in **Figure 6**, AGEGROUP and DEMGROUP increased much more quickly than the other variables. This means that certain subgroups of each of these variables have mean balancing propensities that are much different than the overall mean balancing propensity. The goal of balancing is for all cases to have similar propensities.

Fig 6. Variable-Level UPRIs for Web-First Sample - Balancing Model



In order to see which subgroups were over- or under-represented, we used the category-level UPRIs, shown in **Figures 7 and 8**. **Figure 7** shows that the four youngest age groups, ranging from 00-40 quickly became under-represented, while the four oldest age groups, ranging from 56-75 were over-represented. **Figure 8** shows that whites were over-represented, while blacks were under-represented in the respondent population.

Fig 7. Category-Level UPRIs for AGEGROUP - Web First

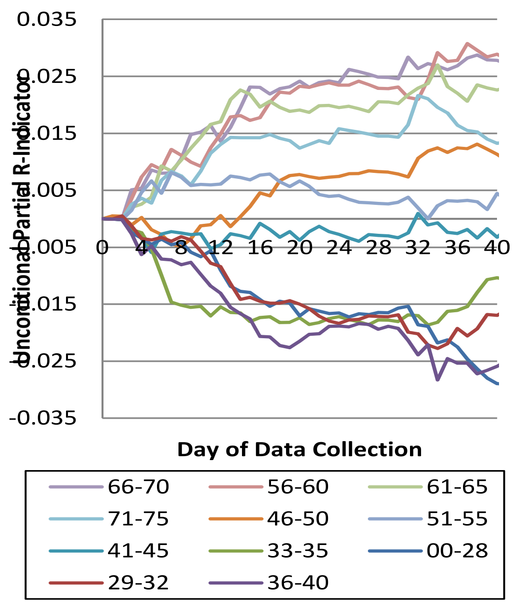
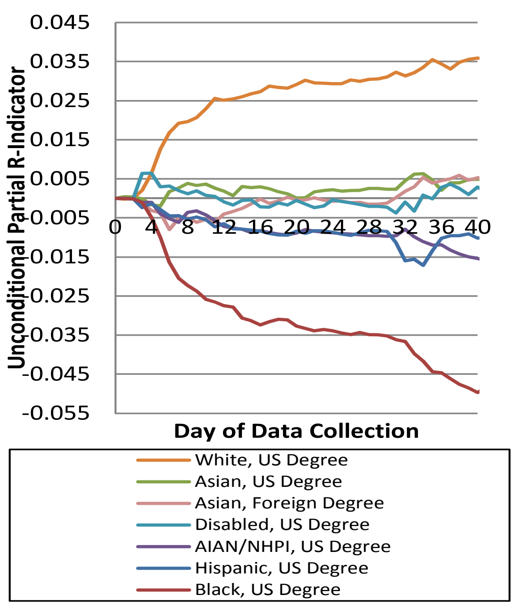
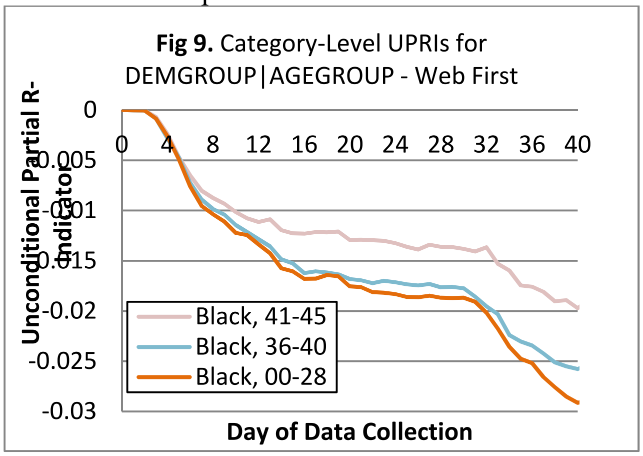


Fig 8. Category-Level UPRIs for DEMGROUP - Web First



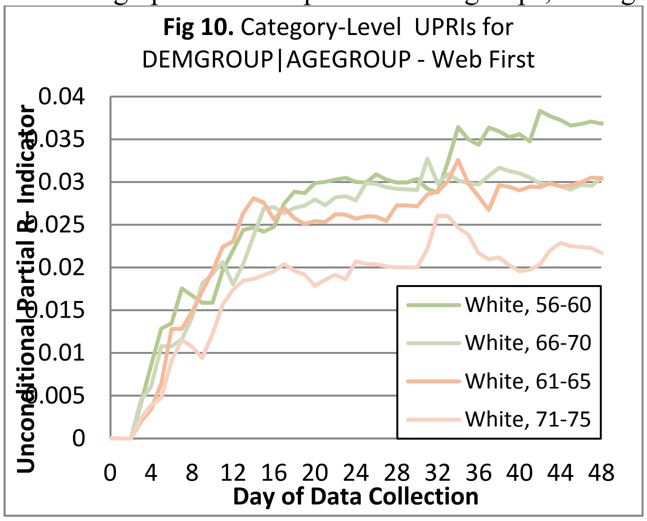
One of the benefits of R-Indicators is that new subgroups can be created, and R-indicators can be recalculated to pinpoint specific subgroups that should be targeted. Here, rather than moving all young sample persons to CATI or moving all black sample persons to CATI for additional targeting, we looked at the cross of AGEGROUP and DEMGROUP to identify which subgroups should be targeted with CATI. **Figure 9** displays the new subgroups by day of data collection. Only the subgroups with the most negative category-level UPRIs were included for clarity, as we were only interested in intervening with the most under-represented cases.



As a result of this information, we intervened by sending the 134 non-respondent cases in the under-represented groups to CATI. A total of 166 cases fell into the under-represented groups shown in **Figure 9**, and only 32 had responded by Day 41. As mentioned on page 10, we assumed the unweighted and weighted response rates for those moved to CATI would match the same population in the CATI First group, as the contact strategy would have been similar after Day 40. ". Cases were selected to have their response status changed from a non-respondent to a respondent every 20 days to reflect the unweighted and weighted response rates of the CATI first study group from Day 41 through 181.

4.1.2. Day 48 Intervention:

Figure 10 shows a similar graph of over-represented subgroups, through Day 48.



Paper questionnaires were mailed on Day 49. We wanted to target over-represented cases by reducing the contact attempts they received. As a result of this graph, we identified the

250 non-respondent cases in these over-represented groups and withheld paper questionnaires. We assumed those cases did not respond for the remainder or data collection.

Monitoring through Day 81:
R-indicators based on these interventions show that representativeness was increased for the sample as a whole through Day 83, and over- and under-representativeness was improved for all groups where interventions took place. **Figures 11** and **12** show the category-level UPRIs for the under-represented cases that were moved to CATI early.

Fig 11. Category-Level UPRIs for DEMGROUP|AGEGROUP - Web First (Actual)

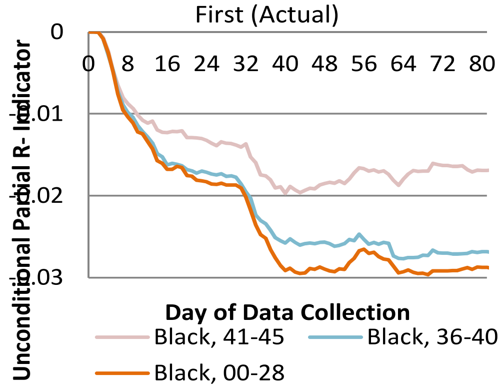
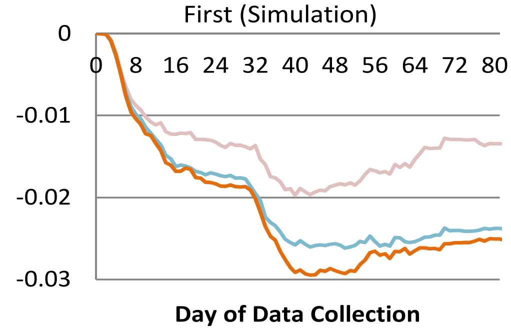


Fig 12. Category-Level UPRIs for DEMGROUP|AGEGROUP - Web First (Simulation)



The actual case results are to the left, while the simulation is to the right. The simulated behavior results in less under-representation in all three groups, though it is only slight. In all three cases, on Day 81, the category level UPRIs are approximately 0.003 higher in the simulation than the actual results. Figures were omitted for the over-represented group for space, but by putting cases on hold and assuming no response, over-representation was reduced appreciably. The smallest improvement in category-level UPRIs was a 0.008 reduction in the 71-75 age group, and the largest reduction was nearly 0.017 in the 66-70 age group. The variable-level UPRIs, shown in **Figures 13 and 14**, also show that both DEMGROUP and AGEGROUP had smaller values in the simulation than in the actual results, meaning their effect on response is reduced through interventions.

Fig 13. Variable-Level UPRIs for Web-First (Actual)

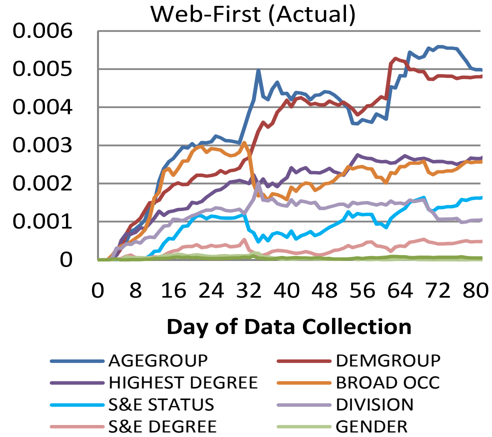
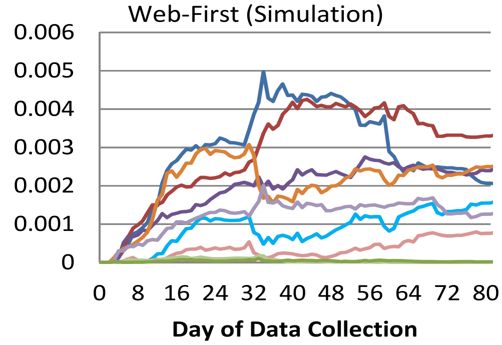


Fig 14. Variable-Level UPRIs for Web-First (Simulation)



4.1.3 Day 83 Intervention

At Day 83, DEMGROUP was the most significant variable, and while an improvement from Day 41, white cases are still over-represented and black cases are still under-represented in the respondent population of the simulation, shown in **Figure 15**.

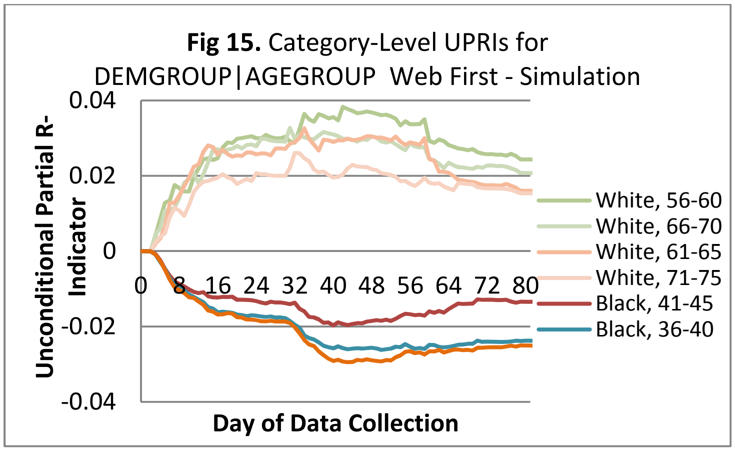


Fig 15. Category-Level UPRIs for DEMGROUP|AGEGROUP Web First - Simulation

Day 83 also coincided with the beginning of production CATI, meaning all non-respondents were moved to CATI, and no data collection pathway interventions could be made on the under-represented cases to increase contact targeting. To continue to improve representativeness, the 250 cases that were not sent a questionnaire were also held out of CATI, reducing contacts made to these sample cases. Again, we assumed that none of these 250 cases responded.

4.1.4 Monitoring through Day 181:

Updating the category level UPRIs for the under and over-represented populations showed that representativeness continued to change due to interventions, to the point where some groups' statuses shifted from under- to over-represented, or vice versa. **Figures 16 and 17** below show how moving cases to a more aggressive mode of data collection does increase response and, therefore, representativeness in the respondent population. By Day 181, the partial R-indicator for the 00-28 black population increased by 0.007, the greatest increase of the three under-represented populations.

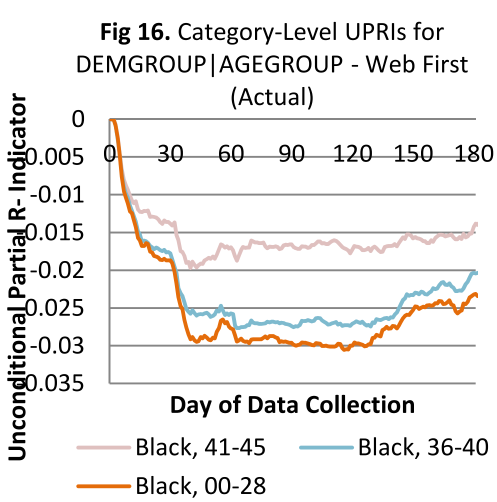


Fig 16. Category-Level UPRIs for DEMGROUP|AGEGROUP - Web First (Actual)

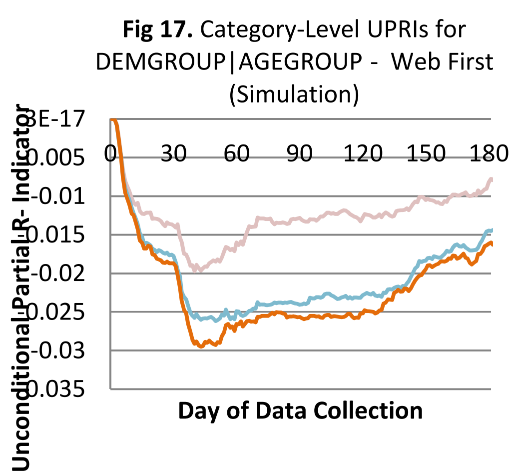


Fig 17. Category-Level UPRIs for DEMGROUP|AGEGROUP - Web First (Simulation)

The over-represented groups showed much larger improvements due to interventions, shown in **Figures 18 and 19**. As previously mentioned, the assumption of no response for over-represented cases put on hold is extreme, but the greater point is that interventions can drive significant reductions in the category-level UPRIs earlier in data collection than a standard data collection strategy.

Fig 18. Category-Level UPRIs for DEMGROUP|AGEGROUP - Web First (Actual)

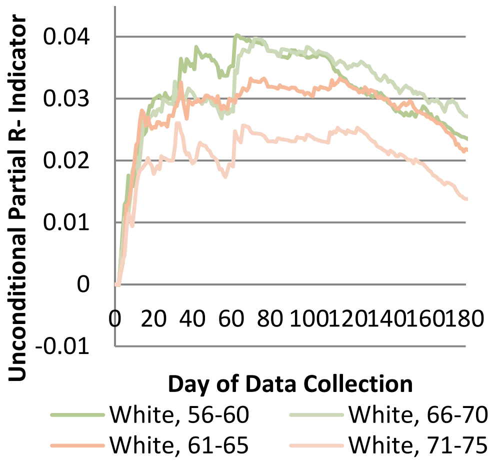
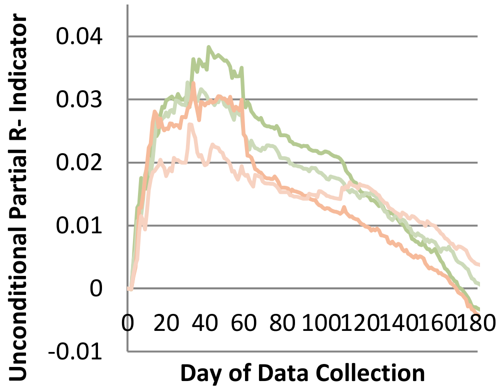


Fig 19. Category-Level UPRIs for DEMGROUP|AGEGROUP - Web First (Simulation)



Another result of the interventions is shown in the variable-level UPRIs. **Figures 20 and 21** show that the interventions have reduced the effect of AGEGROUP and DEMGROUP on response propensity to a similar level as the remainder of the variables. At the same time, no other variables have increased to become strong drivers of response.

Fig 20. Variable-Level UPRIs for Web-First (Actual)

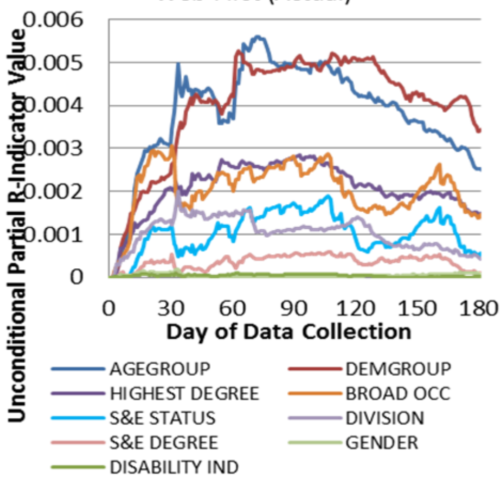
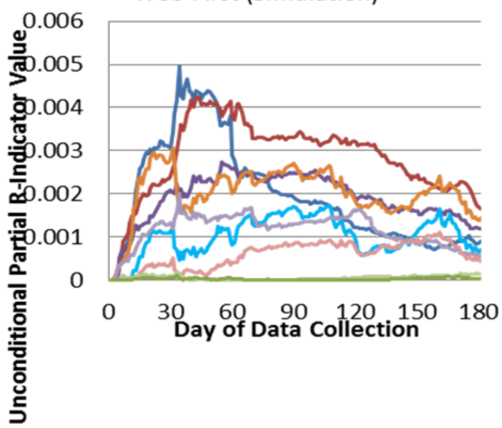


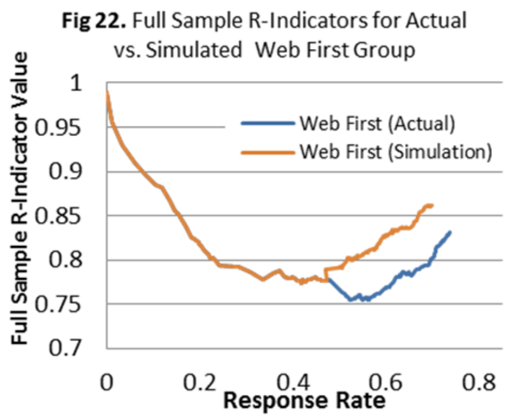
Fig 21. Variable-Level UPRIs for Web-First (Simulation)

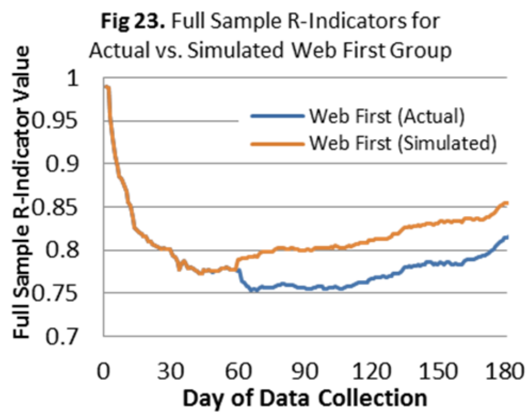


All of these graphs together suggest that the respondent population as a whole became more representative of the full sample population in the simulation.

5. Results: Effect on Overall Representativeness, Response Rate, and Cost Metrics

While a representative sample and balanced response in all subgroups of interest are desirable, if interventions to achieve these goals either reduce the response rate to unacceptable levels, increase the total survey cost significantly, or increase the data collection time period, data monitoring for interventions may not be a viable solution. **Figure 22** shows the overall full sample R-indicator plotted against response rate for the Web First group, comparing the actual results and the simulation through Day 181. While the simulation dataset ultimately had a lower response rate by Day 181, the simulation was more representative. Once the response rate rose to 55%, the R-indicator for the simulation was generally more than 0.05 higher than that of the actual data. **Figure 23** compares the R-indicator over time for the actual and simulated data collection period. **Figure 23** shows that the R-indicator of the simulation increased while the R-indicator of the actual results continued to decrease for approximately 50 more days, until Day 108.





During that time, the simulation was improving the representativeness of the respondent population, ensuring it looked more like the full sample population, while the actual population was receiving "more of the same" types of respondents, making the respondent population look less like the full sample population. While the simulation resulted in higher representativeness throughout the survey, it also resulted in a lower weighted response rate. At Day 181, the response rate for the actual data was 73.86%, while the response rate for the simulation data was 69.95%, a difference of 3.91%.

Lastly, the change in effort, which translates to cost, showed that interventions can be managed to control cost, or at least to ensure that interventions do not necessarily require more resources. Mailing costs and keying costs were reduced by not mailing 489 questionnaires over two questionnaire mailing periods to over-represented cases, and not having to key resulting questionnaires. CATI costs were reduced by keeping the 250 over-represented cases out of production CATI. The under-represented cases moved to CATI early resulted in an additional 1,004 calls being made, of which 94 were complete interviews before production CATI began. **Table 5** summarizes the cost differences.

Table 5

Action	**# of Cases**	**Savings or Expenditure**	**Quantity**	**Cost Difference**
Do not mail questionnaires	250	Savings	Avoid Priority Mailing 489 Questionnaires ($4.96/case)	-$2425.44
			Avoid Keying, Coding, Editing 86 Questionnaires ($22.63/case)	-$1946.18
Do not send cases to CATI	250	Savings	Avoid 1,677 Telephone Calls, 58 Responses ($2.86/call)	-$4796.22
Send cases to CATI Early	134	Expenditure	Make 1,004 Telephone Calls, 94 Responses ($2.86/call)	$2871.44
Total Cost Savings				**$6296.40**

In this specific instance, adaptive design was able to increase representativeness with only a slight loss to response rate, for less money than the actual data collection pathway. In addition, if the 3.91% loss in weighted response rate was acceptable, data collection could have been cut off at Day 181, and the overall R-indicator would have been higher than the R-indicator of the actual results after 286 days of data collection, saving nearly 3 months of data collection, and additional costs during that time.

6. Conclusion

How data monitoring and interventions are used depend on the requirements of the survey sponsors or data users. If resources are not an issue, then perhaps it makes more sense to simply increase targeting of under-represented cases, and worry less about reducing resources spent elsewhere. Decisions should be made based on a given survey's needs. This simulation showed that adaptive design using R-indicators can be a useful targeting method to:

- increase representativeness of the population through case targeting;
- control costs through managing and reallocating resources as opposed to just increasing total resources used on specific target groups;
- attain the same (or higher) level of representativeness with less response;
- attain the same (or higher) level of representativeness in less time.

It is important to keep in mind that adaptive design stems from operational needs, and so data monitoring and the resulting decisions made are directly related to how often data monitoring is used to assess the state of data collection (1). However, it is possible to make decisions using data monitoring, and the results can be quantified through visible improvements in measures like R-indicators.

References:

[1] Axinn, W. Link, C. Groves, R. "Responsive Design Demographic Data Collection and Models of Demographic Behavior." *Demography*. 48 (2011): pp 1127 – 1149.

[2] Schouten, B. Cobben, F. Bethlehem, J. "Indicators for representativeness of survey response." *Survey Methodology*. 35.1 (June 2009): pp 101 – 113.

[3] Schouten, B. Shlomo, N. Skinner, C. "Indicators for monitoring and improving representativeness of response." *Journal of Official Statistics*. 27.2 (2011): pp 231 – 253.

[4] Reist, B. Hall, D. Walker, J. White, M. "The 2010 National Survey of College Graduates Weighting: Leveraging the 2009 American Community Survey." *2013 Joint Statistical Meetings Proceedings* (Submitted for Publication).

[5] Falkenheim, Jaquelina. "Women, Minorities, and Persons with Disabilities in Science and Engineering." Retrieved June 5, 2013 from http://www.nsf.gov/statistics/wmpd/2013/how.cfm

[6] Finamore, J. "NSCG Mode Effects - Cost Spreadsheet (12-14-11)." U.S. Census Bureau Internal Document. December 2011.