National Center for Science and Engineering Statistics

Title: Exploring Adaptive Design Options for the Survey of Doctorate Recipients

Date: August 2021
Final Report

Contractor Awardee: University of Michigan with University of Maryland
Contract Number: 49100420C0020

Exploring Adaptive Design Options for the Survey of Doctorate Recipients

Final Report: August 31, 2021

James Wagner, Brady West, Ai Rene Ong, Robert Schultz

Section 1. Background

The Adaptive Survey Design (ASD) team was engaged to develop adaptive survey design approaches for NCSES-sponsored surveys. Our team focused on the Survey of Doctorate Recipients (SDR), as ASD has been employed extensively on the National Survey of College Graduates (Coffey, et al. 2019). The SDR has a web-mail-telephone design. Key decisions related to the survey include:

- When to switch from one mode to the next
- When to stop effort on cases or, put differently, re-allocate effort to cases judged to be more important

The approach was to identify rules that formalize these decisions statistically. Once these rules were defined, they were to be tested via simulation on existing data. Each of the rules requires inputs. These inputs are predictions about survey outcome variables and survey costs. Our initial work focused on predictions of survey outcome variables (salary and employed by an educational institution) and propensity models (related to costs). With these inputs, it is possible to simulate the results of the proposed decision rules using existing SDR survey data and paradata.

There are some limitations to this approach. First, we can't simulate outcomes for which we do not have data. For example, we do not have randomized assignment of mode sequences in the data. Therefore, we can't use these data to identify whether it is possible to optimally assign different mode sequences to different subgroups with the sample. Second, we did not have direct cost information. Mainly, we did not know the cost of an email invitation, a mailed invitation, a mailed survey, or the costs of CATI attempts and interviews. Instead, we estimated these costs from a variety of published and unpublished resources. Third, the work is based on the 2017 SDR design. It may be that the approach is the most relevant feature of this report. The 2023 SDR may have a different design, different costs, and the behavior of panel members may have changed. Therefore, making design decisions for that iteration of the SDR may involve any or all of the following:

- Repeating the analyses and simulations in this report, but using more recent data;
- Updating simulation approaches to account for design changes; and
- Implementing experiments that will be the basis of future design changes not considered here.

This report will be structured in the following way:

- Section 2. Proposed ASD rules for the mode switch and stopping rule approaches.
- Section 3. Data Management required for the modeling steps.
- Section 4. Description of the simulation studies.
- Section 5. Future Directions.

Section 2. Proposed ASD Rules

Mode switch.

This rule will govern when the mode switch occurs. In the 2017 iteration of the SDR, panel members were assigned to a starting mode. This starting mode was not randomized. It was set based upon either an expressed preference of the panel member (i.e. response to something like the following question: "which mode do you prefer to respond in?") or the observed mode in which they responded at the last wave. The starting modes are either: 1) web, 2) mail, or 3) CATI.

The starting modes were then used to determine the sequence that would be applied to each case. Table 1 shows the three sequences that were the initial plan.

Table 1. Sequences based on starting modes

Starting Mode:	Web	Mail	CATI
Second Mode:	Mail	Web	Web
Third Mode:	CATI	CATI	Mail

Even though these assignments are not randomized, they might be used to explore the possibility that one of these sequences may work better than another. This would require that we believe we are able to model the selection of starting mode. Given the lack of randomization, an experiment would need to be conducted to verify any findings.

Another approach is to turn the mode switch decision into a cost problem. Essentially, the approach is to answer the question -- what is the least expensive path to a completed interview? We will look at expected future costs at each point and identify the lowest cost path to completion. This will require assumptions about costs and predictions of probability of response for each attempt. These predictions will be made by each of the modes in the sequences in Table 1. For example, there are two emails associated with the web starting mode. Then there are three mailed attempts associated with the mail mode. Finally, there are six telephone attempts associated with the CATI mode.

Using these data, we will compare the per interview cost of the next attempt to the attempt immediately following that. Using notation, if *t* denotes the attempt that was just completed, we will compare the costs of attempts at *t+1* and *t+2*. If the current attempt (*t+1*) is more expensive (per expected interview) than the next attempt (*t+2*), then we skip the current attempt (*t+1*) and go directly to the next attempt at *t+2*. It might be that the costs of email are so low that it is difficult to identify cases where switching to mail will make sense. However, if CATI is -- at least in some cases -- much more effective than mail, then switching from mail to CATI might happen more frequently. We could also consider a similar rule that would switch to the next mode if the current mode is more expensive. For example, in the CATI start mode, we might skip CATI and go to mail if interviewing is less expensive by mail.

We note that this assumes that there are no negative synergistic effects between the modes. For example, does offering web first make mail less effective? In other words, are there persons who will complete a mailed survey if the request for mailed completion comes first, but NOT if it comes after the

request to complete a web survey? It may be that there are such persons, although several studies have had difficulty finding them. Our approach assumes that these negative impacts don't happen. An alternative approach would be to randomize mode sequences and then see if a class of persons exists for whom these kinds of synergistic effects occur.

In addition to probability models, we also need some information about costs. We have outlined some cost assumptions that we use that allow us to work out this problem. These are described in the next section. We note that different cost estimates might produce different results.

We will use a different approach for predicting costs in the stopping rule simulations. The approach will predict all future costs at the case level. In order to define this approach, we will use the following notation:

$C_{m,r}$ = Expected costs for mode *m* with result *r*. The available modes are *m*={web, mail, CATI}. The costs are the same independent of the attempt number. For telephone, the costs do depend upon the outcome. Therefore, the result options are only relevant for telephone. For telephone, we assign different costs to attempts that result in an interview and those that do not. Therefore, for telephone, we have two results: r={interview, no interview}.

$R_{m,t}$ = An indicator for response by mode and attempt number *t*. The available modes and attempts (by mode) are the same as those for the expected costs.

Using this notation, the following are the cost estimates for an interview in each mode with T_m attempts for each mode. The number of attempts we will use in our stopping rule simulation for each mode are presented later in Table 10.

$$\widehat{C}_{web,IW} = \sum_{t=1}^{T_m} \left[\prod_{t=2}^{t} \left(1 - \Pr(R_{web,t-1}) \right) \right] \widehat{C}_{web}$$

$$\widehat{C}_{mail,IW} = \sum_{t=1}^{T_m} \left[\prod_{t=2}^{t} \left(1 - \Pr(R_{mail,t-1}) \right) \right] \widehat{C}_{mail}$$

$$\widehat{C}_{CATI,IW} = \sum_{t=1}^{T_m} \left[\prod_{t=2}^{t} \left(1 - \Pr(R_{CATI,IW,t-1}) \right) \right] \left(\frac{C_{CATI,IW,t}}{\Pr(R_{CATI,IW,t})} + \frac{C_{CATI,NonIW,t}}{1 - \Pr(R_{CATI,IW,t})} \right)$$

Thinking about the cost of a web survey, for the first attempt, the term in square brackets is 1.0. The predicted costs are then a function of the cost of each attempt, and the probability that attempt will be required (i.e. 1 minus the probability of an interview). The formula presents total expected costs based on the probabilities of a completed interview at each attempt and the probability of not being interviewed on any of the previous attempts. CATI costs are a function of the probability of an interview since an interview takes more time than attempts that do not result in an interview.

We will estimate *Pr($R_{m,t}$=1)*. We know for each case whether it completed and how many attempts it received. Using this info we can "reconstruct" the history of all attempts across mail, email, and telephone. We know if a case was censored (i.e. never completed an interview) or if they were interviewed we know (at least approximately so) when that interview occurred. Therefore, we were able to estimate discrete-time hazard models that will allow us to estimate *Pr($R_{m,t}$=1)*. The estimated coefficients from these models are presented in Tables 6, 7, and 8. There is one table for each starting mode (web, mail, CATI). In each table, models were estimated for each mode within the sequence. For example, for the web starting mode cases, the first model estimates the probability for each request to complete a web survey. Then, among those cases that did not complete a web survey, there is a model predicting the probability of response for each attempt to request a completed mailed survey. Finally, among those cases that do not respond to either web or mail, there is a model predicting the probability of response at each CATI attempt.

Since we are fixing the costs of attempts, the predicted costs per interview by mode will vary with the probability estimates. We would need to find some cases with very low probabilities of completing by web and relatively high probabilities by mail in order to decide that we should change cases from web to mail earlier than prescribed by the protocol.

For the mode switch simulation we will plan to implement a simple rule that looks at the expected cost of two actions: 1) make the current attempt (*t+1*), or 2) make the next attempt (*t+2*). The following formulae show how these costs will be estimated:

(1) $$\widehat{C}_{m,t+1} = \frac{\widehat{C}_m}{Pr\left(R_{m,t+1}=1\right)},$$ when *t+1* is within the same mode, or

(2) $$\widehat{C}_{m+1,t+1} = \frac{C_{m+1,t+1}}{Pr\left(R_{m+1,t+1}=1\right)},$$ when *t+1* is within the next mode,

where *m+1* denotes the next mode in the sequence. For the next attempt, we can substitute *t+2* for *t+1*. In words, the cost of an interview on the current attempt (*t+1*) is the cost of the next attempt divided by the probability of an interview on that attempt. These "per interview" cost calculations are then used as the basis of a simple rule. Make the attempt with the lowest cost, that is choose the action based on $min\left(\widehat{C}_{m,t+1}; \widehat{C}_{m,t+2}\right)$. Under this simple rule, if the cost of an interview for the current attempt is more expensive than the cost of the next attempt, then skip the current attempt and make the next attempt..

We can further explore a similar rule that would compare the total cost (per interview) of the attempts under the current mode to the costs per interview of the next mode. This rule looks farther into the future than the next attempt and might see additional cost savings.

Stopping rule.

We can apply a similar rule to the one that we deployed on the Health and Retirement Study (HRS). That rule minimizes a function of cost and mean-squared error. Both costs and the survey data are predicted. The following description is adapted from a manuscript in preparation on the rule and its experimental implementation on the HRS.

At a given point in time, there are n_0 sampled elements successfully interviewed and $n - n_0$ elements remaining to be interviewed. We take as our optimizing function ψ , which is a product of the cost, denoted by C (observed/"sunk" and predicted), and a squared error loss function for the sample design, denoted by V. The latter (V) is also based on a combination of the n_0 observed elements and predictions for the $n - n_0$ unobserved elements. For a given stage with n_0 sampled elements successfully interviewed and $n - n_0$ elements remaining, the current value of ψ_0 is given by

$$\psi_0 = \hat{C}\hat{V} \qquad (1)$$

Assuming SRS, $\hat{C} = \sum_{i=1}^{n_0} C_i + \sum_{i=n_0+1}^{n} \hat{C}_i$ and $\hat{V} = \frac{1}{n(n-1)}\left[\hat{Y}_2 - \frac{1}{n}\hat{Y}^2\right]$, where $\hat{Y}_2 = \sum_{i=1}^{n_0} y_i^2 + \sum_{i=n_0+1}^{n} \hat{Y}_i^2$, $\hat{Y} = \sum_{i=1}^{n_0} y_i + \sum_{i=n_0+1}^{n} \hat{Y}_i$, and without loss of generality we take the first n_0 elements to be the successful interviews.

We assume some reasonable model for *Y* given available covariates/paradata to generate predicted values, and treat the estimator of the mean using all of the available data as unbiased. In this case, we have selected two key survey variables (*Y* variables), where "key" is defined as frequently used among the scientific products we reviewed or are deemed to be important by NCSES staff. The variables are salary (continuous) and an indicator of being employed by an educational institution. We are using data from prior waves – including previous versions of the same measurement as well as demographic variables – to predict the current wave survey variables. These predictions may be made using a Bayesian approach in order to include information from previous waves about the relationships between predictors and outcome variables. For example, we might predict 2017 variables using 2015 predictors and include priors derived from a model using 2013 data to predict 2015 responses.

If we stop effort on a given sampled unit $j, j = n_0 + 1, ..., n$, we remove the costs associated with this unit, but also introduce a chance of bias by dropping the unit. Our optimizing function is then given by

$$\psi_j = \left(\hat{C} - \hat{C}_j\right)\left(\hat{B}_j^2 + \left(\frac{n}{n-1}\right)\hat{V}\right) \qquad (2)$$

where the squared bias term is $\hat{B}_j^2 = \left(\frac{1}{n-1}\left(\hat{Y} - \hat{Y}_j\right) - \frac{1}{n}\hat{Y} \right)^2 = \frac{1}{\left(n-1\right)^2}\hat{Y}_j^2 + \frac{1}{n^2(n-1)^2}\hat{Y}^2 - \frac{2}{n(n-1)^2}\hat{Y}\hat{Y}_j$, and $\hat{V} = \frac{1}{n(n-1)}\left[\hat{Y}_2 - \frac{1}{n}\hat{Y}^2\right]$ as defined above. We treat the full sample variance estimator as an unbiased estimator of the variance.

We propose to select the value of *j* from among the $n - n_0$ units that are still active at a given time point (where a decision is to be made) that minimizes ψ_j (large cost reduction, minimal bias from dropping that unit) and set this unit aside. That unit defines an initial **set** of size 1, and we can compute the value of ψ_{s_j} that results from including all elements except those in this set, defined by set s_1:

$$\psi_{s_j} = \left(\hat{C} - \sum_{k \in s_j} \hat{C}_k \right)\left(\hat{B}_{s_j}^2 + \left(\frac{n}{n-s} \right)\hat{V} \right) \qquad (3)$$

Next, we take the remaining $n - n_0 - 1$ active units, and compute ψ_j for each of the remaining units. We pick the unit that results in the smallest value of ψ_j, and add this element to the set s_1, forming the set s_2. Now, we compute ψ_{s_2}. We repeat this process until we identify the set of units to drop (denoted by $s_{\min}$) that *minimizes* ψ_{s_j}, or in other words *maximizes* the function $\Delta_{\max} = \psi_0 - \psi_{s_{\min}}$. Effort will then cease on this set of cases at this time point. If $\hat{\Delta}_{\max} \leq 0$, then no units are stopped (that is, $k \equiv \varnothing$).

In order to implement this rule, we need estimates of the survey values and survey costs. First, we will discuss an approach to developing survey costs for the simulation. The proposed approach is based upon the data that are available. Other options could be available to the contractor – for example, it might be possible to track the time spent calling each case and use this to predict how much time specific call attempts take.

In the paradata file, there are three types of contacts: 1) email, 2) mail and 3) telephone. We discuss our approach to estimating costs for each of these types of contact in the following.

Telephone. We know from previous research that telephone attempts vary in length. We propose to treat attempts that result in an interview as being the longest. Of course, there is a lot of variation among the length of time telephone attempts based on whether there is contact, a refusal, appointment setting, or other results of the attempt. We will ignore this variation and treat all attempts that do not result in an interview as having a cost (smaller than completing an interview).

The SDR takes 18 minutes on the web (median time). CATI likely takes longer. Assume 24 minutes administration time and 6 minutes for reviewing call history, dialing, and introducing the survey. Then we should assume 30 minutes for complete calls and 10 minutes for not complete calls. Assume $35 / hour is the interviewing cost. Ten (10) minutes costs $35/6=$5.83, or 30 minutes costs $17.50; these are the numbers that we used in the simulations. We could also average the rate across all

calls. For example, if 10% of attempts are complete, then assume 30x0.1 + 10x0.9=12 minutes average per attempt regardless of the outcome. 12 minutes costs $35/5=$7 per attempt.

Mail. Next, the mail attempts have a cost that can vary with the size of the mailing. We will ignore this factor and use an average mailing cost. We have counts of the number of mailings. We will assume that these costs are for a package including a paper questionnaire, brochure, introductory letter, etc.

Based on the NSFG experience (paper under review, AAPOR 2021 presentation), the costs of such a package could be $2.72 per mailing for a small job of n=900. A recent paper on the cost of survey mailings from 2017 (Grubert, 2017) reports $1.84 per unit for surveys of 10,000+ with detail about decisions and "bulk savings" included in the description.

Email. Finally, email is the least expensive form of communication. However, for large batch emails, there is still a cost associated with management tasks including developing the email, preparing systems, sending out batches to avoid being labelled as spam. Therefore, we assign a small cost to email. Lacking any information, we assume that an email is about 10% of a mailed package, i.e. $0.18.

With these estimates of costs per attempt by outcome, we need predictions of outcomes at the attempt level in order to estimate costs. We know that "completed interview" and "not completed" outcomes are not equally likely nor are the probabilities of those outcomes consistent over attempts. Therefore, we will predict the probability of each outcome at each attempt using a discrete-time hazard model. We can build a hazard model from the data in the paradata file – but we won't have any time-varying characteristics as these are not reported in the paradata file (it is at the sample-line level). The major predictors will be the number of the attempt and baseline demographic characteristics.

These models can be estimated in two ways: 1) as a Bayesian model with prior information developed from a previous wave, and 2) using the current data only. At the moment, we are focused on the latter approach and using only the 2017 SDR Paradata file.

We can generate cost estimates based on these call-level data. The probabilities of response can be used to generate expected costs as described earlier. The probability of reaching attempt *t* will be the product of 1-Pr(Iw) for all previous attempts.

With these predicted costs, we were able to complete the cost-side of Equation 3. We also need information about the survey variable side of that equation. Since these are not observed for nonrespondents, we will use a prediction generated from regression models used to predict the selected key survey outcome variables. These predictions were made using the same variables from the previous wave, plus a set of demographic variables. Depending upon the outcome variable, either linear or logistic regression were used to make these predictions. Once model coefficients were estimated, a predicted value was calculated for every case in the sample. These predicted values were then used as the survey outcome variables, so that bias of the responding set could be calculated.

This will give us the complete set of inputs required to implement the decision rule. We will simulate the impact of the decision rule on the quality of the estimates and the costs.

Section 3. Data Management and Modeling

In this section, we describe the working environment, data sources, and processing steps/decisions required to process the data used for the modeling steps.

Data Management

The work was completed in the Secure Data Access Facility (SDAF) managed by NORC. This secure enclave gave us access to restricted data that were helpful in the analysis steps. For example, although we began working with the public access SDR data, these data have been coarsened and top-coded, and they have had other limitations implemented as part of a disclosure protection process. Further, the publicly available data are not linkable to the Survey of Earned Doctorates-Doctorate Record File (SED-DRF) file. This linkage is possible in the SDAF. Within the SDAF, we have access to these data resources. We used R and SAS for all of the analysis and simulation steps described in this report.

The data sources used include the following:

- The 2017 SDR. The survey data on the SDAF do not have the disclosure protections implemented for the public release file.
- The 2017 SDR Paradata file. This file was developed by NORC from a variety of sources. The file is a case-level summary of action steps (i.e. phases, number of mailings, number of email attempts, any locating efforts, any special prioritizations, and final status).
- The SED-DRF. The SED is the baseline survey of recent PhD completers. The annual survey is compiled into the DRF, which serves as the sampling frame for the SDR. As such, this file provides useful information for all cases. It is important to note that the SED has changed over the years. As a result, some variables are only available for subsets of the panel.

The 2017 SDR provided several important variables. The two outcome variables that were to be used as key variables in our stopping rule were salary (SALARY) and employment in an educational institution (EMED). After working with the SALARY variable, we found that it was difficult to model the full range of salaries. In particular, salaries above $200,000 were difficult to model. Therefore, we deleted values about $200,000 in order to improve prediction below that level.

The 2017 Paradata file includes a number of useful variables. One difficulty with the case-level file is in reconstructing the sequence of contact attempts. Since we only have the counts of each type of attempt for each case, we cannot exactly reconstruct the sequence of attempts. For example, we assume that mail, email, and CATI attempts could be interleaved across the phases and even within any specific currently assigned mode of interviewing. However, we do not have any way to reconstruct the specific ordering of these attempts. The documentation of the 2017 SDR (from NORC) sheds some light on what was done, but is still not sufficient for us to recreate the exact ordering of attempts for all cases. Further description of how we addressed this issue in the modeling and simulation steps is given below.

The 2017 Paradata file provided the original starting mode assigned to each case. The number of cases assigned to each start mode is described in Table 10. The response propensity models primarily used variables from the 2017 Paradata file with the use of fixed characteristics from either the SED-DRF or the 2015 SDR.

The 2017 SDR includes three important categories of respondents who need to be treated separately for prediction of the survey outcome variables:

1. Panel members who responded in 2015 (n=66,072)
2. Panel members who did not respond in 2015, but participated in previous SDR waves (n=1,165)
3. New panel members who completed the SED in 2013 to 2017 (n=8,744)

We make predictions for the first and third categories. We haven't made predictions for the second category. These could be based on data from earlier SDR waves, the SED-DRF, or some combination of those.

Our strategy was to create a single file that could be used to make predictions for cases in the first and third categories. As a first step, the SED-DRF was linked to the 2017 Paradata file. The SED-DRF contains data from the SED. We used this file to help us fill in the gaps, especially with the third category of SDR panel members. The file was linked to the SDR using the DRF_ID variable. We used six variables recoded from eight variables in the SED-DRF (expected basic annual salary [SALARYV], year completing SED [QUESTYR], post graduate plans [PDOCPLAN], earliest age of functional limitations [DIFAGE]], number of dependents [DEPEND5, DEPEND18, DEPEND19] and father's education [EDFATHER]).

From the SDR 2015 data, we used the following variables as predictors of the two 2017 key variables: SALARY (annualized salary in 2015), EMED (indicator for employer being an education institution in 2015), RACEM (six category race variable), AGE (age of the respondent), JOBSATIS (job satisfaction measured on a four-point scale in 2015), and GENDER (gender of the respondent). These were available for existing panel members, i.e. those who participated in the SDR starting in 2015 or earlier.

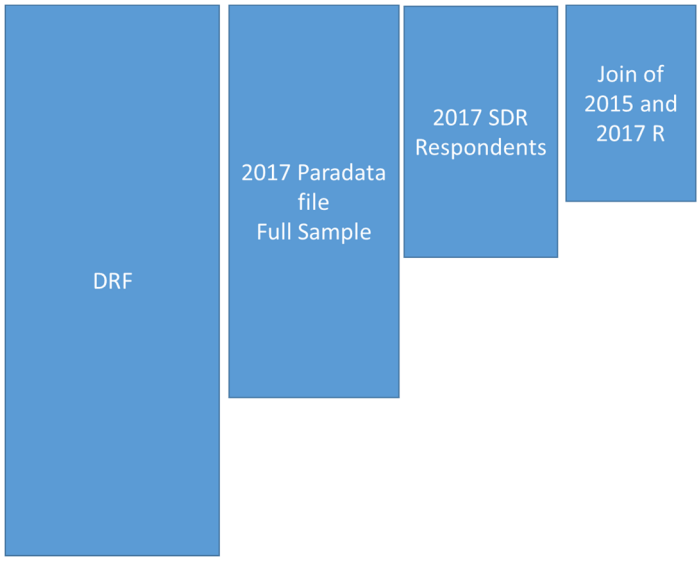
Variables used for the models for the new panel members from the SED-DRF datafile:

1. Expected basic annual salary: A categorical variable was created using quintiles for expected basic annual salary for values $200,000 and lower. Missing items due to logical skips (respondents who did not have post graduation plans) were retained by creating a separate category, "No plans".
2. Year from SED: The difference between 2017 and the year they took SED. Though the new cohort is mostly sampled from SED 2014 and SED 2015, there were a substantial number that were sampled from SED 2013 and a few from SED 2012, SED 2016 and SED 2017.
3. Post graduate plans: This variable was recoded from PDOCPLAN. "Postdoc fellowship" and "postdoc research associate" were collapsed into one category, "postdoc". "Traineeship", "internship", "clinical residency", "other training" and "unspecified other training or study" were collapsed into one category, "further training and education". "Employment" and "unspecified employment" were collapsed into one category, "employment". Military service was kept as its own category. Missing values due to logical skips or respondents skipping the question was recoded as "No plans".
4. Any difficulties: This variable was recoded from DIFAGE. Respondents who skipped the question due to logical skip because they did not report any functional difficulties were coded as "no difficulties", while respondents who answered this question was coded as "yes, have difficulties".
5. Any dependents: This variable was created using DEPEND5, DEPEND18, DEPEND19. If the number of dependents is more than 0, then they are coded as having dependents.

6. Father's education: The variable EDFATHER was mostly kept the same. The only change made was to collapse "Bachelor's degree" and "Associate degree" into one category, and "Master's degree" and "professional degree" into one category. Missing items were coded as "Don't know".

Our strategy was to start from the 124,580 cases in the 2017 SDR Paradata file, which was interpreted as representing the full 2017 SDR sample. We linked the 2017 SDR to this file. Then, we supplemented that file with the variables from the SED-DRF. This file has a complex pattern of missing data depending mainly upon which of the three categories each panel member came from, but also upon other factors such as when the person took the SED, item-level missing data, etc. Figure 1 provides a schematic display of these relationships, but the relative sizes of the files (i.e number of records) are not "drawn to scale."

Figure 1. Data Structure



Modeling Steps

We built two types of models:

1. Predictions of survey response values, and
2. Predicted probabilities of response.

In this section, we describe the steps in selecting models and creating predictions from the final models.

Survey Outcome Variables

We started with the salary variable from the 2017 SDR. The model predicting the responses to the 2017 survey from the 2015 data was the initial model that included data from the largest category of panel members. The predictors included several variables (See Table 2). Standard regression modeling variable selection strategies were used to select the final set of predictors. As noted earlier, we did delete 2017 reported salaries that were greater than $200,000. We considered several transformations of salary and found that the untransformed version, with the extreme values deleted, provided the best fit. This

approach seemed appropriate given our desire to create predictions and not to explain differences in salary. A key predictor in this model is the 2015 SDR salary.

Table 2. Final Estimated Coefficients Predicting 2017 SDR Salary using 2015 SDR data

Predictor	Est. Coef	SE	P-value
SDR 2015 salary	.87	< 0.01	<.01
Age	-284.55	10.45	<.01
Employer in 2015 is an education institution (ref: No)	-3629.37	238.21	<.01
Male	11169.78	235.05	<.01
Job satisfaction in 2015 (ref: Very satisfied)			
Somewhat satisfied	-568.15	242.02	.02
Somewhat dissatisfied	574.12	442.01	.20
Very dissatisfied	-241.54	851.53	.78
R^2	.68		

A separate model was estimated for the new panel members joining the SDR for the first time in 2017. These cases had predictions of salary based upon data from the SED-DRF. The variable of most value was the self-reported "predicted" salary, i.e. the salary they expected to have once they started their first job after completing the PhD. Table 3 shows the estimated coefficients from this model.

Table 3. Final Estimated Coefficients Predicting 2017 SDR Salary using 2015-2016 SED data and other predictors

Predictor	Est Coef	SE	Pval
Expected salary (ref: Between $0 to $42000)			
$42000 to $52999	7006.58	1369.00	< .01
$53000 to $79999	17158.24	1469.58	< .01
Above $80000	49960.58	1667.76	< .01
Don't know/Missing	18348.16	1814.50	< .01
Years since SED	3589.17	514.10	< .01
Age (centered)	-243.92	58.75	< .01
Male	-4633.83	759.67	< .01
Race (ref: Hispanic)			
Non-Hisp Asian	7224.03	1366.48	< .01
Non-Hisp Black	99.13	1777.53	.96
Non-Hisp Native	-4997.69	4194.35	.23
Non-Hisp Pacific	4263.53	6722.53	.53
Non-Hisp White	3367.72	1257.03	.01
Not U.S. citizen	-7436.33	953.67	< .01
Primary field of study (ref: Biological, agricultural, environmental life science)			
Computer, information sciences	28826.24	1826.75	< .01
Mathematics, statistics	9046.29	1979.51	< .01
Physical sciences	4821.29	1158.14	< .01

Psychology	3899.15	1391.30	.01
Social sciences	611.99	1246.13	.62
Engineering	14441.13	1129.97	< .01
Health	9383.71	1627.82	< .01
Post-graduate plans (ref: Postdoc)			
Military	-624.97	5758.32	.91
Further training, education	8672.87	3896.98	.03
Employment	1815.84	1150.27	.11
No plans	-5735.99	1702.56	< .01
Any disability	-2881.26	1428.29	.04
R^2	.28		

The next survey outcome variable we predicted was employment by an educational institution (EMED). The modeling strategy was similar to that of the salary variable – we predicted the 2017 SDR response using the 2015 SDR variable and other survey items for those 2017 SDR respondents who responded at both time points. Table 4 presents the estimated coefficients for this logistic regression model.

Table 4. Final Estimated Coefficients Predicting 2017 SDR Employment by an Educational Institution using 2015 SDR data and other predictors

Predictor	Est Coef	SE	Pval
2015 salary	-2.14×10^{-6}	4.47×10^{-7}	< .01
Age	.01	1.74×10^{-3}	< .01
Employer in 2015 is an education institution (ref: No)	5.61	.04	< .01
Race (ref: White only)			
Asian only	-.17	.05	< .01
AIAN only	.46	.28	.10
Black only	-.03	.08	.67
NHPI only	-.38	.42	.37
Multiple races	-.05	.12	.68
$Pseudo\ R^2$	.69		

For the new panel members, we used several variables from the SED as predictors, including a variable describing future plans. The estimated coefficients are presented in Table 5.

Table 5. Final Estimated Coefficients Predicting 2017 SDR Employment by an Educational Institution using 2015-2016 SED data and other predictors

Predictor	Est Coef	SE	Pval
Expected salary (ref: Between $0 to $42000)			
$42000 to $52999	-.12	.10	.22
$53000 to $79999	-.45	.10	< .01
Above $80000	-1.57	.12	< .01
Don't know/Missing	-.72	.13	< .01

Age (Centered)	.01	4.43×10^{-3}	.01
Race (ref: Hispanic)			
Non-Hisp Asian	-.36	.10	< .01
Non-Hisp Black	-.10	.13	.44
Non-Hisp Native	-.46	.30	.12
Non-Hisp Pacific	-.40	.47	.39
Non-Hisp White	-.08	.09	.35
Not U.S. citizen	.20	.07	< .01
Primary field of study (ref: Biological, agricultural, environmental life science)			
Computer, information sciences	-.22	.13	< .01
Mathematics, statistics	.64	.14	.10
Physical sciences	-.06	.08	< .01
Psychology	-.08	.10	.46
Social sciences	.79	.09	.38
Engineering	-.57	.08	< .01
Health	.49	.12	< .01
Post-graduate plans (ref: Postdoc)			
Military	-1.34	.51	.01
Further training, education	-.29	.27	.29
Employment	-.41	.08	< .01
No plans	-.26	.12	.03
Any dependents	.17	.06	.01
Father's education (ref: Less than highschool)			
Highschool	-.04	.12	.75
Some college	-.03	.12	.83
Bachelor's degree/Associate's degree	-.28	.11	.01
Masters or professional degree	-.29	.11	.01
Phd	-.06	.13	.64
Don't know/missing	-.23	.22	.30
$Pseudo\ R^2$	.09		

Response Propensity Models

In the 2017 SDR, there were three "starting" modes assigned – web, mail, and CATI. These are indicated for each case on the 2017 Paradata file. For each of these three groups (defined by starting mode), we created an attempt-level file. The creation of an attempt-level file involved a series of assumptions outlined in the data management section. This file was used to estimate discrete time hazard models for each "starting" mode (i.e. sequence of modes). The predictors for this model were drawn from the 2017 Paradata file (including results from 2015 SDR) and other sources. Only records that conform to the mode switch sequence from the SDR 2017 Methods Report were used for these analyses. The results are reported in Table 6.

The variables used are as follows:

- Starting mode: This variable is used to determine the mode sequence to follow. Cases that were assigned as "Congressional Refusals", "Hostile", "Reluctant", "Deceased" or "Out of scope" were not included in the analyses.
- No. of contacts: This variable was constructed from "ST_MAILING_COUNT17", "ST_EMAILS17", "ST_DIALSESSIONS17", "IN_MAILING_COUNT17", "IN_EMAILS17", "IN_DIALSESSIONS17", "LS_MAILING_COUNT17", "LS_EMAILS17", and "LS_DIALSESSIONS17". These variables are a count of contact attempts made in each mode in the start, interim and last phase. This variable was constructed differently for each start mode, following the mode sequence that was in the SDR2017 Methods Report. For example, no. of contacts for the web start mode was the sum of "ST_EMAILS17", "IN_MAILING_COUNT17" and "LS_DIALSESSIONS17".
- Mode preference: Respondents who skipped that question were recoded as "Don't know" to retain the cases for modelling
- Needed locating in 2017: This is EVER_LOC17, which is a flag that indicates whether the respondent ever needed locating in 2017
- Responded in SDR2015: This was constructed from LASTRESP17, a variable that captures the respondent's most recent response before SDR2017. Responses were collapsed into three categories which are "Responded to SDR2015", "Did not respond to SDR2015", "New cohort".
- Primary field of study: This is PAD17_14, Primary Analysis Domain in the paradata, used as is with no recoding
- Besides the above, demographic variables were also used (AGE17, SEX17, RACETH17, CURCIT17)

Table 6 presents the estimated coefficients for cases that were assigned to the web mode as the starting mode. The order of the modes was Web-Mail-CATI. The model for the second mode is estimated using only those cases that were not completed during the first mode. The model for the third mode is estimated using only those that were not completed during the first or second mode. In each model, the number of the attempt is an important predictor. The coefficients for the number of the attempt are positive for web and mail -- indicating that later attempts are more likely to produce interviews than early attempts -- and negative for the third mode.

Table 6. Final Estimated Coefficients Predicting Response to the 2017 SDR for the Web Starting Mode

	1st mode: Web			2nd mode: Mail			3rd mode: CATI		
Predictor	Coef.	SE	P-val	Coef.	SE	P-val	Coef.	SE	P-val
No. of contact attempt	.19	.01	<.01	.35	.02	<.01	-.13	.01	<.01
Mode preference (ref: CATI)									
Web	1.02	.11	<.01	.38	.09	<.01	-.13	.13	.32
Mail	.22	.11	.04	2.42×10^{-3}	.10	.98	-.39	.14	<.01
None	.69	.11	<.01	.14	.10	.15	-.14	.14	.32
Don't know	-1.64	.11	<.01	-2.15	.10	<.01	-2.75	.14	<.01
Needed locating in 2017	-1.09	.03	<.01	-1.08	.03	<.01	-.52	.04	<.01

Primary field of study (ref:Bio)									
CIS	.04	.04	.27	-0.7	.05	.19	-.19	.10	.06
MS	.12	.03	<.01	.04	.04	.38	.03	.09	.69
Phys	.08	.02	<.01	.01	.03	.73	.05	.05	.30
Psychology	-.06	.02	<.01	-.01	.03	.61	-.11	.06	.06
Social sciences	.01	.02	.68	8.61x 10^{-3}	.03	.81	-08	.06	.13
Engineering	-.01	.02	76	-.01	.03	.59	3.89 x 10^{-3}	.05	.94
Health	-.11	.03	<.01	.05	.04	.27	.12	.08	.13
Responded in SDR 2015 (ref: Responded)									
Did not respond	.36	.04	<.01	.14	.05	<.01	-.31	.05	<.01
New cohort	.56	.02	<.01	.75	.03	<.01	-.02	.08	.79
Age (centered)	.01	5.36 x 10^{-4}	<.01	8.09 x 10^{-4}	7.96 x 10^{-4}	.31	1.08 x 10^{-3}	1.53 x 10^{-3}	.49
Female	-.01	.01	.52	.04	.02	.03	-.03	.03	.46
Race (ref: Hispanic)									
Non-Hisp Asian	-.16	.03	<.01	-.02	.04	.49	.04	.07	.57
Non-Hisp Black	-.26	.04	<.01	-.13	.05	.01	.09	.08	.27
Non-Hisp AIAN	.13	.07	.06	-.05	.10	.63	-.08	.94	.65
Non-Hisp NHPI	.02	.13	.89	.22	.18	.22	-.14	.14	.75
Non-Hisp White	.12	.02	<.01	.07	.03	.03	.05	.06	.41
Not U.S. citizen	-.08	.02	<.01	-.17	.02	<.01	.14	.04	<.01
$Pseudo\ R^2$	.18			.29			.24		

*CIS: Computer and information sciences; MS: Mathematics and statistics; Phy: Physical sciences; Bio: Biological, agricultural, environmental life science; AIAN: American Indian/Alaska Native; NHPI: Native Hawaiian/Other Pacific Islander

Table 7 presents the estimated coefficients for the cases assigned to start with the mail survey mode.

Table 7. Final Estimated Coefficients Predicting Response to the 2017 SDR for the Mail Starting Mode

	1st mode: Mail			2nd mode: Web			3rd mode: CATI		
Predictor	Coef.	SE	P-val	Coef.	SE	P-val	Coef.	SE	P-val
No. of contact attempt	1.04	.04	<.01	-.12	.05	.02	-.07	.02	.01
Mode preference (ref: CATI)									
Web	1.27	.36	<.01	.09	.13	.50	.26	.23	.25
Mail	2.23	.36	<.01	-.01	.14	.94	-.23	.24	.33
None	1.53	.37	<.01	-.01	.15	.97	-.18	.27	.49
Don't know	-.09	.37	.08	-2.66	.16	<.01			

Needed locating in 2017	-1.73	.19	<.01	-2.39	.23	<.01	-.65	.13	<.01
Primary field of study (ref:Bio)									
CIS	.16	.20	.41	-.48	.17	.01	-.47	.26	.07
MS	.21	.13	.09	-.10	.12	.42	-.36	.21	.09
Phys	.11	.08	.14	-.06	.07	.39	-.15	.12	.21
Psychology	.13	.09	.13	-.06	.09	.49	.02	.14	.91
Social sciences	-.12	.09	.19	-.06	.08	.48	-.14	.14	.29
Engineering	-.03	.09	.72	-.05	.07	.52	-.28	.13	.03
Health	-.09	.15	.56	-.41	.15	<.01	-.07	.18	.70
Responded in SDR 2015 (ref: Responded)									
Did not respond	1.20	.17	<.01	.87	.33	.01	-.23	.34	.49
New cohort	2.59	.21	<.01	3.50	.38	<.01	.81	.71	.26
Age (centered)	.02	2.43×10^{-3}	<.01	4.56×10^{-3}	2.18×10^{-3}	.04	.01	3.76×10^{-3}	<.01
Race (ref: Hispanic)									
Non-Hisp Asian	.06	.13	.64	-.10	.09	.25	-.20	.16	.22
Non-Hisp Black	-.46	.18	.01	-.24	.11	.03	-.55	.20	<.01
Non-Hisp AIAN	-.09	.36	.81	-.43	.33	.18	-.01	.40	.98
Non-Hisp NHPI	-12.12	203.33	.95	.48	.41	.24	-.36	1.00	.78
Non-Hisp White	.31	.12	.01	-.05	.09	.52	-.23	.15	.13
Not U.S. citizen	-1.17	.11	<.01	.04	.06	.55	.10	.10	.36
$Pseudo\ R^2$	.25			.31			.20		

Table 8 presents the estimated coefficients for the cases assigned to start with the CATI mode.

Table 8. Final Estimated Coefficients Predicting Response to the 2017 SDR for the CATI Starting Mode

	1st mode: CATI			2nd mode: Web			3rd mode: Mail		
Predictor	Coef.	SE	P-val	Coef.	SE	P-val	Coef.	SE	P-val
*No. of contact attempt	-.10	.04	.02	-.07	.15	.63	-	-	-
Mode preference (ref: CATI)									
Web	-.63	.16	<.01	.12	.24	.61	.49	.53	.35

Mail	.58	.24	.01	-.63	.32	.05	-.04	.61	.95
None	.14	.20	.50	-.05	.33	.89	.11	.67	.87
Don't know	-3.08	.51	<.01	-2.51	.32	<.01	-1.91	.50	<.01
Needed locating in 2017	-2.82	.51	<.01	-1.14	.21	<.01	-.67	.35	.06
Responded in SDR 2015 (ref: Responded)									
Did not respond	-.14	.37	.70	.25	.32	.45	-.24	.75	.75
New cohort	1.67	.65	.01	.72	.30	.02	-.50	.55	.36
Age (centered)	.01	5.9×10^{-3}	.01	1.27×10^{-3}	6.85×10^{-3}	.85	8.16×10^{-6}	1.28×10^{-2}	.99
Female	-.04	.14	.76	-.08	.16	.62	-.73	.34	.03
$Pseudo\ R^2$	.22			.32			.26		

*Only two respondents received more than one mailing attempt, therefore no. of attempts was not included in the model for mail mode

Predictions from these models were attached to the 124,580 cases in the paradata file. In the end, we had about 70,000 cases with predicted values for the two survey variables and one of the three response propensity models. The remaining cases did not have predicted values. The major reasons for not having predicted values include not completing the 2015 SDR, being a new panel member with no values for predicted positions, and the Paradata file variables revealing that the case did follow the prescribed mode sequence. The major reasons and the counts of cases are presented in Table 9.

Table 9. Most Frequent Explanations for Inability to Create Predicted Values

Reason	Count
Did not respond to SDR2015	10,923
Did not follow the mode sequence (Web start mode)	38,671
Did not follow the mode sequence (Mail start mode)	3,491
Did not follow the mode sequence (CATI start mode)	1,388

Section 4. Simulation Design and Results

Mode Switch

The simulation is based upon the 2017 Paradata file and the description of the design in the "2017 Survey of Doctorate Recipients: Methodology Report" (referenced as "Methodology Report" from here). We note that the report summarizes the design, but that individual cases may have experienced departures from that overall design. Some of these could be identified in the 2017 Paradata file while others could not be identified.

In order to create an attempt-level file for the propensity models, we first divided all cases by "start" mode – web, mail, and CATI. Working from the Methodology Report and the 2017 Paradata file, we constructed a design specification that most – but not all – cases received. We also truncated the tails of the effort. That is, some cases received more effort than indicated by the protocol in the table. We truncated effort to levels received by a large majority of cases (80-90%). The resulting specification of the design is presented in Table 10. The first row lists the total sample size available in the paradata file. The second row is the subset of cases for which we have predicted survey variables. These are the cases that are included in the simulation studies. For each start mode, the number of attempts in each mode in the sequence are described. For example, the "web start mode" cases start with two email attempts. After these attempts, cases are switched to mail. There are three mailed requests to complete the survey. Then, finally, cases are switched to CATI and receive 6 attempts.

Table 10. Study Design by Mode Sequence ("Start Mode"): Recruitment Effort by Mode within Sequence

	Web start mode	**Mail start mode**	**CATI start mode**
Full Sample	(n=66,572)	(n=5,602)	(n=660)
Sample with Predicted Values	(n=42,858)	(n=3,598)	(n=366)
First Mode | Attempts	Email | 2	Mail | 3	CATI | 6
Second Mode | Attempts	Mail | 3	Email | 2	Email | 2
Third Mode | Attempts	CATI | 6	CATI | 6	Mail | 3

*Note: Cases were assigned to each of the mode sequences based on respondent behavior and expressed preferences as described in the SDR2017 methods report -- they were **not** randomized to these sequences.

Starting from this overall design, we created an "attempt-level" file. Each record includes the start mode, the mode of contact, and the number of the attempt (i.e. first attempt, second attempt, etc.). We also included the cost of the attempt on each record using the cost information described in a previous section. We also included a set of predictors from the Paradata file and the SED-DRF. We tried to include paradata indicators as our purpose is to separate "easy" and "difficult" cases. Finally, we created a binary flag indicating whether the case was completed on that attempt or not. From this file, we estimated a discrete-time hazard model. The results are presented in the previous section.

We note that this model would underestimate the final response rate as we placed prior limits on effort that were not observed in practice. For example, some cases may have received 8 CATI attempts. Our simulation structure excludes those attempts and their results. We simulated the truncated effort by removing all attempts after the limits presented in Table 10. All interviews that occurred on attempts

after these limits were excluded from the calculation of key statistics. All attempts after these limits were excluded from cost estimates. The results from this analysis serve as a "control" or comparison in the simulation study. From this file, we can calculate the final response rate, total costs, and the mean and variance of the survey outcome variables of interest. Table 11 show the results from this control simulation using the truncated effort described in Table 10. Full results, including standard errors and percentiles of the simulation results are presented in Appendix 1.

Table 11. Truncated Original Design Results, "Control"

	Final Response Rate	**Completed Interviews**	**Total Cost**	**Mean Salary**	**Proportion Employed by Educational Institution**
Combined	90.5%	42,354.3	$293,234.70	$95,625.72	0.523
Web Start Mode	91.5%	39,201.0	$224,616.40	$95,591.11	0.527
Mail Start Mode	79.2%	2,849.1	$58,792.73	$96,167.86	0.473
CATI Start Mode	83.2%	304.1	$9,825.63	$95,008.15	0.517

Our first simulation study implemented a simple rule -- if the expected cost of an interview (attempt cost / estimated response probability) is higher on the current attempt than the next attempt, then skip the current attempt. Here, the attempt costs are $0.18 for an email, $1.84 for mail, and $17.50 for CATI (i.e. the cost of conducting an interview via the telephone). The simulation involved stepping through – for each case – the ordered attempts. First, the expected cost for completion of each attempt is compared to the expected cost of completion for the next attempt on the ordered list. If the next attempt is less expensive, then the current attempt is deleted (i.e. not made). Then, for each attempt that is made, a random UNIFORM(0,1) draw will be compared to the estimated probability of response to that attempt. If the draw is less than the estimate, the case is then coded as "complete." After a case is "complete," all the attempts after the attempt on which the completion occurs are deleted. Once all attempts had been treated, the final response rate, costs, and survey estimates were tabulated and stored. This process was repeated 1,000 times. The results are not presented as this rule led to increases in costs without any reduction in errors. This is because in the web start mode (the largest group) only one email (the second email) is sent, only one letter is sent (the third letter) and all CATI attempts are made. In the mail start mode, the web attempts are completely dropped. Therefore, this rule produces an unintended consequence.

Our second simulation study implemented a more complex rule -- if the expected cost of an interview is higher in the current mode (when considering all planned attempts) than in the next mode, then skip the current mode. This required calculating for each case the expected probability of being interviewed in a mode. The cost of an interview was then estimated using the cost of the expected number of

attempts in each mode and the costs of each of those attempts. A simulation of using this rule was conducted and the results are reported in Table 12.

Table 12. Simulated Mode Switch Rule Results

	Final Response Rate	**Completed Interviews**	**Cost (% Savings)**	**Mean Salary (Bias)**	**Proportion Employed by Educational Institution (Bias)**
Combined	90.3%	42,262.2	$273,056.50 (6.9%)	95,614.08 ($11.64)	0.523 (0.000)
Web Start Mode	91.5%	39,201.8	$225,664.00 (-0.5%)	95,593.60 (-$2.49)	0.527 (0.000)
Mail Start Mode	76.7%	2,758.9	$47,854.00 (18.6%)	95,970.83 (197.03)	0.474 (-0.001)
CATI Start Mode	82.5%	301.5	$538.48 (94.5%)	95,018.03 (-$9.88)	0.517 (0.000)

In general, this rule produced savings mainly by eliminating CATI attempts from the CATI start mode cases and the mail attempts from the mail start mode cases. Most of the savings occurred in the CATI start mode. There were no savings in the web start mode as no attempts were dropped.

Stopping Rule

For the stopping rule proposed in Section 2, we use the predicted survey variables, predicted response probabilities, and information about costs. For this rule, predictions of the survey variable values are critical. We used predictions for two variables: salary and employment at an educational institution. As described earlier, we have these predictions for most, but not all, of the cases.

The next step was to simulate the impact of the stopping rule. We used the same data structure employed for the mode switching rule simulation. This allowed us to predict future costs for the case as a combination of probability of completion and the cost per attempt to create a predicted costs at the case level. These costs will be used as an input to the rule.

In the simulation, we looked at the impact of implementing the rule prior to several different attempts. An important question to answer regarding the use of the rule is when and how often to implement it. Implementing the rule earlier allows for saving additional costs by stopping cases earlier. Implementing the rule later allows more information to be collected that may improve the efficiency of decisions about stopping. We tried several alternatives for when the rule was first implemented. We did this separately for each starting mode (web, mail, and CATI). For example, in the “Starting Mode: Web” Table, the row labelled “3 (mail)” is the result of a simulation where the stopping rule was implemented

prior to the 3rd attempt. The results in that row show a benchmark where no cases are stopped and the results after the stopping rule is implemented but the data collection is finalized.

We stopped trying implementations of the stopping rule at later attempts once only small changes were detected. For example, for the web starting mode, implementing the stopping rule at the 5th attempt (i.e. the third mail attempt) had very little impact on estimates or costs.

Table 13. Stopping Rule Simulation Results

Table 13a. Starting Mode: Web

	Benchmark Estimates (Truth)				**Estimates After Stopping Rule**			
Attempt Number	Resp. Sample size	Cost	Mean Salary	Prop. Higher	Resp. Sample size	Cost (% Savings)	Mean Salary (Bias)	Prop. Higher (Bias)
3 (mail)	39,227	$214,001.01	$95,569.11	0.527	37,224	$44,044.00 (79.4%)	$95,623.57 ($54.46)	0.528 (0.001)
4 (mail)	39,192	$220,829.32	$95,632.19	0.527	38,524	$103,567.99 (53.1%)	$95,623.78 (-$8.42)	0.528 (0.001)
5 (mail)	39,211	$221,047.04	$95,656.24	0.527	39,211	$221,010.22 (0.02%)	$95,656.24 ($0.00)	0.527 (0.000)
6 (cati)	39,239	$216,546.54	$95,638.02	0.527	39,239	$216,511.56 (0.02%)	$95,638.02 ($0.00)	0.527 (0.000)

Table 13b. Starting Mode: Mail

	Benchmark Estimates (Truth)				Estimates After Stopping Rule			
Attempt Number	Resp. Sample size	Cost	Mean Salary	Prop. Higher	Resp. Sample size	Cost (% Savings)	Mean Salary (Bias)	Prop. Higher (Bias)

3 (mail)	2,836	$58,830.51	$95,680.66	0.476	2,611	$29,729.76 (49.47%)	$95,917.93 ($237.27)	0.478 (0.002)
4 (web)	2,867	$57,900.60	$95,999.09	0.477	2,859	$56,085.80 (3.13%)	$96,001.83 ($2.74)	0.477 (0.000)
5 (web)	2,839	$60,472.93	$95,996.24	0.471	2,839	$60,437.77 (0.06%)	$95,996.24 ($0.00)	0.471 (0.000)

Table 13c. Starting Mode: CATI

	Benchmark Estimates (Truth)				Estimates After Stopping Rule			
Attempt Number	Resp. Sample size	Cost	Mean Salary	Prop. Higher	Resp. Sample size	Cost (% Savings)	Mean Salary (Bias)	Prop. Higher (Bias)
3 (cati)	302	$9,341.44	$94,746.39	0.518	301	$9,306.45 (0.37%)	$94,733.73 (-$12.65)	0.519 (0.001)
4 (cati)	302	$9,795.47	$95,321.31	0.521	301	$9,772.14 (0.24%)	$95,310.57 (-$10.74)	0.522 (0.001)
7 (mail)	301	$10,175.60	$95,478.75	0.517	301	$10,167.88 (0.08%)	$95,478.75	0.517

In general, we note that the stopping rule allows for relatively large savings (53-79% in web start mode, 49% in mail start mode) with relatively small biases (less than 1%). These cost savings result from stopping relatively small numbers of cases, as evidenced by the relatively small reductions in the number of respondents. In the web start group, there were 2,003 fewer interviews in the simulation of stopping at the 3rd attempt, where savings were up to 79%. The simulation of stopping at the 4th attempt saved 53% on costs and collected 668 fewer interviews.

Section 5. Future Directions

In this section, we propose additional simulation work that can be done, as well as some possible experimentation for future waves of the SDR.

Now that we have built our simulation structure, we can work to improve it. We would like to extend our predictions so that all cases have predicted survey variables. This may require that we build additional models. For example, we may need to predict the survey variables for cases who were not interviewed in the previous wave and who are not new from the SED. Another improvement would be to use machine learning techniques to improve predictions. These techniques can be used to improve both predictions of key survey variables and response probabilities. Examples of techniques we can implement include Bayesian Additive Regression Trees (BART). We have used these models to predict costs in the HRS and the NSFG. Also, we can use Least Absolute Shrinkage and Selection Operator (LASSO) as a technique to select a model and control overfitting in the prediction of response propensities.

Further, we are currently developing multivariate stopping rules. We would use simulation to test these on the SDR using the simulation structure we have already created. The multivariate rule will base decisions upon several variables and seek to optimize outcomes across these.

In addition to these simulation steps, we would like to propose several experiments that would help elaborate adaptive designs for the SDR. There are three experiments that we are proposing. In each case, we could work closely with NCSES and survey vendors to provide detailed guidance on the design, including sample sizes. The experiments are:

Randomize the starting mode. The starting mode in 2017 was assigned based upon expressed preferences from panel members about "preferred" modes of completion, or based upon the actual mode of completion. Without an experimental comparison we can't determine whether the mode sequence assigned was more effective than an alternative sequence. In general, it appears that assigning cheaper modes first is more cost effective. But without randomization of assignment of sequences across cases, we can't make that claim with evidence. This might have been done prior to SDR 2017 or in either 2019 or 2020. If so, the results of such an experiment might be used to identify cases for which an alternate sequence might work better (e.g. persons who will not do a web survey but will complete on paper). An experiment with those subgroups might be useful to confirm the findings.

Experiment with a mode switching rule. Our analysis suggests that ordering the modes from least expensive to most expensive is the best option. We did not identify places where mode switching should occur earlier. However, this has been an effective technique for the National Survey of College Graduates (Coffey, et al., 2021). It merits further exploration for the SDR. First, it would be helpful to incorporate real cost data (possibly updating cost estimates in real time, but certainly update probability estimates in real time). These estimates might differ from the ones used here such that mode switching becomes more attractive.

It is also possible that the estimated probabilities of response for each attempt would change given new data. These are directly related to costs. Changes in probabilities across the modes could lead to identifying cases that should switch earlier.

Experiment with the stopping rule. We need to show some simulation results to justify this choice. An experiment would allow us to evaluate whether stopping cases actually leads to a "re-allocation" of effort to other cases judged to be more important to estimates. We have found this to be the case in experimentation on the Health and Retirement Study. We would expect similar results in this setting, particularly with the use of CATI.

References

Coffey, S., B. Reist and P. V. Miller (2019). "Interventions on-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates." Journal of Survey Statistics and Methodology.

Grubert, E. (2017). "How to Do Mail Surveys in the Digital Age: A Practical Guide." Survey Practice 10(1): 1-10.

Appendix 1. Simulated Control Results Including 2.5 and 97.5 Percentiles

		Mean	2.5 Percentile	97.5 Percentile
Combined	Final Response Rate	90.5%	90.2%	90.6%
	Completed Interviews	42,354.3	42,277.0	42,434.0
	Total Cost	$293,234.70	$289,920.30	$296,873.80
	Salary	$95,625.72 (186.9)	$95,556.14 (186.5)	$95697.36 (187.1)
	Employed by Educational Institution	0.523 (0.002)	0.523 (0.002)	0.524 (0.002)
Web Start Mode	Final Response Rate	91.5%	91.3%	91.6%
	Completed Interviews	39,201.0	39,132.0	39,271.0
	Total Cost	224,616.40	221,816.70	227,757.70
	Salary	95,591.11 (192.94)	95,525.30 (192.63)	95,658.04 (193.24)
	Employed by Educational Institution	0.527 (0.002)	0.526 (0.002)	0.528 (0.002)
Mail Start Mode	Final Response Rate	79.2%	78.3%	80.0%
	Completed Interviews	2,849.1	2,818.0	2,879.0
	Total Cost	58,792.73	57,512.72	60,026.05
	Salary	96,167.86 (776.64)	95.733.29 (769.93)	96,613.21 (783.32)
	Employed by Educational Institution	0.473 (0.008)	0.468 (0.008)	0.478 (0.008)
CATI Start Mode	Final Response Rate	83.2%	81.3%	85.1%
	Completed Interviews	304.1	297.0	311.0
	Total Cost	9,825.63	9,589.61	10,092.50
	Salary	95,008.15 (2,399.53)	94,066.96 (2,349.15)	95,968.22 (2,450.24)
	Employed by Educational Institution	0.517 (0.025)	0.507 (0.025)	0.528 (0.025)

Appendix 2. Simulated Next Mode Stopping Rule Results Including 2.5 and 97.5 Percentiles

		Mean	2.5 Percentile	97.5 Percentile
Combined	Final Response Rate	90.3%	90.1%	90.4%
	Completed Interviews	42,262.2	42,189.0	42,341.0
	Total Cost	273,056.50	269,846.90	276,281.90
	Salary	95,614.08 (187.0)	95,542.90 (186.7)	95,681.11 (187.3)
	Employed by Educational Institution	0.523 (0.002)	0.523 (0.002	0.524 (0.002)
Web Start Mode	Final Response Rate	91.5%	91.5%	91.6%
	Completed Interviews	39,201.8	39,130.0	39,269.0
	Total Cost	225,664.00	221,810.80	227.439.30
	Salary	95,593.60 (192.9)	95,530.57 (192.7)	95,655.79 (193.2)
	Employed by Educational Institution	0.527 (0.002)	0.526 (0.002)	0.528 (0.002)
Mail Start Mode	Final Response Rate	76.7%	75.8%	77.6%
	Completed Interviews	2,758.9	2,727.0	2,792.0
	Total Cost	47,854.00	46,432.59	49,259.13
	Salary	95,970.83 (788.81)	95,458.52 (781.73)	96,456.09 (796.53)
	Employed by Educational Institution	0.474 (0.008)	0.468 (0.008)	0.479 (0.008)
CATI Start Mode	Final Response Rate	82.5%	80.4%	84.6%
	Completed Interviews	301.5	294.0	309.0
	Total Cost	538.48	481.99	599.36
	Salary	95,018.03 (2,409.9)	94,002.67 (2,359.31)	96,010.52 (2,461.53)
	Employed by Educational Institution	0.517 (0.025)	0.506 (0.025)	0.528 (0.026)