Supporting Statement for Promoting Opportunity Demonstration (POD)
OMB No. 0960-0809

B. COLLECTIONS OF INFORMATION EMPLOYING STATISTICAL METHODS

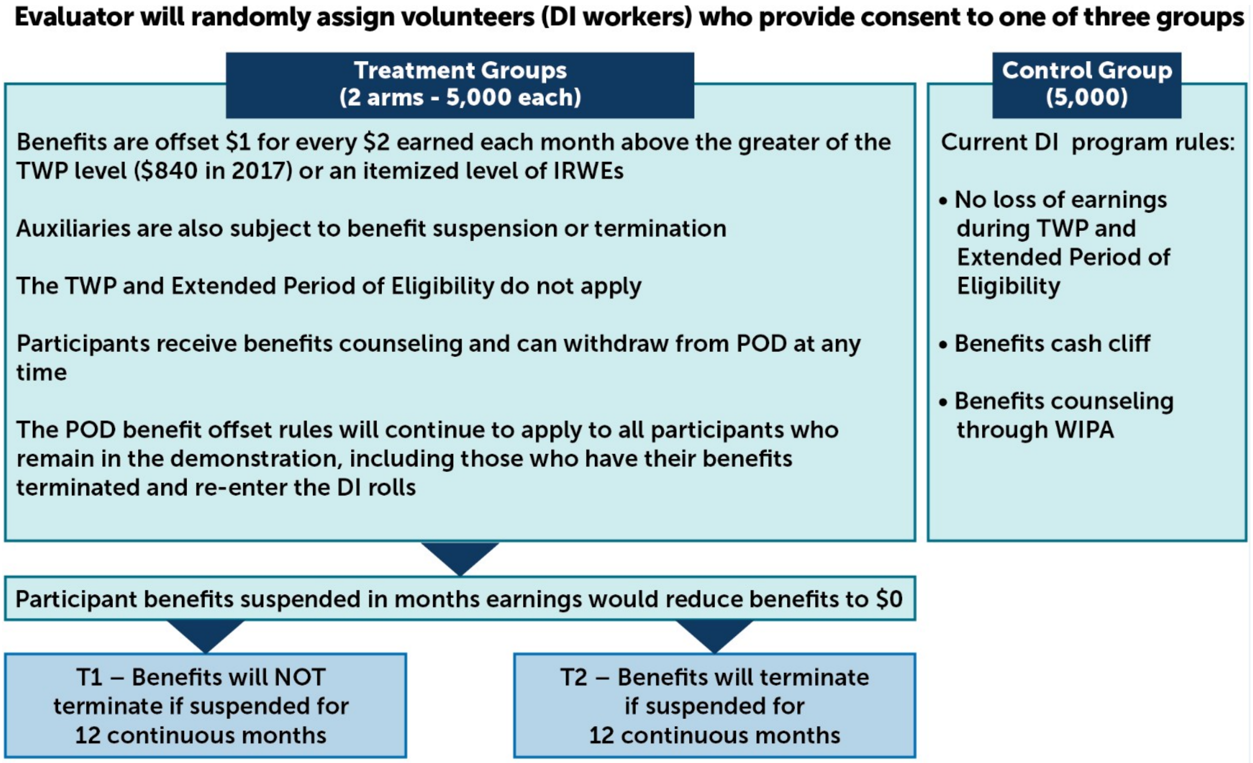
As part of the Bipartisan Budget Act of 2015 (BBA), policymakers required the Social Security Administration (SSA) to carry out the Promoting Opportunity Demonstration (POD) to test a new benefit offset formula for Social Security Disability Insurance (SSDI) beneficiaries who volunteer to be in the demonstration. We intend the new rules, which also simplify work incentives, to promote employment and reduce dependency on benefits. POD is part of a broader effort by policymakers to identify new approaches to help beneficiaries and their families, many of whom are low income, increase their incomes and self-sufficiency through work. A thorough and comprehensive evaluation of POD is, therefore, critical to enable SSA and policymakers to assess the promise of this option for improving beneficiary outcomes and program administration for the context in which it is implemented. The evaluation will provide empirical evidence on the impact of the intervention for the SSDI beneficiaries in the study and their families in several critical areas: (1) increased employment; (2) increased number of employed beneficiaries who have substantive earnings; (3) reduced benefits; and (4) increased beneficiary income (earnings plus benefits).

Based on the early pilot period, SSA established a goal to enroll at least 9,000 SSDI beneficiaries and up to approximately 15,000 SSDI beneficiaries across the eight selected POD states (Wittenburg et al. 2018). During the design period, SSA made several changes to the recruitment strategy to reach these objectives.

Based on the requirements of the BBA (Public Law 114-74, Section 823), participation in POD must be voluntary and include informed consent. As we discuss below, this feature prescribes how SSA can implement the demonstration, which has implications for the interpretation of the evaluation findings. We are implementing POD in eight purposively selected states: Alabama, California, Connecticut, Maryland, Michigan, Nebraska, Texas, and Vermont. As discussed in greater detail below, the POD implementation contractor, Abt Associates (Abt) chose these states according to substantive and practical criteria related to their capacity to carry out the demonstration. Abt is engaging with implementation partners in each state through subcontracting arrangements. (State implementation partners may include, for example, state VR agencies and Work Incentives Planning and Assistance providers.) Hereinafter, we will refer to Abt and its state partners as the “POD implementation team.”

The evaluation of POD uses random assignment to create two treatment arms and one control arm. We refer to beneficiaries randomized into the demonstration as treatment and control subjects. Both treatment arms include a benefit offset of $1 for every $2 earned above the larger of the Trial Work Period level (defined as $840 in 2017) and the amount of the subject’s Impairment Related Work Expenses. They differ in their administration of policies involving cases that reach full benefit offset (that is, benefits reduced to zero). In both treatment arms, POD initially suspends benefits. However, in one arm the suspension is not time-limited while, in the other arm, POD terminates benefits after 12 consecutive months of suspension. Beneficiaries in the control arm are subject to the current law. Exhibit B.1 summarizes the services that the two treatment groups and the control group receive. In addition to requiring the recruitment of volunteers, Public Law 114-74 specifies that study volunteers assigned to a treatment group have the right to revert from the new POD rules to current SSDI rules at any point. Although both self-selection and the capacity to revert to current law may constrain the generalizability of this evaluation's findings, an intention-to-treat analysis based on the random-assignment design will yield unbiased, internally valid impact estimates of the offer of new POD rules for the pool of POD volunteers.

Exhibit B.1. POD study design



Evaluator will randomly assign volunteers (DI workers) who provide consent to one of three groups

Treatment Groups (2 arms - 5,000 each)

Benefits are offset $1 for every $2 earned each month above the greater of the TWP level ($840 in 2017) or an itemized level of IRWEs

Auxiliaries are also subject to benefit suspension or termination

The TWP and Extended Period of Eligibility do not apply

Participants receive benefits counseling and can withdraw from POD at any time

The POD benefit offset rules will continue to apply to all participants who remain in the demonstration, including those who have their benefits terminated and re-enter the DI rolls

Control Group (5,000)

Current DI program rules:

- No loss of earnings during TWP and Extended Period of Eligibility
- Benefits cash cliff
- Benefits counseling through WIPA

Participant benefits suspended in months earnings would reduce benefits to $0

T1 – Benefits will NOT terminate if suspended for 12 continuous months

T2 – Benefits will terminate if suspended for 12 continuous months

WIPA = Work Incentives Planning and Assistance program.

SSA is conducting the study and Mathematica is carrying out components of the evaluation on behalf of SSA. Hereinafter, we will refer to this as "the POD evaluation team" or, when clear from context "the evaluation team." The evaluation will include a comprehensive assessment of the implementation of POD and its effectiveness for the subjects enrolled in the eight POD states. The evaluation team will base this assessment on results from the following analyses:

- A **process analysis** will describe the components of POD's infrastructure and assess the functioning and implementation of each component. It will document the program environment; beneficiaries' perspectives on the POD offset; and the fidelity of program operation to the offset design. It will also seek to determine the extent of any problems the POD evaluation team might detect and whether they will affect the impact estimates.
- A **participation analysis** will examine recruitment, withdrawal, and use and nonuse of POD's services and the benefit offset. The first component of the analysis will compare the study subjects' characteristics with those of nonparticipants in the recruitment pool. The second component will compare treatment subjects who remain in their treatment with those

who withdraw from it separately for the two treatment groups. The third component will examine how beneficiaries in the treatment groups use demonstration services and the offset, including the extent to which they report earnings each month, receive benefits counseling, and earn enough to have benefits reduced or terminated under the offset.

- An **impact analysis** will leverage the experimental design to provide internally valid quantitative estimates of the effects of the benefit offset and benefits counseling on the outcomes of subjects. Because the study is using random assignment, the treatment and control groups will have similar observed and unobserved characteristics, on average, when they enter the study. Hence, the evaluation's impact estimates will provide an unbiased assessment of whether the benefit offset can help SSDI beneficiaries who volunteer for the study achieve greater economic self-sufficiency and other improvements in their lives. In addition, information from the process and participation analyses will inform understanding of POD's impacts, and the results from the impact evaluation will support the calculation of POD's costs and benefits.
- A **cost-benefit analysis** will assess whether the impacts of the POD treatments on subjects' outcomes are large enough to justify the resources required to produce them. By placing a dollar value on each benefit and cost of an intervention, a cost-benefit analysis can summarize in one statistic all the intervention's diverse impacts and costs.

As discussed more extensively in Part A, any broader implications drawn based on the impact estimates found in the evaluation will be undertaken cautiously due to the Public Law 114-74 requirement that the demonstration include volunteers who provide informed consent. The evaluation team will use the participation analysis to learn about the types of SSDI beneficiaries who enrolled in POD and, therefore, to whom the results are applicable. In addition, findings from the process study, participation, and impact analysis will build a clear understanding of potential replicability for a similar sample of volunteers. Finally, the cost-benefit analysis for POD will produce cost-effectiveness results on altering the SSDI work rules for the set of volunteers who enroll as study subjects, providing SSA with valuable information about the net costs of the demonstration.

As mentioned in the Supporting Statement Part A, SSA requests clearance for five data collection efforts for POD: recruitment materials and baseline survey; two follow-up surveys; four rounds of qualitative data from POD implementation and operations staff; two rounds of semi-structured interviews with treatment group subjects; and two data collection reporting forms for POD implementation.

B.1 Respondent universe, sampling, and expected response rates

B.1.1. Recruitment materials and baseline survey

The POD implementation team implemented the demonstration in eight states, and the POD evaluation team recruited volunteers from a population of beneficiaries who meet the study's basic eligibility criteria (described below). As discussed in greater detail in Section B.2, recruitment for the POD demonstration occurred over a 12-month period, from January 2018 through the end of 2018. The following subsections provide more information about how the states were selected and the study volunteers were recruited. Importantly, because states were selected in a purposive manner and applicants self-selected based on an unknown mechanism, the study sample cannot be construed as being drawn from a larger target population with well-defined probabilities. Hence, it will not be possible to make statistical inferences about any larger population than the volunteers included in the demonstration.

Selection of states. The POD implementation team chose the eight states implementing POD (listed previously) according to substantive and practical criteria; these states were not randomly sampled from a larger target population. Factors they used to select these states included: (1) having sufficient numbers of SSDI recipients to meet the demonstration's enrollment target; (2) covering a diverse set of areas that we expect will reflect a wide range of beneficiary experiences, economic conditions, and contextual factors affecting implementation; and (3) being committed to providing work-incentive counseling, administering the offset, and delivering other services and technical assistance necessary for successful implementation of POD.

Selection of eligible beneficiaries. SSA established seven basic eligibility criteria to identify SSDI beneficiaries who may qualify for the demonstration. These criteria require that applicants to the demonstration: (1) be currently receiving SSDI benefits or have their benefits temporarily suspended due to work; (2) be the primary recipient of SSDI benefits, as opposed to drawing benefits through another person's entitlement; (3) be age 20 or older by the January 2018 and younger than age 62 by June 2021 (the end of the evaluation period); (4) reside in one of the eight study states; (5) are not participating in another SSA demonstration; (6) are not considered be at risk of receiving overpayments, according to a profiling model developed by SSA; and (7) do not have a pending review of continuing eligibility for SSDI benefits.

The POD evaluation team conducted recruitment and outreach activities to solicit volunteers from a set of approximately 420,000 beneficiaries identified in SSA's program data as meeting the basic eligibility requirements noted above starting in January 2018. We randomly sampled this target group for direct outreach from the program data. (Details about sampling methods are in Section B.2.1.1, and details about recruitment and outreach activities are in Section B.3.1.) However, the respondents who volunteer for the demonstration are a small and self-selected set of beneficiaries from the solicitation pool. The evaluation team screened out any volunteers who declined to consent; moved out of the demonstration catchment area; sent an incomplete baseline questionnaire; or returned the forms after recruitment ended for the site. The evaluation team expected to receive responses from 16,500 beneficiaries in total, of whom up to 15,000 could become study "subjects," meaning they consented to participate and met all of the eligibility criteria. Although the evaluation team solicited volunteers in a way that mimics hypothetical outreach efforts under a voluntary national program (see Part A for more details), study subjects ultimately represent a non-random subset of beneficiaries who they contacted. Hence, the ultimate pool of subjects might not constitute a statistically representative population of SSDI beneficiaries.

The members of the POD evaluation team randomly assigned eligible study subjects to the two treatment groups and a control group. Further, they structured the random assignment process so that they grouped together study subjects over time and then assigned in batches. This allowed the evaluation team to stratify the random assignment process to improve the balance across study groups, as discussed in Sections B.3.1 and B.4.1.

Response rates. Among study subjects, the response rate for the baseline survey was 100 percent, by construction. In this section, we focus first on the rate of enrollment into the study, and then on those who choose to remove themselves from the study.

As a starting assumption for the pilot, the POD evaluation team assumed a recruitment yield of 5 percent, based on experience from previous SSA demonstrations (Wittenburg et al. 2018). This estimate is lower than the 6 percent yield from Stage 2 of the Benefit Offset National Demonstration (BOND), a demonstration with an intervention and goals similar to those in POD. The evaluation team used a more conservative estimate for the recruitment yield for POD because the BOND offset was financially more attractive to the beneficiaries recruited for that demonstration. (Sections B.3.1 and B.4.1 further discuss efforts that we took to maximize recruitment yields, given the resources available for the evaluation.)

In recruitment reports to SSA, we will use SSA program files to provide details about how volunteers compare to those who do not volunteer. A major advantage of SSA program files is that they include detailed information about beneficiaries' demographic, impairment, and program characteristics on the universe of all SSDI beneficiaries. Hence, we have a rich source of information to understand differences between those who volunteered, those who did not respond to the survey, and those who responded but did not volunteer for the demonstration. Our plan is to compare the demographic, impairment, and other baseline characteristics of study subjects to those of non-volunteering beneficiaries in the catchment areas. We will also separately tabulate non-volunteering beneficiaries who were contacted, those whom the study attempted to contact but did not reach, and those whom the study did not attempt to contact. These comparisons will help SSA understand the ways in which POD subjects differ from all SSDI beneficiaries and therefore how lessons from POD may or may not apply to the SSDI beneficiary population.

Based on theory and on findings from the BOND evaluation, we anticipate that eventual POD subjects will disproportionately include those who are most likely to benefit from the new rules, including those who have relatively higher benefit amounts, have more substantial earnings at baseline, have no SSI or other benefits (such as private disability insurance), and are near or past their Grace Period. Such beneficiaries will generally have stronger incentives to volunteer for POD relative to other beneficiary groups. In addition, we expect that beneficiaries who have a severe vision impairment and, hence, a higher blind SGA amount will have less incentive to participate, given the advantages of the higher SGA amount under current rules relative to the TWP amounts in POD.

The differences in volunteer rates by characteristics and, specifically, our inability to observe impacts for non-volunteers, are important for interpreting impact findings, especially in noting the limitations of generalizing beyond the study sample. For instance, theory predicts that some beneficiaries who earn between the TWP and the SGA amounts under current law will choose to work less under POD, but we might not observe such behavior among volunteers, because such beneficiaries are unlikely to volunteer in the first place. Hence, any attempt to draw broader conclusions about the impacts based on the POD evaluation findings must carry a caveat based on the unobserved impacts for this group that are not represented in the study.

Despite these limitations, the demonstration findings will provide information on how those who do volunteer benefit from the new POD rules, which has important implications for assessing potential replicability or scalability. The impact findings will provide important information on whether the new rules result in earnings increases or benefit decreases for the subset of beneficiaries, which provides insights to POD as a proof of concept. Additionally, the process-related findings will provide insights into how the simplified operations change the administration of benefits, which might inform broader changes in policy that could be tested in a subsequent evaluation.

Following enrollment, we track attrition from the evaluation. Attrition occurs when demonstration subjects contact the evaluation team and ask the team not to include their data in the evaluation when they leave services. Evidence from previous Stage 2 BOND findings indicates that 0.7 percent of treatment subjects voluntarily withdrew from the evaluation five years following enrollment. Other SSA demonstrations have had similarly low attrition rates. In part, this reflects a major advantage of using administrative records in the evaluation, given that administrative data are regularly updated for all beneficiaries and do not deter study subjects from participating in the way that survey data can.

It is important to note that a special feature of POD is the requirement of Public Law 114-74 that allows treatment subjects to revert back to current law at any point, which we refer to as "service withdrawals." Service withdrawal from treatment services occurs when a treatment subject asks to revert to current law. The incentives for T2 subjects to withdraw are especially important because it is possible some might enter the demonstration with the opportunity to receive the protection from termination under the T1 treatment and subsequently drop out if they are randomly assigned to T2. Consequently, service withdrawals are an important outcome for the evaluation. When people withdraw from services, we will continue to track their outcomes in administrative records unless they request to no longer be part of the evaluation when they withdraw from the treatment arm (as noted above, it is relatively rare to ask to withdraw from the evaluation completely). Based on incentives, we expect the service withdrawal rate for T1 to be higher than for the T2 group, though the magnitude of the difference is an empirical question that the demonstration will answer. Understanding the degree of service withdrawals is important for addressing the Congressional requirements for service withdrawals. This information is also important to inform SSA designs for work incentives as it provides some indication of how willing people might be to continue with services if termination is still possible, which we can test by comparing withdrawal rates between the T1 and T2 groups.

B.1.2. Follow-up surveys

Both of the two follow-up surveys will yield information about the outcomes of a target population consisting of the up to 15,000 study subjects enrolled in the demonstration. The POD evaluation team will field the Year 1 survey to a 50-percent subsample, as discussed in Section B.2.1.2. The evaluation team will field the Year 2 survey to all study subjects (i.e., up to 15,000 subjects). They expect to achieve at an 80 percent response rate for both surveys.

B.1.3. Qualitative data from POD implementation and operations staff

As discussed in Part A, the POD evaluation team will conduct four rounds of qualitative data collection, one in each of the first four years of the evaluation. In rounds 1 and 3, the evaluation team conducts in-person interviews during site visits; in rounds 2 and 4 (the interim years between site visits) they will conduct telephone interviews. Although, the evaluation team conducts in-person visits for rounds 1 and 3 of data collection, they will collect information by telephone to reduce the costs of data collection in these rounds if they find key informants geographically dispersed. POD sites will likely vary in geographic location, organization, and staffing arrangements, so the specific number of interviews might vary across sites. The evaluation team expects to interview an average of 5 respondents at each site for a total of 40 each round (160 total).

The evaluation team purposively selects these respondents from within the eight POD states to provide information about an array of implementation experiences. They select respondents based on their role and knowledge of POD at each stage of project implementation. Thus specific respondents selected may vary by round; however, the evaluation team expects them to include the VR agency director, POD work incentives counselor, staff at VR agencies or other state implementation partners, technical assistance provider, and additional program partners identified by the POD implementation team. The evaluation team asks directors from the implementation team's state partners to identify people at the site who can provide the required information for each round of interviews. Because they expect sites to provide the information needed to assess program implementation and fidelity as the POD implementation team communicated to them, the evaluation team expects a 100 percent response rate for the qualitative interviews.

B.1.4. Semi-structured interviews with treatment group subjects

The POD evaluation team conducts two rounds of semi-structured telephone interviews with 9 POD participants in each of the 8 sites for total of 72 interviews in each round (144 total across the two rounds). They interview beneficiaries from both treatment groups and from key subgroups of interest (for example, beneficiaries who requested to withdraw, low earning offset users, and high earning offset users). The beneficiaries selected for this evaluation component constitute a purposive sample of POD treatment group members selected using information from the POD implementation team's management information system. The evaluation team selects subjects using quota sampling within strata established by state; random-assignment group; and key subgroup measures to obtain 9 interviews per state and round. Although the evaluation team uses random subsampling within strata to reduce the sample size in each stratum from the full universe, they replace any non-respondents with alternate beneficiaries who meet the same criteria. That is, if a potential interviewee cannot be located or declines to participate, the evaluation team will randomly select another from that person's same stratum to maintain the quota set for the stratum. Because the evaluation team purposively selects respondents, they use information about them for illustrative purposes only in the evaluation reports. They will not statistically generalize findings based on this sample to any larger population of treatment group subjects.

B.1.5 Implementation Data collection

Data collection for POD's implementation will take place over a five-year period and will include two data collection activities: (1) POD Monthly Earnings and IRWE Reporting Form; and (2) POD EOYR Form. Of the up to 10,000 POD participants assigned to the POD treatment groups, SSA estimates the implementation team will require 40 percent (approximately 4,000

POD participants) to report earnings and IRWEs because of earnings that exceed $840 per month, the threshold set for POD. Of the up to 4,000 participants, the implementation team expects 65 percent (approximately 2,600) to report earnings and IRWEs.

B.2. Procedures for sampling and estimation, degree of accuracy, specialized procedures, and periodic data collection cycles

B.2.1. Statistical methodology for stratification and sample selection

As noted in Section B.1, the two main ways in which the POD evaluation team uses statistical methods to select samples for the POD demonstration are (1) to select stratified random samples of beneficiaries to target when conducting recruitment for the demonstration, and (2) to select simple random samples of study subjects to include in the first follow-up survey. We discuss each type of sampling in the following subsections.

B.2.1.1. Recruitment materials and baseline survey: selecting stratified random samples of beneficiaries to target in recruitment efforts

The POD evaluation team conducted recruitment for the demonstration from January 2018 through December 2018, gradually rolling out throughout each of the 8 states. As part of recruitment, the evaluation team required all beneficiaries to have their materials post-marked by December 31, 2018 to be eligible for POD random assignment. The first three months included a pilot phase to refine assumptions about the study population. The last nine months included the full rollout. As discussed in Section B.3.1, the evaluation team's general approach to recruitment in both the pilot phase and full rollout phase included a combination of direct mailings; telephone follow-ups; and indirect methods to contact potential recruits. The evaluation team sent mailings to potential volunteers in each selected area using structured random sampling techniques from lists of beneficiaries in the given area who meet SSA's basic eligibility criteria.[1]

During the pilot phase of recruitment, the POD evaluation team selected eligible beneficiaries using structured simple random sampling that includes implicit stratification on several key characteristics available in the program data. This approach allowed the evaluation team to obtain information on the potential recruitment yield of a broad range of beneficiary characteristics. Examples of such characteristics are state, age, duration of SSDI receipt, and primary impairment type. Implicit stratification resulted in a sample for which the distribution of these key characteristics is approximately the same as in the full population.[2] The evaluation

[1] As discussed in Section B.3.1, the evaluation team also purposively selected a subset of these 420,000 beneficiaries when conducting telephone follow-up outreach efforts for the recruitment effort in the pilot. Following the pilot, SSA directed the evaluation team to use post-card follow-up based on pilot findings.

[2] Implicitly stratifying by state maximized the statistical power of pooled estimates that incorporate data on beneficiaries in all POD states. However, when information about the number of eligible beneficiaries is available, the evaluation team may consider explicitly stratifying (rather than implicitly stratifying) by state if the anticipated size of the eligible population differs substantially by state. Explicit stratification allows the evaluation team to set a floor on the number of beneficiaries targeted for recruitment within each state, which improves statistical power for within-state estimates while potentially resulting in only modest reductions in statistical power for pooled, multi-state estimates. The evaluation team makes decisions to switch from implicit stratification to explicit stratification in conjunction based on an updated statistical power analysis conducted during the study's design phase.

team implemented implicit stratification using the serpentine sorting technique of Chromy (1979).

Once we concluded the pilot phase, the POD evaluation team continued to draw structured random samples, but adapted the sampling approach to change the sampling rate or incorporate explicit strata when selecting the remaining potential subjects during the full rollout period. Over the course of the pilot and full rollout period, the evaluation team continued to monitor and analyze recruitment yields, defined as the number of enrolled subjects divided by the number of beneficiaries targeted in the recruitment efforts. They used this information in combination with preexisting data on beneficiary characteristics from SSA program records to fine-tune the recruitment process. For example, the evaluation team conducted chi-squared tests to assess whether recruitment yields differed significantly by age; duration of SSDI receipt; primary impairment type; and other factors. The evaluation team also conducted a multivariate analysis of factors associated with the decision to volunteer using a linear or logistic regression analysis similar to that described later in Section B.2.3.

Based on these results, as well as responsiveness to variation of recruitment efforts tested during the pilot phase (discussed in Section B.4.1), the evaluation team adapted its sampling approach to expand the sample (Wittenburg et al. 2018). The adaptations were necessary because the initial yields were below the 5 percent yield rates noted above. Additional details about this approach are in Sections B.3.1 and B.4.1. These adaptations enabled the evaluation team to make future mailings and outreach more targeted so that they could recruit at least 9,000 subjects and up to 15,000 subjects for the demonstration.

B.2.1.2. Follow-up surveys: selecting stratified random samples of subjects for inclusion in the first follow-up survey

As already noted, the POD evaluation team will field the first follow-up survey to a 50-percent subsample of the subjects enrolled in the demonstration. Data collection efforts for this survey will span 15 months, with a rolling release of sample that will mirror the 12 months of study recruitment plus another 3 months to complete interviewing in the final sample release. The evaluation team will aggregate subjects from a given month into cohorts and release them on a monthly basis so that they interview each beneficiary as close to their one-year anniversary as practical. Assuming a roughly even pace of enrollment, each monthly release will contain approximately half of the subjects enrolled in the demonstration during the corresponding recruitment month. The evaluation team will select the subjects for inclusion using a structured random sampling technique similar to what we previously described for the study recruitment effort. The evaluation team will use implicit stratification so that the distribution of key characteristics in each release is similar to the distribution of those characteristics in the full group of subjects that they could consider for the given release.

The evaluation team will field the second follow-up survey to all POD subjects. Data collection will span 15 months that roughly correspond to two-year anniversaries of study subjects’ enrollment dates plus a 3-month closeout period. The evaluation team will group the data into monthly releases and attempt to survey all subjects in each release. There is no need for stratification or subsampling because the second follow-up survey includes all demonstration subjects.

B.2.1.3 Implementation Data collection

The implementation team requires all beneficiaries assigned to POD treatment groups whose earnings exceed the POD threshold to report earnings to SSA. They do not plan on sampling for the implementation data collection.

B.2.2 Estimation procedure

The POD evaluation team expects that the estimation methods differ more substantively by analysis type than by data collection effort. Therefore, the following subsections describes the methods that Mathematica plans to use for each of the major analysis components of the POD evaluation.

Process Analysis. As discussed in Part A, the POD evaluation team uses site visit and semi-structured interview data to provide a detailed description of the POD programs: how they implement the POD program; the contexts in which they operate; the program operations and their fidelity to design; and subjects' perspectives of POD. These detailed descriptions assist in interpreting program impacts, and identifying program features and conditions necessary for effective program replication or improvement. The evaluation team gathers information using a range of techniques and data sources to describe the programs and activities fully. The evaluation team uses the Consolidated Framework for Implementation Research (CFIR; Damschroder et al. 2009) to guide the collection, analysis, and interpretation, of qualitative data. The CFIR is an implementation framework that guides systematic assessment of the multilevel and diverse contexts in interventions implementation and helps understanding of the myriad factors that might influence intervention implementation and effectiveness.

Using the CFIR allows the evaluation team to structure the analyses of POD implementation to produce results based on objective, reliable qualitative data across the key domains related to the program environment; early implementation, program operations; fidelity; systems change; and beneficiary perspectives. For each of these domains, the evaluation team develops measurable constructs that align with research questions for the POD process analyses shown in Part A and Appendix C. Based on this framework, they will be able to produce tables that summarize the major process findings in the study's reports. The CFIR may also allow the evaluation team to make systematic use of qualitative data as part of the impact analysis, permitting an examination of impacts vary with certain implementation constructs.

Participation Analysis. The participation analysis examines recruitment, withdrawal, and use and non-use of POD's services and the benefit offset. The first component of the analysis compares the study subjects' characteristics to those of nonparticipants in the recruitment pool. The second component compares treatment subjects who remain in their treatment to those who withdraw from it, separately for the two treatment groups. The third component examines how subjects in the treatment groups use demonstration services and the offset, including the extent to which they report earnings each month, receive benefits counseling, and earn enough to have benefits reduced under the offset. These analyses make comparisons of these intermediate outcomes across subgroups of subjects using chi-squared tests and multivariate linear regressions. For binary variables, we compare the main estimates from a linear probability model to those from logistic regression for the primary impact outcomes and for the key participation analysis outcomes, especially considering that some of the participation analysis outcomes have low prevalence and are more sensitive to the model specification. We describe the specifics of the control variables for the model in the impact section below.

Impact Analysis. The objective of the impact analysis is to provide statistically valid and reliable estimates of the effects of POD on the economic outcomes of subjects enrolled during the recruitment phase. The POD evaluation team relies on the experimental design to estimate the causal impacts of the new benefit offset policies available to subjects through the demonstration. Random assignment enables estimation of the net impact of those policies by comparing average outcomes across the treatment and control groups. The evaluation team's analysis will focus on intent-to-treat estimates, which measure how the *offer* of the POD offset shaped volunteers' behavior after they enrolled in the demonstration.

Primary outcomes. To avoid concerns about data mining and reduce the extent of false positives, the POD evaluation team pre-specifies a parsimonious set of primary outcomes in important study domains. A preliminary list of those domains includes employment; earnings above a substantive threshold; SSDI benefit payments received; and total income (benefits + earnings). The evaluation team will refine this list and to define specific outcomes prior to conducting the analysis. Evaluation reports also include results for secondary outcomes related to services received, employment, receipt of benefits, and income. However, the evaluation team considers these results exploratory. This approach, which is based on guidance for evaluations conducted by the Department of Education's Institute for Education Sciences, strikes a balance between addressing the multiple comparisons problem while maintaining the evaluation's ability to detect policy-relevant impacts. By limiting the number of primary outcomes tested in the impact analysis, this approach reduces the likelihood of false positives without undermining the evaluation's statistical power to detect true impacts on any single outcome.

Model. The main impact model is a weighted linear regression model that pools together data from all states and includes interactions to allow for state-level heterogeneity in beneficiary characteristics, program implementation, and contextual factors. The evaluation team uses the following interacted regression model to estimate the pooled impact of each POD treatment arm:

$$[1] \qquad y_{is} = c_s + d_1 T1_i + d_2 T2_i + b'_s X_{is} + u_{is}$$

where y_{is} is the outcome of interest for individual i in state s, c_s is a state fixed effect, $T1_i$ and $T2_i$ denote assignment to the two POD treatment groups, X_{is} denotes a set of K covariates included to improve precision, and u_{is} is an error term. The set of covariates includes gender, age, whether completed high school, whether non-white, whether living in poverty, whether health status is poor, categorical variables for primary impairment, years since onset of disability, recent annual earnings, whether earned $1,000 or more in at least one of previous 12 months, completion of TWP, completion of the Grace Period, monthly SSDI benefits, number of months receiving SSDI, and whether SSDI only or a concurrent SSI beneficiary. The evaluation team estimates impacts separately for pairwise comparisons between all three random-assignment groups (T1 versus control, T2 versus control, and T1 versus T2).

The results from the linear regression model analysis will have an immediate and straightforward interpretation for understanding the effects of the benefit offset. Estimates of d_1 and d_2 represent the intent-to-treat impacts of being assigned to each treatment arm relative to the control group; the evaluation team will mean-center the covariates so that c is readily interpretable as the mean level of the outcome of the control group. The team can also use these estimates transparently to produce adjusted means for the two treatment groups (based on $c + d_1$ and $c + d_2$, respectively) to facilitate a graphical representation suitable for a wide audience. In contrast, nonlinear models such as logit produce estimates that (1) are less immediately interpretable and (2) tend to closely align with results from a linear model when converted into more meaningful impact estimates (Wooldridge 2010). However, for each binary outcome that addresses a primary research question, the evaluation team will estimate a logistic regression model to verify that the logistic and linear regression model estimates have the same direction and significance level as well as to measure the difference between the two estimates.

We will also examine estimates from a logistic regression for the key participation analysis outcomes and compare those estimates to the linear probability model. The key outcomes for the participation analysis, such as enrolling in the demonstration, are likely to have low prevalence, in which case the logistic and linear probability models could yield somewhat different estimates (though they are unlikely to substantively differ).

The POD evaluation team will use a model similar to equation [1] to estimate impacts for select subgroups defined by baseline characteristics and to compare impacts across these subgroups. This model will include interaction terms between each of the treatment indicators ($T1_i$ and $T2_i$) and binary indicators for each subgroup. The corresponding coefficients will measure the subgroup-specific impacts of the POD offset policies, including for impact estimates for individual POD states or groups of states.

Standard errors. Because the implementation team chose POD states purposively and the evaluation team will not sample volunteers who enrolled in the demonstration probabilistically from a known population, estimates of precision focus on inference about the baseline sample of subjects who the evaluation team randomly assigns. Therefore, variances can be straightforwardly estimated using two techniques. First, for linear models, the evaluation team will estimate standard errors of the impacts based directly on the study's randomized design using the methods developed by Mathematica for the Department of Education (Schochet 2016). Second, if using nonlinear models for binary outcomes, the evaluation team will explore the distribution of the error terms to establish the proper nonlinear model to apply (for example, logit or probit). In this case, they would then use other analysis methods to estimate model-based standard errors for all impacts, including both linear and nonlinear outcomes, based on standard econometric techniques (Cameron and Trivedi 2005).

Cost-benefit analysis. The POD evaluation team will conduct the cost-benefit analysis for POD using an approach it successfully adopted in other evaluations, including BOND (Bell et al. 2011) and the Youth Transition Demonstration (Fraker et al. 2014). To do this, they will develop a comprehensive accounting framework that incorporates a range of perspectives to guide cost-benefit data collection, analysis, and reporting. These perspectives include those of the POD treatment group members, the federal government, and society as a whole. The evaluation team will use the study impact estimates in conjunction with actuarial estimates and program data to develop line-item entries of costs and benefits from each perspective. Examples of such line items include program administration costs, benefit receipt, labor market productivity, estimated tax revenues, and other direct cost of POD services. As with the impact analysis, the cost-benefit analysis only directly applies to the POD study subjects, who have voluntarily participated in the demonstration. The evaluation team will use the participation and process analyses to assess what broader inferences can be made from the cost-benefit analysis.

Because some benefits and costs occur at different times, the POD evaluation team will make two adjustments when aggregating them into line items for cost-benefit accounting. First, they will use a price deflator, such as the implicit gross domestic product price deflator, to convert all benefits and costs occurring in later years into constant dollars. Second, they will use a discount rate to convert all future benefits and costs to their present value. This discount rate will equal the rate SSA uses in its actuarial projections of the SSDI Trust Fund balance.

As in most cost-benefit analyses, there will be some uncertainty surrounding estimates of the benefits and costs of the POD treatments that is not easily quantifiable. Some amount of this uncertainty may arise due to imprecision in estimating the impacts that feed into line item in the accounting framework above, but other uncertainty may arise based on assumptions and the modeling structure. Hence, the POD evaluation team will develop methods to establish reasonable bounds on the estimates that meet SSA's needs for the cost-benefit analysis.

B.2.3. Degree of accuracy needed for the purpose described in the justification

The expected sample size of eligible study subjects supports an analysis that can reliably distinguish the overall impacts of each POD policy innovation from other factors shaping subjects' outcomes. Calculating minimum detectable impacts (MDIs) is a standard way to characterize the expected precision of the evaluation's results, given the sample sizes and research design. MDIs quantify the smallest true impact we are likely to find to be significantly different from zero, based on a two-sided statistical test of differences. The POD evaluation team has calculated MDIs separately for outcomes measured in the program data for all study subjects and outcomes measured for the subsets of subjects who complete the two follow-up surveys.

MDIs for outcomes measured in program data for all subjects. The POD evaluation team expects that impacts based on program data are likely to have sufficient precision to assess policy-relevant impacts of each of the two POD policy packages reliably. The first two columns of Exhibit B.2 includes illustrative outcomes for: (1) employment; and (2) substantive earnings, defined as annual earnings greater than 12 times substantial gainful activity (SGA). Both outcomes are available in the SSA program data and therefore available for all subjects enrolled in the study. The exhibit summarizes the MDIs for the pool of evaluation subject as a whole, as well as for analyses of potential impacts for potential subgroups of interest to SSA, as well as subjects within each demonstration state. The evaluation team expects the POD MDI for employment to be 2.2 percentage points and the MDI for having annual earnings over the SGA amount to be 1.5 percentage points. These MDIs are for pairwise comparisons between two random-assignment groups, and they suggest that the evaluation would be able to identify even relatively small impacts for both outcomes reliably, compared to current law. For example, the MDI for the employment rate is 4.4 percent the size of the assumed prevalence under current law (2.2/50 = 0.044). The MDIs are all about 15 percent smaller for a comparison of both treatment groups combined versus the control group. In Wittenburg et al. (2018), we also showed MDIs for smaller sample sizes, which is relevant given we just completed recruitment. The main finding is that MDIs do not substantively change for key outcomes, such as over SGA-level earnings. For example, the report showed that the evaluation team expects the overall MDI for estimated impacts on SGA-level earnings for each treatment arm to be 2 percentage points.

Impacts of POD would need to be somewhat larger for the evaluation team to detect for subgroups based on beneficiary characteristics or for particular states reliably, but the evaluation team expects these subgroup analyses to still have enough precision to be informative. For example, impacts would have to be almost 25 percent higher for the evaluation team to detect for a subgroup comprising 66 percent of study subjects reliably, which might represent subjects with a recent work history or could reflect a decision to pool data from a subset of states based on the process study. Impacts would need to be correspondingly larger for them to detect for more focused subgroups containing smaller percentages of the study subjects reliably.

MDIs for outcomes measured using survey data for subsets of subjects. The POD evaluation team expects to have less precision for survey-based outcome measures because the analysis sample sizes will be smaller, but they expect to be able to detect modest-sized impacts overall and for all but the smallest subgroups. The final two columns of Exhibit B.2 report minimal detectable impacts (MDIs) on the share of subjects actively searching for work at the time of each of the two follow-up surveys. Based on the expected response rates of 80 percent, the evaluation team expects a respondent sample size of up to 6,000 for the Year 1 survey (which they will field to half of the subjects) and up to 12,000 for the Year 2 survey. For comparative purposes, they assume the prevalence of work search in the control group at both points in time to be 15 percent – the same as the prevalence of annual earnings greater than 12 times the SGA amount. MDIs for pairwise comparisons of work search are 2.5 percentage points based on the Year 1 survey and 1.7 percentage points based on the Year 2 survey.

Exhibit B.2. Minimum detectable impacts for POD evaluation

	Outcomes measured in program data		Outcomes measured using survey data	
Group/subgroup	Annual employment rate	Annual earnings > 12 × SGA	Searching for work at time of Year 1 survey	Searching for work at time of Year 2 survey
MDIs for pairwise comparison of two study groups				
All subjects	2.2 pp	1.5 pp	2.5 pp	1.7 pp
66% subgroup	2.7 pp	1.9 pp	3.0 pp	2.1 pp
50% subgroup	3.1 pp	2.2 pp	3.5 pp	2.5 pp
33% subgroup	3.8 pp	2.7 pp	4.3 pp	3.0 pp
Subjects in a large state	4.9 pp	3.5 pp	5.5 pp	3.9 pp
Subjects in a medium state	6.9 pp	4.9 pp	7.8 pp	5.5 pp
Subjects in a small state	11.9 pp	8.5 pp	13.4 pp	9.5 pp
MDIs for comparison of both treatment groups combined vs. control group				
All subjects	1.9 pp	1.3 pp	2.1 pp	1.5 pp
66% subgroup	2.3 pp	1.7 pp	2.6 pp	1.8 pp
50% subgroup	2.7 pp	1.9 pp	3.0 pp	2.1 pp
33% subgroup	3.3 pp	2.3 pp	3.7 pp	2.6 pp
Subjects in a large state	4.2 pp	3.0 pp	4.7 pp	3.4 pp
Subjects in a medium state	5.9 pp	4.2 pp	6.7 pp	4.7 pp
Subjects in a small state	10.3 pp	7.4 pp	11.6 pp	8.2 pp
Key assumptions				
Assumed outcome prevalence in the control group	50%	15%	15%	15%
Total sample size	15,000	15,000	6,000	12,000

Note: Additional assumptions for all MDIs: large state = 3,000 per group; medium state = 1,500 per group; small state = 500 per group; the evaluation requires at least an 80 percent chance of correctly identifying true impacts as statistically significant using two-tailed statistical tests with a 5 percent significance level; the POD evaluation team will estimate impacts using regression models that include baseline covariates explaining 40 percent of the variation in employment outcomes; and analysis weights or adjustments for heteroscedasticity will not substantially alter variance estimates. Further assumptions about MDIs for survey outcomes are that: the evaluation team will field the first survey to half of the study subjects; it will field the second survey to all subjects; and approximately 80 percent of potential respondents will complete the survey(s) fielded to them. Estimates of MDIs at different sample sizes are available in Wittenburg et al. (2018).

SGA = substantial gainful activity, pp = percentage point.

B.2.4. Unusual problems requiring specialized sampling procedures

During the pilot phase of recruitment, the evaluation team gathered information about volunteer interest and ran recruitment experiments to understand how volunteer rates vary by different outreach methods and beneficiary characteristics. See Sections B.3.1 and B.4.1 for additional details. They then used this information to adapt their stratified sampling procedure used to identify beneficiaries to target in the recruitment effort, as discussed in Sections B.2.1.1 and B.3.1.

B.2.5. Periodic cycles to reduce burden

The POD evaluation team does not plan to use a less-than-annual periodicity of data collection for this evaluation because it expects that doing so would increase, rather than decrease, respondent burden. Implementation data collection must occur monthly for SSA to obtain the information on earnings and IRWEs needed to calculate monthly SSDI benefits. Less than monthly data collection for POD implementation would not satisfy the requirements of POD.

B.3. Methods to maximize response rates and data reliability

B.3.1 Recruitment materials and baseline survey

Designing accessible informative recruitment materials. The law requires informed consent from POD subjects, and helps the evaluation team produce more reliable estimates of POD's impacts on the study population for whom the treatment is salient. Consequently, the evaluation team's recruitment plans centered around providing beneficiaries with the information needed to make informed choices. Their approach relied on a combination of (1) direct outreach to prospective subjects and (2) indirect outreach via trusted stakeholders – organizations commonly engaged with beneficiaries, especially around employment issues. These two efforts occurred concurrently in each area within which the evaluation team rolls out POD during the recruitment phase.

The POD evaluation team designed direct mailings to (1) spark beneficiary interest in participating in POD; and (2) meet current ethical and regulatory standards for providing full, objective information about the demonstration rules and their consequences for subjects. The outreach materials included a letter and a brochure (as shown in appendix A). The evaluation team wrote these materials in a way that seeks to entice beneficiaries who are likely to benefit from POD while dissuading beneficiaries for whom assignment to a treatment arm is likely to be detrimental. They structured the materials to be succinct, yet complete and accessible, information. All outreach material that the evaluation team developed is readily accessible to people with low education levels and met all 12 basic elements outlined in the U.S. Department of Health and Human Services Informed Consent Checklist (1998) §46.116 Informed Consent forms. To complement the mailings, the evaluation team attempted to contact a subset of beneficiaries (approximately 25 percent of the total sample) to: (1) verify that they received a packet and reviewed the contents; (2) provide further explanations of POD; and (3) offer help completing the forms. As discussed in Section B.4.1, during the pilot recruitment phase, the evaluation team used random assignment to select the subset of beneficiaries receiving mailings that also receive telephone follow-ups. Following the pilot period, SSA directed the evaluation team to use follow-up post cards instead of phone calls as the follow-up post cards were the most cost effective way to increase yields.

At the same time, the evaluation team organized indirect outreach efforts to key stakeholders to establish the legitimacy of POD in each implementation area, a critical component of successful recruiting (Derr et al. 2015). Indirect outreach included an information dissemination campaign just prior to the start of recruitment, targeted to stakeholder organizations that serve beneficiaries in various ways, including state vocational rehabilitation agencies; organizations administering SSA-funded training and work-incentive programs; and various advocacy organizations. The campaign provided information about the nature of POD and the types of beneficiaries who will find it attractive. It included: letters, flyers, posters, and brochures for SSA offices and other local organizations serving beneficiaries; an informational website about POD; and notifications from SSA to beneficiaries (for example, in information packets for new beneficiaries). In addition to these efforts, the evaluation team presented and recorded three webinars to explain and promote POD to local staff at stakeholder organizations in each recruitment area

Adaptive sampling to identify potential subjects from program data. As noted previously, the evaluation team implemented a three-month pilot period to assess the efficacy of the recruitment effort, particularly with respect to assumptions about volunteer rates for POD. Based on the actual volunteer rates observed, the evaluation team adapted the sampling procedure. As documented in Wittenburg et al. (2018), the evaluation team made changes to reflect updated actual recruitment yields.

Over the remainder of the recruitment period, the POD evaluation team continued to adapt its sampling approach based on progressively more information about recruitment yields. These types of adaptations allowed the evaluation team to meet SSA's revised goals of recruiting at least 9,000 subjects and up to 15,000 subjects more efficiently, particularly when done in combination with other updates to recruitment plans that might be undertaken based on the recruitment experiments described in Section B.4.1.

Use of incentives. An important factor in recruitment is the use of incentive payments, which the evaluation team employed throughout the entire recruitment period. As described in Section A.9, the use of incentive payments can also reduce longitudinal attrition to the follow-up surveys. The POD evaluation team paid a monetary incentive to each POD volunteer who returned a completed baseline self-administered questionnaire and consent form, regardless of whether they gave consent for random assignment and further participation in the study. This monetary incentive was $25.

Data reliability. The POD evaluation team developed the baseline survey instrument and recruitment materials using materials developed and fielded on recent similar SSA demonstrations such as BOND and the Promoting Readiness of Minors in Supplemental Security Income (PROMISE) initiative. Several experts on the POD evaluation team, including survey researchers, disability policymakers and practitioners, reviewed the draft instrument and recruitment materials and helped to refine them further. The evaluation team also field-tested the instrument and consent form with a small number of SSDI beneficiaries as described in section B.4 below.

Addressing item nonresponse. As discussed previously (Section B.1.1), there was no unit-level nonresponse for the baseline survey by construction. The past experience of the POD evaluation team suggests that item-level nonresponse will be low for the baseline survey, although some item nonresponse is inevitable. To address this when running subsequent analysis that draw on the baseline survey, the evaluation team will consider using one of several standard imputation techniques, as described in Allison (2001), depending on the pattern item nonresponse observed in the final study sample.

Balance of study sample and integrity of random assignment. The POD evaluation team uses a management information system to maintain balanced study groups and minimize on-site contamination risks. Built-in eligibility and duplication checks prevent staff from enrolling beneficiaries who did not completed a baseline survey and provide written consent or who might have previously enrolled.

In addition, the POD evaluation team developed explicit or implicit randomization strata that may improve the extent to which the distribution of subjects' characteristics is similar in each study group. The evaluation team uses explicit stratification to seek exact balance for batches of subjects they randomized based on a select set of their characteristics determined in conjunction with SSA. Within in each explicit stratum, they use structured selection to stratify on additional characteristics implicitly. Implicit stratification does not guarantee that every one of the additional characteristics will balance perfectly between study groups in any given sample, particularly for rare characteristics. In addition, the evaluation team expects thy can only achieve exact balance through explicit stratification when randomization batches are large enough to include at least three subjects in each explicit stratum. Therefore, during the design phase, the evaluation team developed key factors for explicit stratification based on past research and the experience of the POD evaluation team. They also weigh the tradeoffs between (1) achieving balance on a broader set of characteristics for explicit stratification, and (2) randomly assigning a fewer batches of subjects, each containing a greater number of cases.

Further, the POD evaluation team monitors the random assignment process, checking the balance of study groups using data on subjects recorded in both SSA program data and the baseline survey. Such characteristics could include the same ones used to stratify sampling for recruitment (as discussed in Section B.2.1.1) and other measures determined to be of substantive importance during the design phase of the evaluation. Substantial or statistically significant differences (based on t-tests and chi-square tests, as appropriate) in characteristics across subgroups and assignment status could reveal the need for the evaluation team to adjust the random assignment procedure.

B.3.2 Follow up surveys

Designing and fielding of survey. The follow-up survey is unique to the current evaluation and the evaluation team will use it across all POD study sites, ensuring consistency in the collected data. Staff on the POD evaluation team extensively reviewed the survey, and will thoroughly test it in a pretest involving no more than nine individuals (as described in Section B.4.2). The evaluation team designed the survey fielding methods to both maximize response rates and improve the quality of response data among those who complete surveys. For example, the mixed-mode of administration will offer potential respondents the flexibility to complete the survey in a manner that is most convenient for them; and the evaluation team designed the length of the survey to balance the evaluation's need for information about a variety of topics with the quality of the responses obtained from the survey. Additionally, the evaluation team will assure the respondents of the privacy of their responses to questions, which should yield higher quality data. Finally, the evaluation team will store responses collected by both the web and telephone versions of the survey in a single database, eliminating the need for merging and related data cleaning.

Response rates. The POD evaluation team will address several challenges that can depress response rates to a follow up survey. They will offer the follow up survey in several modes (web, telephone and mail) to encourage response. Subjects will receive an advance letter by mail and via email with a link to the web survey and their unique survey login information so they can easily access the web survey. The evaluation team will mail materials to subjects based on their SSA identified mailing preferences (special notice options), including in Spanish.

The POD evaluation team designed the survey interview to be brief so as not to discourage response or full completion. As discussed in greater detail in Part A, they will offer response incentives to encourage subjects' participation; they will tailor these incentives to entice response by the most cost effective mode (a higher response is offered for web survey mode, followed by telephone and hard copy modes). The evaluation team will also offer a shortened mail survey for subjects who do not respond via web or telephone to complete on paper.

The evaluation team will train bilingual (English and Spanish speaking) telephone interviewers to address subject's questions clearly and effectively and to gain cooperation and avoid refusals. They will also train interviewers in how to work with subjects who need assistance to complete surveys such as via TTY lines. In addition, the evaluation team will train staff (also bilingual) on techniques for locating subjects who are no longer at the address and telephone number provided at enrollment.

Data reliability. The POD evaluation team developed the follow up survey instrument and contact materials using materials developed and fielded on recent similar SSA demonstrations such as BOND and PROMISE. Several experts on the POD evaluation team – including survey researchers, disability policymakers and practitioners, reviewed the draft follow up instrument and contact materials and helped to refine them further. The evaluation team also field-tested the instrument with a small number of SSDI beneficiaries as described in section B.4 below.

Item nonresponse. Although the POD evaluation team's past experience conducting surveys for similar evaluations suggests that rates of item nonresponse on the follow-up survey will be very low, some item nonresponse is inevitable. The follow-up survey primarily collects data on outcome measures for the evaluation team to use in the impact analysis. Imputation of outcome data could lead to biased estimates due to imperfect matches on observables when using a hot-deck procedure (Bollinger and Hirsch 2006). Hence, depending on the pattern of missing data observed, the evaluation team will consider alternative multivariate imputation techniques or omitting subjects with missing data on a given outcome when analyzing that outcome.

Individual-level nonresponse. As with almost any survey, some nonresponse in the follow-up survey is inevitable. The evaluation team will not be able to locate some sample members and others will not be able or willing to respond to the survey. The POD evaluation team expects to attain a response rate of at least 80 percent. In the event that response rates are lower, they will conduct a nonresponse analysis using various data items from the SSA program records and baseline survey. The nonresponse bias analysis will consist of the following steps:

- *Compute response rates for key subgroups*. The evaluation team will compute the response rate for the subgroups using the American Association for Public Opinion Research (AAPOR) definition of the participation rate for a nonprobability sample: the number of respondents who provided a usable response divided by the total number of individuals from whom they requested participation in the survey (AAPOR 2016).[3] The evaluation team will conduct comparisons of the response rate across key subgroups, including most notably the two treatment groups and the control group, as well as subgroups used to stratify sampling for the recruitment and survey field efforts. (The evaluation team will define the latter set of subgroups using pre-enrollment characteristics from SSA program records and the baseline survey.) The goal of this analysis is to determine whether response rates in specific subgroups differ systematically from that of other subgroups or from the overall response rate. This could inform the evaluation team's development of nonresponse weights for use in the analysis.

- *Compare the distributions of respondents' and non-respondents' characteristics*. Again using data from program records and the baseline survey, the evaluation team will compare the characteristics of respondents and non-respondents. They will assess the statistical significance of the difference between these groups using *t*-tests. This type of analysis can be useful in identifying patterns of differences in observable characteristics that might suggest nonresponse bias. However, this approach has low power to detect substantive differences when sample sizes are small, and the large number of statistical tests conducted can also result in high rates of Type I error. Consequently, the evaluation team will interpret the results of this item-by-item analysis cautiously.
- *Identify the characteristics that best predict nonresponse and use this information to generate nonresponse weights*. This is a multivariate generalization of the subgroup analysis described previously. The evaluation team will use logistic regression models to assess the partial associations between each characteristic and response status; propensity scores obtained from such models provide a concise way to summarize and correct for initial imbalances (Särndal et al. 1992). Given the rich set of program and baseline survey data available for this analysis, the evaluation team will use a mixture of substantive knowledge and automated "machine learning" methods to identify covariates to include in the final weights. Examples of automated procedures they could use to produce these weight efficiently include: (1) using pre-specified decision rules, such as those described by Imbens and Rubin (2015) and Biggs, de Ville, and Suen (1991) to select covariates and interactions between them; and (2) identifying and addressing outliers by, for example, trimming weights in a way that minimizes the mean-square error of the estimates (Potter 1990).
- *Compare the nonresponse-weighted distribution of respondent characteristics with the distribution for the full random assignment sample.* In this last step, the evaluation team will compare the weighted distribution of respondent baseline characteristics to the unweighted distribution of the full set of study subjects that went through random assignment. They will make these comparisons for the whole sample and for key subgroups, as described earlier in this subsection. This step will include validation of the nonresponse weights using outcomes measured in the program data for the full sample (but not used in the construction of the weights). This analysis can highlight measures in which the potential for nonresponse bias is greatest, even after weighting, in which case they should exercise greater caution in the interpretation of the observed findings.

[3] This OMB submission uses the terms response and nonresponse, rather than participation and nonparticipation (as in the AAPOR definitions), to avoid confusion with "participation in POD."

B.3.3. Qualitative data from POD implementation and operations staff

Response rates. The POD evaluation team expects that sites will provide the information needed to assess program implementation and fidelity, as the implementation contractor communicated this expectation to the VR agencies and other state implementation partners in the participating sites; thus, we anticipate high levels of cooperation for the qualitative interviews. To explain the purpose of the evaluation and further encourage cooperation, SSA sent a letter to the directors of each program site. The evaluation team follows up with the POD site director in a phone call to describe the information they will gather from site staff and partners during site visits and in-person interviews. To minimize burden on site staff and maximize staff availability, the evaluation team works with POD site directors to determine the most convenient times to convene the interviews. The evaluation team limits the interviews to approximately one hour so that the data collection imposes only a modest burden on respondents.

To facilitate a smooth interview process and improve the completeness of the data collected, the POD evaluation team mails an information packet to the site director and partner agency contacts containing the final site visit and interview schedule. The evaluation team sends out the packet about two weeks before the site visit takes place, and it contains the lead site visitor’s contact information so the respondents can reach the visitors in the event of a schedule change, or other issues that might arise before the interviews. The evaluation team also sends email reminders to site directors and partner agencies several days before the site visit confirming their arrival day and time. Providing the local sites with adequate information ahead of time in a professional manner helps build rapport; facilitates a more fluid interview process; and establishes that interviewees are available and responsive.

Data reliability. Interviewers use an interview guide, based on the interview topic list provided in Attachment C, to conduct the semi-structured staff interviews. The interviewer takes notes and obtain permission to record each interview. The POD evaluation team uses separate discussion guides for each potential respondent type (for example, VR administrator, POD work incentives counselor, VR agency staff, or staff from other state implementation partners) so they do not ask respondents about activities or issues that do not apply to them. Rather than ask respondents to answer detailed questions about specific operations such as provision of benefits counseling and tracking the benefit offset, the evaluation team reviews program documentation and memos ahead of the interviews to minimize burden and supplement information provided by respondents to the interviews. During the in-person site visits, the evaluator also conducts direct observations of site operations using a standard template (Attachment C) to systematically document the process of enrollment and provision of services and assess implementation fidelity. The evaluation team uses this information to supplement information gathered from the interviews with an objective assessment of specific operations such as provision of benefits counseling and tracking the benefit offset. After completion of all interviews conducted for a particular POD site, the evaluator develops a summary of all of the information collected during the site visit.

B.3.4. Semi-structured interviews with beneficiaries

Response rates. Interviewers contact selected study subjects to explain the purpose of the interview and schedule a convenient time for a telephone interview. To facilitate high response rates, interviewers stress the private nature of the interview; the importance of this information for future program improvements; and interviewer flexibility in selecting a time that meets the needs of the respondents. As noted previously (Section B.1.4), in the event of a refusal, the POD evaluation team selects new respondents using the same criteria used to select the initial pool of potential respondents purposively. To facilitate cooperation, the evaluation team offers potential respondents a $25 incentive, such as a card for a national retailer.

Data reliability. To improve the quality and reliability of the data collected, the POD evaluation team undertakes the following steps. First, senior members of the evaluation team test the interview protocol, and refine it to establish: that they can cover key topics in the designated time; that questions are clear and unambiguous; that they cover all key topics; and that they ask appropriate topics of key respondent groups (for example, those in the two treatment arms and other sub-groups of interest such as low earning offset users). Second, the evaluation team selects interviewers who are well versed in conducting telephone interviews and train them in the use of the protocol. Training focuses on: ensuring the interviewers understand the interview protocols fully; are able to adapt the protocol based on existing information available about the respondent; are able to clarify questions and probe for additional details to gather comprehensive information on all topics; and fully understand how to document data from the interviews in a systematic and consistent fashion using a standardized template and coding scheme. The coding scheme includes constructs that allow the POD evaluation team to generate consistent measures of implementation fidelity for the process study. Third, before the interview, interviewers use data from each respondent’s baseline and follow-up surveys to assemble a preliminary profile of the respondent that can guide the interview and allow for more time for efficient follow-up on key topics. Fourth, senior members of the POD evaluation team monitor each interviewer’s initial telephone calls to verify that they are using the correct interview techniques and following the interview protocol with fidelity. Finally, a senior member of the evaluation team reads a subset of the transcribe interview notes to check that the interviewer collected and recorded the relevant data. Interviewers conduct follow-up telephone calls with respondents to collect missing data as necessary, and they add to the interview summaries notes on key themes and cross-cutting findings that appear in interviews with multiple subjects.

B.3.5 Implementation data collection

Abt Associates takes several steps to maximize response to the monthly earnings and IRWE reporting form and annual reporting form. The implementation team gives POD treatment subjects detailed instructions for completing these forms in materials provided after random assignment. Work incentives counselors discuss these instructions with treatment subjects and will remind treatment subjects to report earnings and IRWEs. Abt Associates provides quarterly reminders by mail and monthly reminders via email and text to POD treatment subjects who give consent for the project to email or text them. Abt Associates also call treatment subjects who previously reported earnings above the POD threshold to remind them to report if they do not report their monthly earnings timely. POD treatment subjects have the option of submitting their monthly and annual earnings and IRWE information on paper or online.

B.4. Tests of procedures

B.4.1 Recruitment materials and baseline survey

Pretesting of baseline survey. The POD evaluation team conducted a pretest of the baseline questionnaire with five respondents and then revised the instrument based on findings from the pretest. The pretest provided an accurate estimate of respondent burden as required by OMB, and the evaluation team also assessed flow and respondent comprehension.

Refining and testing random assignment procedure. As noted previously (Sections B.1.1 and B.3.1), random assignment is at the individual level, and the POD evaluation team used stratified, batch random assignment to improve the balance of the treatment and control groups. Before starting the recruitment phase, the evaluation team evaluated the usefulness of this randomization approach with fabricated data to (1) verify that it balances appropriately on the stratifying variables using the anticipated batch sizes; and (2) assess the risk of imbalance on other variables. Based on this testing, they developed a random assignment procedure and reported on-going results in recruitment reports to SSA to assess baseline equivalence. The reports showed comparisons of treatment and control group subjects.

Experiments during the pilot recruitment period. The recruitment experiments conducted during the pilot period tested how the yield rates of eligible volunteers vary in response to alternative outreach strategies. These experiments included all beneficiaries selected for recruitment targeting in the pilot areas. The POD evaluation team implemented the experiments using a factorial design to test variants of materials, follow-up activities, and incentives.

Following the pilot, SSA expanded the catchment area to increase the number of POD subjects into additional counties in Texas. As documented in Wittenburg et al. (2018), the expansion included 13 new counties in Texas, which increased outreach to 420,000 beneficiaries in total over the full demonstration period.

B.4.2. Follow up surveys

The POD evaluation team conducted a pretest of the follow up survey instrument with five respondents and then revised the instrument based on findings from the pretest. The pretest provided an accurate estimate of respondent burden as required by OMB. In addition, the evaluation team also assessed flow and respondent comprehension by debriefing each respondent to determine if any words or questions were difficult to understand and answer. Like actual study subjects, the evaluation team gave participants in the pretest of the follow-up survey an incentive for their time.

B.4.3. Qualitative data from POD implementation and operations staff

The POD evaluation team bases site visit protocols on those used for related evaluations, and it uses the first site visit of each round as a pilot to test interview protocols and the assessment tool used to assess implementation quality. During this pilot site visit, the process study lead, who also serves as a state liaison, conducts cognitive tests of the semi-structured interview protocol to establish that respondents interpret questions as intended and have the information necessary to answer the questions. The evaluation team assesses the qualitative interview questions by conducting up to nine cognitive tests of the interview protocols with a range of program staff. Interviewers ask respondents to explain how they arrived at their answers and probe to determine how respondents arrived at their answers and whether any items were difficult to answer. Interviewers review their notes following the interviews and summarize issues or potential problems in a brief memo. At the conclusion of the pilot site visit, the evaluation team makes corrections when necessary to improve the data collection tools and procedures. Before the start of data collection, the process study lead trains the other interviewers on how to conduct cognitive interviews. Senior research staff also assess the site visit agenda, including the data collection activities they conduct, and how they structure these activities, to check that they can feasibly conduct them as part of the site visits and yield the desired information.

B.4.4. Semi-structured interviews with beneficiaries

The POD evaluation team designs and tests the protocol for semi-structured telephone interviews with beneficiaries so that these interviews yield high-quality data that provide richer detail (compared with data from the program and survey data). Senior members of the evaluation team test the protocols during the first couple of interviews and subsequently refine them as needed. During this process, the evaluation team pays careful attention to whether the protocol covers key topics of interest to the evaluation; whether the protocol covers all these topics in the designated time; and whether they worded questions and probes clearly, made them easy to understand, and optimally sequenced them to solicit responses with sufficient levels of detail. The evaluation team also reviews the protocols to determine that they ask appropriate topics of the respondent given their group assignment and other characteristics (for example, those in the two treatment arms and other sub-groups of interest such as low earning offset users). In light of initial interviews, the evaluation team revises and streamlines the interview protocol and the related templates for data recording.

B.4.5 Implementation data collection

SSA adapted the monthly earnings and IRWE reporting form and annual form, which the implementation team uses for the implementation data collection from standard procedures and protocols used to collect earnings and IRWE documentation from SSDI beneficiaries for SSDI benefit calculation. The implementation team determined that they would not need additional pre-testing to pre-test these data collection forms.

B.5. Statistical agency contact on technical and statistical issues

We list the evaluation team members providing input on technical and statistical issues discussed in this information clearance request in Exhibit B.3, and we list the implementation team members consulted on technical and statistical issues related to data collection in Exhibit B.4.

Exhibit B.3. Individuals consulted on the study design

Name	Phone number	Affiliation
Dr. David Wittenburg	609-945-3362	Mathematica
Dr. Kenneth Fortson	510-830-3711	Mathematica
Noelle Denny-Brown	617-301-8987	Mathematica

Martha Kovac	609-275-2331	Mathematica
Dr. David Stapleton	202-484-4224	Mathematica
Dr. Heinrich Hock	202-250-3557	Mathematica
Dr. Debra Wright	703-504-9480	Insight Policy Research

Exhibit B.4. Individuals consulted on the implementation data collection

Name	Phone number	Affiliation
Ms. Sarah Gibson	617-520-2810	Abt Associates
Mr. Eric Friedman	617-520-2876	Abt Associates
Mr. Brian Sokol	617-349-2532	Abt Associates
Ms. Susan O’Mara	757-620-5451	Virginia Commonwealth University
Ms. Michelle Wood	301-634-1777	Abt Associates

REFERENCES

American Association for Public Opinion Research (AAPOR). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Ninth edition. Oakbrook Terrace, IL: AAPOR, 2016.

Allison, Paul D. *Missing Data (Quantitative Applications in the Social Sciences, Vol. 136)*. Thousand Oaks, CA: Sage Publications, 2001.

Bell, Stephen, David C. Stapleton, Daniel GU bits, David Wittenburg, Michelle Derr, David Greenberg, Arkadipta Ghosh, and Sara Ansell. "BOND Implementation and Evaluation: Evaluation Analysis Plan." Cambridge, MA: Abt Associates, and Washington, DC: Mathematica Policy Research, 2011.

Biggs, David, Barry de Ville, and Ed Suen. "A Method of Choosing Multiway Partitions for Classification and Decision Trees." *Journal of Applied Statistics*, vol. 18, no. 1, 1991, pp. 49–62.

Bollinger, Christopher R., and Barry T. Hirsch. "Match Bias in the Earnings Imputations in Current Population Survey: The Case of Imperfect Matching." *Journal of Labor Economics*, vol. 24, no. 3, July 2006, pp. 483–520.

Chromy, James R. "Sequential Sample Selection Methods." *Proceedings of the American Statistical Association, Survey Research Methods Section.* Washington, DC: American Statistical Association, 1979, pp. 401-406.

Cameron, A. Collin, and Pravin K. Trivedi. *Microeconometrics: Methods and Applications.* New York: Cambridge University Press, 2005.

Damschroder, Laura J, David C. Aron, Rosalind E. Keith, Susan R. Kirsh, Jeffery A. Alexander, and Julie C. Lowery. "Fostering Implementation of Health Services Research Findings into Practice: A Consolidated Framework for Advancing Implementation Science." *Implementation Science,* vol. 4, 2009, pp. 50-65.

Derr, Michelle, Denise Hoffman, Jillian Berk, Ann Person, David Stapleton, Sarah Croake, Christopher Jones, and Jonathan McCay. "BOND Implementation and Evaluation Process Study Report." Washington, DC: Mathematica Policy Research, February 12, 2015.

Fraker, Thomas, Arif Mamun, Todd Honeycutt, Allison Thompkins, and Erin Jacobs Valentine. "Final Report on the Youth Transition Demonstration Evaluation." Washington, DC: Mathematica Policy Research, December 29, 2014.

Imbens, Guido, and Donald Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press, 2015.

Potter, Francis J. “A Study of Procedures to Identify and Trim Extreme Sampling Weights.” In *Proceedings of the American Statistical Association, Section on Survey Research Methods.* Alexandria, VA: American Statistical Association, 1990, pp. 225-230.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. *Model-Assisted Survey Sampling*. New York: Springer-Verlag, 1992.

Schochet, Peter Z. “Statistical Theory for the RCT-YES Software: Design-Based Causal Inference for RCTs.” Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Research, 2016.Westfall, Peter H., and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* New York: John Wiley, 1993.

Wittenburg, David, Kenneth Fortson, David Stapleton, Noelle Denny-Brown, Rosalind Keith, David R. Mann, Heinrich Hock, and Heather Gordon. “Promoting Opportunity Demonstration (POD): Design Report.” Report submitted to the Social Security Administration. Washington, DC: Mathematica Policy Research, May 18, 2018. Available at https://www.ssa.gov/disabilityresearch/documents/POD%205%202_Evaluation%20Design%20Report_10-4-2018.pdf (accessed December 24, 2018).

Wooldridge, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2010.