Contract Number:
SS00-11-31390

Mathematica Reference Number:
06975.300

Submitted to:
Social Security Administration
Office of Retirement and Disability Policy
Office of Program Development and Research
496 Operations
Pole 3-C-25
Baltimore, MD 21235

Project Officer: John T. Jones

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director: David Wittenburg

Work Incentive Simplification Pilot (WISP):

Recommendations of the Technical Advisory Panel Regarding the Evaluation Design

Final Report

April 25, 2012

David Wittenburg
David R. Mann
David C. Stapleton

MATHEMATICA
Policy Research

CONTENTS

EXHIBITS

ACRONYMS

BOND	Benefit Offset National Demonstration
CDRs	Continuing Disability Reviews
CMS	Centers for Medicare & Medicaid Services
EN	Employment Network
EPE	Extended Period of Eligibility
EXR	Expedited Reinstatement
GP	Grace Period
IRP	Initial Reinstatement Period
IRWE	Impairment-Related Work Expenses
MBI	Medicaid Buy-In
OACT	Office of the Actuary
SGA	Substantial Gainful Activity
SNAP	Supplemental Nutrition Assistance Program
SSA	Social Security Administration
SSDI	Social Security Disability Insurance
SSI	Supplemental Security Income
SVRA	State Vocational Rehabilitation Agency
TAP	Technical Advisory Panel
TTW	Ticket to Work
TWP	Trial Work Period
VR	Vocational Rehabilitation
WIPA	Work Incentives Planning and Assistance
WISP	Work Incentive Simplification Pilot

EXECUTIVE SUMMARY

The Social Security Administration (SSA) is in the early stages of designing the Work Incentive Simplification Pilot (WISP), a demonstration to test major simplifications to the Social Security Disability Insurance work incentives. WISP would replace the complex SSDI rules associated with returning to work with a simplified process designed to be easier for beneficiaries to understand and less costly for SSA to administer. Under WISP, benefits could be suspended for work activity above "Substantial Gainful Activity", but not terminated for work, and Medicare coverage would continue indefinitely to all SSDI beneficiaries unless their benefits are terminated for some other reason.[1] WISP could cut SSA administrative costs and improve several beneficiary outcomes related to employment. If successful, WISP's provisions could eventually replace SSDI's existing system for reporting and developing earnings.

To obtain recommendations for the WISP evaluation design, SSA contracted with Mathematica Policy Research to develop and administer a technical advisory panel (TAP). Mathematica composed the TAP with seven members from the academic, nonprofit, and governmental fields with a wide range of evaluation and policy experience. Mathematica provided the TAP with background information on WISP in a briefing document. The TAP provided input on evaluation design options during an all-day meeting and in written input, facilitated by Mathematica staff with input from SSA.

As a starting point for the TAP's deliberations, SSA outlined five general research questions for the evaluation design:

1. How do the various WISP provisions affect the work behavior and benefit payments of SSDI beneficiaries?
2. How do the various WISP provisions affect Medicare costs, utilization rates, and use of private insurance?
3. How much does an automated system of earnings reporting and benefit determination used by beneficiaries affect SSDI workloads?
4. What is the impact of WISP on administrative costs and improper payments?
5. What is the potential for induced entry and how would SSA measure it?

The TAP provided several recommendations for an evaluation of WISP to support decisions about nationwide implementation (**Exhibit ES-1)**. The TAP acknowledged that the five research questions sufficiently cover the important policy issues. They recommended SSA consider effects on several additional outcomes, such as use of work supports, work attempts, health, satisfaction with services, use of other health insurance, and consumption. They suggested the use of both administrative data and follow-up surveys to evaluate these outcomes.

The TAP recommended a random assignment design be applied to a nationally representative beneficiary sample to ensure rigorous estimates for national implementation. Random assignment could be at the site and/or individual level, but the best option depends on operational

[1] In 2012, SGA is generally defined as any activity that is comparable to unsubsidized paid work for monthly wages of at least $1,010 for non-blind individuals or $1,690 for blind individuals.

considerations for WISP that remain to be determined. They noted that beneficiary understanding of WISP is important for generalizing findings to the national level and it would be best for SSA to lead the outreach effort. They also noted that non-experimental approaches are appropriate for evaluating some outcomes, particularly for small subgroups.

The TAP also recommended that the evaluation last at least five years. WISP's impacts on SSA administrative procedures should be reported after the first two years of the project. Beneficiary outcomes, especially employment and benefit receipt, should not be reported before the third year. The timing of assessments and reporting is critical because impacts on several key outcomes take longer to emerge and premature reporting can be misleading.

Finally, the TAP had detailed suggestions in several other areas, including recommendations for dissemination, sampling and data collection. The TAP's recommendations should be a useful starting point for the evaluation as SSA continues to develop its plans for the full demonstration.

ES Table 1. Summary of TAP Recommendations

Research Questions and Outcomes	• SSA five questions capture the primary outcomes, although methodological challenges exist in measuring induced entry. • Comprehensive evaluation would consider effects on several additional outcomes, such as use of work supports, work attempts, health, satisfaction with services, use of other health insurance, and consumption.
Dissemination of Findings	• Outcomes should be reported at intervals based on anticipated impacts. • WISP evaluation should last at least five years • Interim and final reports should inform policymakers and other stakeholders.
Evaluation design	• Evaluation should use experimental design. • A non-experimental approach could be used to assess findings for supplemental outcomes, particularly when samples are limited. • Random assignment could be at the site or individual level, or a hybrid, but the best option depends on operational considerations that remain to be determined. • Beneficiary understanding of WISP is essential to the external validity of the evaluation, and it would be best for SSA to lead the outreach effort.
Sampling and Subgroups	• The random sample should be nationally representative and sufficiently large to detect policy relevant impacts. • Potential subgroups to be oversampled include those who have achieved certain work milestones, such as the Trial Work Period (TWP). • A sufficiently large sample of new SSDI beneficiaries is critical for projecting long-term WISP impacts.
Data	• Use all available administrative data—including those from SSA and outside agencies (e.g., for Medicare and Medicaid)—to measure outcomes when feasible. • Surveys would be needed for some outcomes, such as health. • Fidelity of WISP program services should be tracked closely using a combination of quantitative and qualitative data

ACKNOWLEDGEMENTS

Many people made significant contributions to this report and project. We are grateful to the seven technical advisory panel members—Burt Barnow, Kelly Buckland, Henry Claypool, Carolyn Heinrich, James Smith, Jeffery Smith, and Susan Webb—whose thoughtful discussion and recommendations are the foundation of this report. Several staff at SSA provided valuable input throughout the project, which was very helpful in facilitating the TAP's recommendations and improving the content of this report. John Jones, the project officer, provided technical and administrative oversight throughout the project, including assistance in responding to all technical questions about WISP and organizing timely and very helpful feedback on this report. Renee Ferguson and Scott Muller also provided comments and support throughout the project that were especially helpful in clarifying the WISP's intervention parameters for consideration by the TAP. Finally, Cindy Duzan, Kristine Erwin-Tribbitt, Susan Kalasunas, Sharon Kuczak, Susan Wilchke, and Robert Weathers provided written comments on this report and helpful feedback during the TAP meeting and/or project kickoff meeting. At Mathematica, we would like to thank Daniel Weaver and Cindy McClure, who provided excellent logistic and administrative support both on this report and throughout the project.

David Wittenburg
David R. Mann
David C. Stapleton

I. INTRODUCTION

A. Work Incentive Simplification Pilot (WISP) Overview

Social Security Disability Insurance (SSDI), which is administered by the Social Security Administration (SSA), is the nation's primary earnings-replacement program for workers who become unable to work. In making SSDI disability eligibility determinations, SSA assesses whether a person (1) has a medically determinable disability expected to last at least 12 months or result in death and (2) is unable to engage in substantial gainful activity (SGA), defined in essence as the ability to earn a minimum monthly amount. SSA defines SGA as the performance of significant physical and/or mental activities in work for pay or profit, or in work of a type generally performed for pay or profit.[2] In 2012, SGA is generally defined as any activity that is comparable to unsubsidized paid work for monthly wages of at least $1,010 for non-blind individuals or $1,690 for blind individuals. The SGA amount is used in initial SSDI eligibility assessments and in ongoing eligibility assessments for those who attempt to return to work. SSDI beneficiaries are automatically entitled to Medicare health coverage after 24 months of entitlement to SSDI.

SSDI beneficiaries who return to work must report their earnings to SSA in a way that is complex to administrate and is likely not well understood by beneficiaries. For SSA, there are substantial administrative burdens associated with collecting earnings information on a monthly basis from beneficiaries. If there is an error in reporting information to SSA, beneficiaries could receive an overpayment or underpayment in their SSDI benefit check (Livermore 2003). For beneficiaries, a concern is that the program rules and administrative processes might limit beneficiaries' interest in returning to work (Stapleton et al. 2006).

SSA is in the early stages of designing the Work Incentive Simplification Pilot (WISP), a demonstration to test major simplifications to the SSDI work incentives. WISP would replace the complex SSDI rules associated with returning to work with a simplified process that would be easier for beneficiaries to understand and less costly for SSA to administer. WISP could cut SSA administrative costs and improve several beneficiary outcomes related to employment. If successful, WISP's provisions could eventually replace SSDI's existing system for reporting and developing earnings.

The goal of this project is to develop evaluation design options to test the WISP as a national demonstration. SSA has designed the basic parameters of the WISP intervention, including new administrative procedures and work incentives for beneficiaries. However, some aspects of the WISP intervention still need to be specified, including the overall scope of the intervention and the specific automated procedure to process benefits. Following the design phase, SSA might conduct a pre-test of the WISP and then move to a national demonstration.[3] For WISP to proceed to the design phase, Congress would first need to renew SSA's demonstration authority.[4] This could be

[2] The SGA amount includes the total (unsubsidized) earnings net of allowable expenses that SSA classifies under impairment-related work expenses (IRWE).

[3] According to the original statement of work, SSA may pre-test the demonstration. The pre-test would examine the feasibility of implementing the design, acquire initial impact estimates, and provide guidance on how best to implement a national demonstration.

[4] SSA's demonstration authority expired in 2005 and would have to be updated for SSA to operate WISP.

problematic because, even though WISP is largely advantageous to beneficiaries, it has one feature that could temporarily reduce their incomes, as will be seen in Section II.[5]

B. Technical Advisory Panel (TAP) for WISP

SSA contracted with Mathematica Policy Research to develop and administer a technical advisory panel (TAP) that would make recommendations for the WISP evaluation design. The TAP was composed of members from the academic, nonprofit, and governmental fields with a wide range of evaluation and policy experience (Exhibit I.1). Three TAP members had a research background with experience in designing and evaluating interventions for SSA and other government agencies. The remaining four TAP members had extensive policy and program knowledge based on their work in non-profit and government agencies. These four TAP members had expertise in an array of policy or programmatic issues that might influence WISP, such as SSA administrative policies, health care reform, rehabilitation support programs (e.g., Vocational Rehabilitation (VR)) and consumer perspectives. The combined experience of the TAP provided SSA with insights from a diverse group of experts using a multidisciplinary approach.

The TAP provided input on evaluation design options during an all-day meeting and in written input, which were both facilitated by Mathematica staff with input from SSA. The TAP was initially sent briefing materials, which outlined the key parameters of the WISP intervention and issues for consideration. Mathematica staff then facilitated an all-day meeting with the TAP and SSA staff on February 10, 2012. During the meeting, the TAP provided input on several evaluation design options, including the types of outcomes that should be tracked; the timeline for tracking outcomes; the evaluation design for examining program impacts; strategies for sampling beneficiaries and subgroups; potential data sources; and other intervention features that could have implications for the evaluation (e.g., how to inform beneficiaries and service providers about WISP). Following the meeting, the TAP provided written input using a "TAP Input Form."[6]

This report summarizes the TAP's recommendations. The TAP reached a consensus recommendation on several topics, including the need to use random assignment to evaluate impacts on a nationally representative sample. However, in other areas, such whether to use individual or site level random assignment and whether to include specific subgroups, there was not a consensus and/or enough information on the WISP operational procedures to make a full recommendation. In these areas, we summarize the TAP's input and discuss options for SSA to proceed in considering a final evaluation design.

[5] Allowing demonstration treatment subjects to opt out of WISP would address this problem. However, WISP is not envisioned as a voluntary alternative to current rules, so allowing treatment subjects to opt out would undermine the value of the demonstration for predicting outcomes under a national program.

[6] Appendix A includes the written summary of the TAP Input Forms. Six of the seven TAP members returned the forms. The report includes the feedback of all TAP members during the TAP meeting and the written input from the six forms.

Exhibit I.1. TAP Members for the WISP Project

	Affiliation	Relevant Experience
Academic		
Burt Barnow	Amsterdam Professor of Public Service, George Washington University	Evaluation design methodologies, SSA policy, previous SSA demonstrations, and employment services.
Carolyn Heinrich	Sid Richardson Professor of Public Affairs, University of Texas at Austin	Evaluation design methodologies and employment services
Jeffrey Smith	Professor of Economics, University of Michigan	Evaluation design methodologies, SSA policy and employment services.
Nonprofit		
Kelly Buckland	Executive Director, National Council on Independent Living	Disability policy, SSA policy, rehabilitation services, and consumer perspectives
Susan Webb	Vice President and Director, Employment Services/Business Development, Arizona Bridge to Independent Living	Disability policy, SSA policy, previous SSA demonstrations, rehabilitation services, and consumer perspectives
Government		
Henry Claypool	Director, Office of Disability, U.S. Department of Health and Human Services	Disability policy, previous SSA demonstrations, health policy, and consumer perspectives
James Smith	Budget and Policy Manager, Vermont Division of Vocational Rehabilitation	Disability policy, previous SSA demonstrations, rehabilitation administrative experience, and consumer perspectives

C. Organization of This Report

The remainder of this report includes four sections. Sections II and III include background information on WISP to inform the TAP's recommendations, as initially presented in TAP briefing materials. Section II provides information on SSDI's current work incentives and on the WISP intervention. Section III highlights the initial SSA research questions for the evaluation, current beneficiary work outcomes, and initial projections of potential WISP effects. Section IV provides a detailed summary of the TAP's recommendations. Section V provides concluding remarks, including highlights of the major considerations from the TAP's deliberations as SSA moves forward with a full evaluation design.

Throughout the report, we use the terms WISP "treatment" and "control" to refer to research groups in a future evaluation. A WISP treatment subject refers to an SSDI beneficiary who is part of the demonstration and receives the WISP intervention. A WISP control subject refers to an SSDI beneficiary who is part of the demonstration, but receives current SSDI work incentives (i.e., these beneficiaries do not receive any WISP services).

II. SSDI PROGRAM RULES AND THE WISP INNOVATION

The proposed WISP intervention would modify several existing SSDI program rules. These modifications have important implications for the evaluation design, especially the outcomes that could be emphasized in the eventual WISP evaluation. The remainder of this section provides information on SSDI work incentives and the proposed WISP changes. We highlight eight features of the existing SSDI program that will be modified under WISP. We also summarize how these changes could influence SSDI beneficiaries who participate in other programs.

A. SSDI Program Rules

The WISP intervention does not change the SSDI application process, though it makes several changes to the on-going eligibility requirements. These changes have important implications for SSDI work incentive provisions, on-going SSA administrative processes to assess eligibility, and other SSA and non-SSA programs that directly or indirectly affect SSDI beneficiaries.

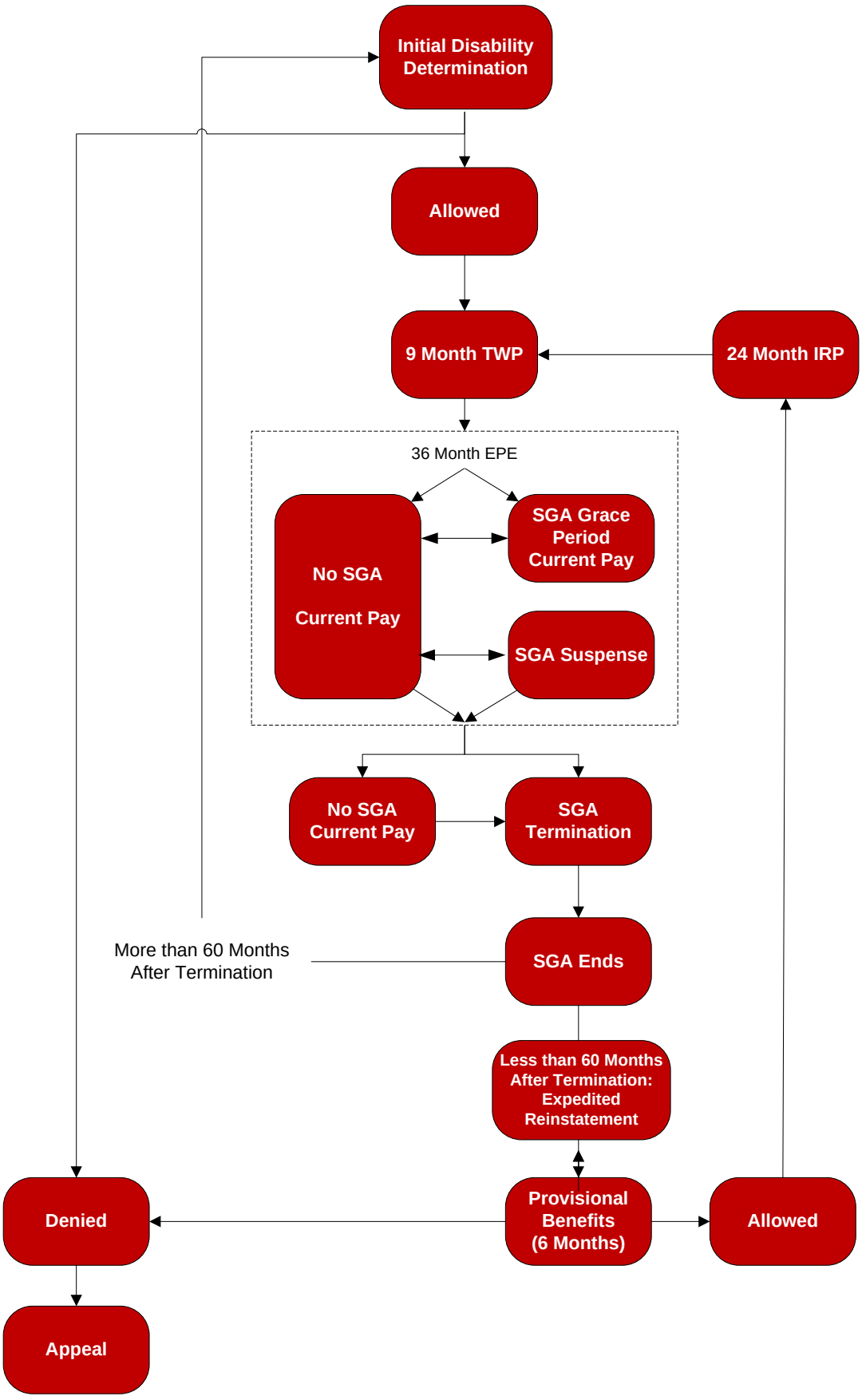
1. Trial Work Period (TWP), Extended Period of Eligibility (EPE), and Expedited Reinstated (EXR)

The complexity of the ongoing SSDI work incentive rules is illustrated in Exhibit II.1. For beneficiaries who have been in the program at least 12 months, SSDI includes work rules that affect how much a beneficiary can earn and retain benefits. There are three critical time frames over which a beneficiary's benefits can be affected by work and eventually terminated, though some can apply for expedited benefit reinstatement:

1. **The trial work period (TWP)** tests an SSDI beneficiary's ability to work without affecting benefits. The TWP may begin no earlier than the later of the month of filing or the month of entitlement to SSDI, and ends when the beneficiary has completed 9 trial work months during any window of 60 consecutive months. In 2012, earnings of $720 or more per month or 80 hours of self-employment constitute trial work. During these months, the beneficiary is entitled to full benefits even if engaged in SGA. The TWP ends after the ninth month of trial work.

2. **The extended period of eligibility (EPE)** begins immediately after completion of the TWP and lasts until benefits are terminated. During the first 36 months of the EPE, called the re-entitlement period, benefits are suspended—that is, not due to the beneficiary—during any month if the beneficiary engages in SGA, except that each beneficiary has three Grace Period (GP) months, which occur with the first month of SGA in the EPE, in which full benefits are due even if the beneficiary engages in SGA. If SGA ends before the end of the re-entitlement period, the beneficiary is again entitled to full benefits, provided that benefits have not been terminated for some other reason, including medical recovery.

Exhibit II.1. Overview of the Current SSDI Provisions



3. **Termination and expedited reinstatement (EXR).** Benefits are terminated with the first month of SGA-level work after the EPE re-entitlement period, or as soon thereafter as the GP months are used up. After termination, benefits do not resume simply because SGA ends. If SGA ends within 60 months of termination, the beneficiary may apply for expedited reinstatement of benefits, and can receive up to six months of provisional benefits while SSA reviews the application. If the beneficiary reaches SGA after 60 months, the beneficiary may reapply under the same process as for first-time applicants. In either case, the beneficiary must go through a reapplication and redetermination process to obtain eligibility again. Former beneficiaries who successfully reapply for benefits during EXR enter an initial reinstatement period (IRP). As with the EPE, during the IRP benefits are paid only when the beneficiary does not engage in SGA. Beneficiaries in IRP also receive Medicare coverage regardless of earnings. The IRP ends only after 24 months of benefits are paid, and it can last indefinitely if the beneficiary continues to engage in SGA.

The cash cliff that exists after the GP months under existing SSDI program rules gives beneficiaries a strong incentive to keep earnings below the SGA level, especially if the beneficiary is unable to earn well above the SGA amount. To illustrate, consider a non-blind beneficiary who receives a monthly benefit of $1,000. If after completing the TWP and GP, the beneficiary earns $1,020 a month, he or she is not entitled to benefits, resulting in a total monthly income from earnings and benefits of $1,020. If the beneficiary were to instead earn $20 less, he or she would receive an SSDI payment of $1,000 and accrue a significantly higher total income of $2,000. Intentionally keeping earnings just below the SGA amount in order to avoid benefit suspension or termination is sometimes referred to as earnings "parking." Exploiting a 1999 change in the SGA level from $500 to $700, Schimmel et al. (2011) estimated that in a typical month between 2002 and 2006, 0.2 to 0.4 percent of all SSDI beneficiaries were parked below the SGA level. Further, the authors cite several reasons why these estimates might understate the extent of parking. While this estimate represents a small portion of the overall SSDI beneficiary population, it includes the beneficiaries who would most likely change their behavior in response to WISP's provisions.

2. Continuing Disability Reviews (CDRs) and Earnings Evaluation

Any employment disincentive due to confusion about the work incentives is magnified by problems related to their administration by SSA. There are two key administrative aspects that are especially important to the administration of benefits under current law that would change under WISP:

4. **Continuing disability reviews (CDRs).** SSA conducts periodic CDRs based on expected medical improvement or on evidence of work. There are two types of CDRs: medical and work. Medical CDRs are used to make contact with beneficiaries to update their personal and medical information and to determine if they still meet the SSA definition of disability. SSA keeps a record for when a medical CDR should occur, which is based on the beneficiary's expected medical improvement. Those with a high probability of medical improvement ("medical improvement expected") are supposed to have a medical CDR within three years, and many are supposed to occur in less than two years. Other beneficiaries with lower probabilities of expected medical improvement have longer medical CDR windows (for example, seven years). Due to SSA's limited operational resources, individuals in benefit suspension usually do not receive timely medical CDRs. The work CDR process is an important SSA administrative process for ensuring accurate benefit changes based on the beneficiary's earnings. Beneficiaries are

required to report earnings changes to their local field office. While work CDRs can be prompted by several events, most are generated by SSA's CDR Enforcement Operation (enforcement operation).[7] This process involves periodic data matches between SSA's administrative databases and Internal Revenue Service earnings data. A work CDR is also performed at the end of the TWP as a condition for providing the EPE benefit.

5. **Evaluation of earnings when due versus when paid.** SSDI beneficiaries must provide information on earnings from wages and self-employment for the months when earned rather than the amount for the month when paid. This provision can complicate TWP and SGA determinations because beneficiaries' pay stubs and other wage records are dated when payment occurs. Hence, paychecks and other transactional evidence that might be generated in a later month must be translated into evidence of SGA during the month the work was performed.

The administrative burden of tracking earnings is significant to SSA staff and SSDI beneficiaries. SSA staff must explain complicated work incentive policies and reporting responsibilities to all SSDI beneficiaries and applicants. When a work activity report is received, SSA must evaluate each month of work activity by contacting beneficiaries (and sometimes their employer) to resolve any questions about the work incentive provisions. The time needed to effectively inform each claimant or beneficiary of SSA's work incentives is significant, and the complicated nature of work incentive policies can inadvertently discourage beneficiaries from returning to work. Furthermore, determining a beneficiary's work activity can require multiple follow-up discussions with the beneficiary or employer, which can result in lengthy case processing times and overpayments.

Competing demands on field office staff has, at times, led to long backlogs in work CDRs and problems in processing benefits. These backlogs often result in overpayments, where the beneficiary enters the EPE, engages in SGA, completes the GP months, and yet continues to receive a monthly benefit check (Livermore 2003). Eventually SSA conducts the work CDR, notifies the beneficiary of an overpayment, and demands repayment. This can lead to an appeal and perhaps a gradual repayment plan. Underpayments also occur, although less often. The overall result is substantial additional uncertainty for the beneficiary plus a substantial administrative burden for both the beneficiary and SSA. Some overpayments are not repaid in full—another cost to SSA, but a gain to the beneficiary. Overpayments, as well as underpayments, may also occur because of errors in the process.

Given the complexity of the work incentives and their administrative challenges, it is not hard to imagine that some beneficiaries would avoid earning enough to complete the TWP or earn above SGA, especially if their potential income gain, once they lose their benefits, is only modest (Stapleton et al. 2006). Many beneficiaries might think it is better to keep earnings low and avoid the hassle and uncertainty.

[7] A work CDR can also be generated from a third party report, such as earnings reported by a state Vocational Rehabilitation Agency.

3. Medicare, Ticket to Work (TTW) and Other Provisions

SSDI beneficiaries are also eligible for two other supports that would be affected by or interact with WISP:

6. **Medicare continuation.** As noted in Section I, SSDI beneficiaries are automatically entitled to Medicare after 24 months of entitlement to SSDI. However, their Medicare eligibility can change based on their SSDI eligibility status. Specifically, once the TWP is completed, if eligibility to cash benefits terminates due to work activity, Medicare eligibility continues for 78 months after the first month of SGA occurring after the 15th month of the EPE, provided that the beneficiary has not medically improved. After this time frame, Medicare coverage is only available by paying a monthly premium.
7. **Ticket to Work (TTW).** SSDI beneficiaries interested in returning to work can use supports from the TTW program. Each SSDI beneficiary receives a "ticket" that can be used to purchase employment services from a state vocational rehabilitation agency (SVRA) and/or another qualified provider, known as an Employment Network (EN). Providers are not obligated to accept the ticket. Because payments to providers are based on the beneficiary's success in earning enough to give up his or her benefits, it is only in the provider's interest to accept tickets from beneficiaries who are likely to work and earn enough to trigger TTW payments.[8]

SSA also funds Work Incentives Planning and Assistance (WIPA) grants to provide job placement, benefits planning, and career development to beneficiaries, but SSA's authority to offer these grants expires on June 30, 2012.[9] If the WIPAs are re-authorized, they could play an important role in informing WISP treatment subjects about WISP's work rules. If not, SSA would need to inform WISP treatment subjects about the new rules in some other way. [10]

B. WISP Parameters

WISP aims to simplify SSDI's work incentive structure and administrative procedures. These simplifications are meant to encourage work and ease administrative burden. All WISP subjects would be assigned to a treatment or control group. SSDI beneficiaries would not be asked to volunteer for WISP, because a national change would not allow beneficiaries to opt out of the new rules. Below, we highlight how WISP changes the SSDI program rules described in Section II.A and provide a side by side comparison of the WISP rules and current rules in a summary table.

---

[8] ENs can receive TTW "outcome" payments when a Ticket holder enters zero cash benefit status because of work. In addition, under the Milestone-Outcome payment system elected by most ENs, the EN can receive a limited number of milestone payments based on progression of the Ticket holder's earnings toward a level at which cash payments would be zero. A description of both payment options is available at https://yourtickettowork.com/web/ttw/en-payments-options (accessed April 12, 2012).

[9] SSA Commissioner Astrue alerted Congress on March 9, 2012 that SSA will shut down the WIPA program on June 30, 2012 unless Congress acts.

[10] While there are no plans to provide special outreach or benefits counseling services to WISP subjects, WIPA organizations, if their funding is renewed, would need to have protocols for identifying and supporting WISP subjects who contact WIPAs for supports.

1. TWP, EPE, and EXR Changes

Although initial disability determination and Medicare eligibility would remain the same, WISP's provisions would substantially alter the work incentives SSDI beneficiaries currently face. WISP would eliminate many positions along SSDI's current path of work incentives.

1. **Elimination of the TWP.** WISP would eliminate the TWP. The current work incentive rules would be simplified so that beneficiaries would receive their full SSDI benefit during months they do not engage in SGA and no income benefit during months they engage in SGA. As under current law, WISP subjects who initially qualify for benefits must not engage in SGA until they have satisfied the medical eligibility criteria for SSDI entry.
2. **Elimination of the EPE.** WISP would also eliminate the EPE. As above, WISP treatment subjects would continue to receive their full SSDI benefit during months they do not engage in SGA, and no benefit when they do, until their benefits are suspended or terminated for some other reason.
3. **EXR phased out and IRP eliminated.** When WISP is first introduced, some beneficiaries will be in their 60-month EXR following termination for work. If reinstated during that period after WISP is introduced, they will be immediately subject to WISP rules; the IRP will be eliminated. The EXR will be eliminated after all those in the EXR at WISP start-up have been reinstated or have exhausted their 60 months.[11]

A schematic diagram that captures the first three simplifications appears in Exhibit II.2. WISP would replace much of the current rules and procedures for SSDI beneficiaries who have earnings, as described in Exhibit II.1, with the simplified rules and procedures in Exhibit II.2. Suspension of benefits would occur as soon as earnings paid during a month exceed the SGA level (even if the beneficiary was not actually engaged in SGA, as currently defined, during that month). Suspension would continue without limit until monthly earnings fall below the SGA level or benefits are terminated for another reason.

2. CDR and Earnings Evaluation Changes

WISP would also reduce administrative burdens for SSA and beneficiaries in three key areas:

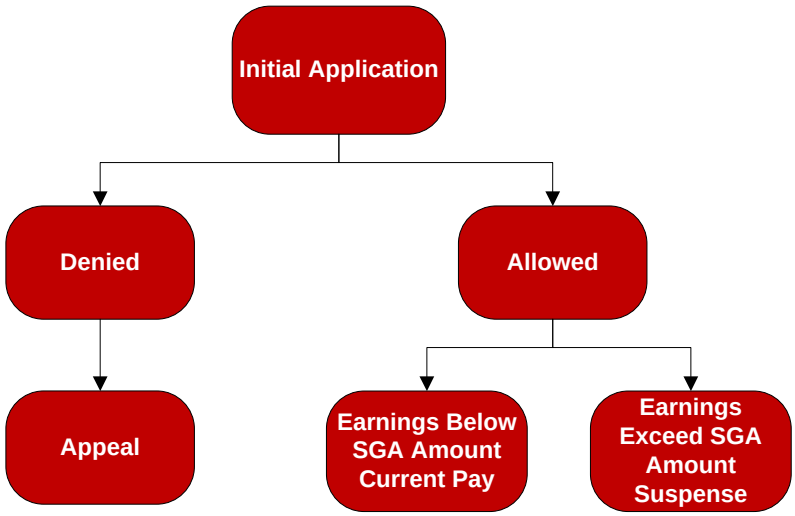
4. **Changes for CDRs.** Under WISP, work CDRs would be eliminated, but payment determinations based on WISP income counting rules would still be conducted. A full work CDR takes about 186 minutes, whereas a WISP payment determination would only take an estimated 20 to 40 minutes. SSA expects that many payment actions under WISP will be fully automated and will not require any intervention by SSA staff. Medical CDRs would be performed as normally scheduled based on the medical improvement diary. WISP subjects in benefit suspension would be eligible for a medical CDRs, just a beneficiaries are under current law when in their benefits are suspended for work. As WISP subjects' benefits can only be terminated due to medical improvement, not

[11] Those already in the IRP at the conversion to WISP will also be subject to WISP rules during the remainder of what would otherwise be the IRP, as benefits are suspended during the IRP in months when the beneficiary engages in SGA.

performing medical CDRs for WISP subjects in benefit suspension would allow some beneficiaries who medically improve to continue receiving Medicare and other benefits.

5. **Evaluation of earnings when paid and calculation of earnings.** Under WISP, monthly entitlement to SSDI income benefits would be based on earnings when paid, rather than when earned (as is done under current rules). In addition, under WISP, SSA would no longer exclude subsidies or special conditions when calculating a beneficiary's earnings.[12]

Exhibit II.2. WISP Provisions Simplify Incentives Within SSDI



A full evaluation of WISP would take account of the effects of these simplifications on beneficiary behavior and SSA's operational systems. The changes would likely result in more work reviews given the elimination of work CDRs and longer beneficiary eligibility periods. The changes should also affect over and underpayments made by SSA to WISP subjects and potentially ease the monthly income reporting for WISP subjects.

3. Medicare and TTW changes

WISP would simplify other long-term supports that could influence return-to-work decisions by SSDI beneficiaries. Of particular importance are the changes to the Medicare eligibility rules. Additionally, the changes in monthly payments noted above also have implications for TTW payment rules, though these changes would not necessarily fundamentally change the provision of services. The two key changes are as follows:

6. **Permanent Medicare continuation for those who do not improve medically.** A major change under WISP is the permanent continuation of Medicare eligibility. Medicare coverage would continue indefinitely to all SSDI beneficiaries unless their benefits are SSDI benefits are terminated for some other reason before they are eligible for Medicare because of age. Specifically, under WISP, Medicare coverage would be

[12] This change would be consistent with how SSA records earnings for the Supplemental Security Income program (described below in Section II.C).

extended from the 93 months after TWP provided under current law to an indefinite period.[13]

7. **Changes to TTW payments using earnings paid.** WISP would alter the TTW reimbursements to providers under WISP. Administratively, SSA would make TTW payments to ENs based on beneficiary's reports of earnings when paid rather than when earned as under current rules. Additionally, while not an administrative change to the TTW program, the elimination of the TWP could affect beneficiary and provider decisions to participate in the TTW given the way TTW payments are structured to ENs.[14]

4. Summary Comparison of WISP to Current SSDI Program Rules

Exhibit II.3 provides a side-by-side comparison of existing SSDI provisions and how WISP would alter them. While WISP simplifies incentives, the comparison illustrates the large number of SSDI rules that would eventually be affected by a demonstration, which is particularly relevant to the discussion in Section IV about informing beneficiaries and key stakeholders about WISP.

C. Other SSDI Program Interactions

The WISP rules described above also have implications for programs that interact substantially with SSDI, particularly Supplemental Security Income (SSI) and Medicaid. A substantial share of SSDI beneficiaries concurrently qualify for SSI. Social Security Administration (2010) reported that 27 percent of SSDI beneficiaries also received SSI benefits in December 2011. SSI has its own unique work incentives, which may interact with SSDI work incentives for concurrent beneficiaries who return to work.[15] In addition, the vast majority of SSI recipients are eligible for Medicaid. When SSI benefits are suspended for work, recipients are allowed to maintain their Medicaid eligibility

[13] Beneficiaries under WISP would be eligible for Medicare just as they currently are, with all the same rules regarding premiums and enrollment periods. The only thing that would change is that eligibility for Medicare would not terminate due to work activity. WISP beneficiaries would retain eligibility to Medicare (all parts) until death or they achieve medical improvement and no longer meet medical eligibility requirements. Implementation of the Affordable Care Act (ACA) could have an effect on Medicare expenditures for beneficiaries who are working. For instance, if they are more likely to have employer coverage, then Medicare will more likely be second payer. This is an area that needs further exploration as health reform and WISP proceed.

[14] Under current TTW rules, ENs can receive payments based on a milestone or outcome basis. Under the milestone system, SSA makes initial payments in two phases based on earnings above trial work (Phase 1) and earnings above SGA (Phase 2). In the outcome payment approach, SSA makes payments for earnings above SGA and benefit payments equaling zero. Under WISP, taking away the TWP could impact outcomes payments for ENs, as some payments are only made when work is above SGA and benefits suspended. If the evaluation finds that taking away the TWP results in parking, it could have impacts on EN service. For more information on EN payment terms for 2012, see https://yourtickettowork.com/web/ttw/en-payments-options (accessed April 17, 2012).

[15] The main SSI work incentive concerns how earnings are treated in determining the size of the SSI benefit. Because SSI is a means-tested program, each dollar of income from almost any source (including SSDI) reduces the amount of the SSI benefit dollar-for-dollar after a $20 income disregard. The major exception is earnings. There are special disregards for earnings, including a minimum monthly earnings disregard of $65 added to the $20 income disregard (if it is not already used against other income). Only half of earnings above these disregards are counted against the SSI benefit. That is, for every $2 in earnings above the disregards, SSI benefits are reduced by just $1 instead of $2. Unlike SSDI benefits, SSI benefits are not suspended and eventually terminated if the beneficiary engages in SGA; instead, the $1-for-$2 reduction continues until the SSI benefit is zero.

Exhibit II.3. Comparison of Current Program Rules to New Rules Under WISP

Item	Present Law	WISP
Work Incentive Provisions		
Trial Work Period (TWP)	• In 2012, a trial work month is any month in which total earnings are $720 or more.[16] The TWP continues until a beneficiary has worked nine trial work months within a 60-month period.	• The TWP, EPE, and grace period would be eliminated. Beginning with the first month of benefit entitlement, benefits would be paid for any month when the beneficiary is not engaged in SGA and would not be paid for months when the beneficiary is engaged in SGA.
Extended Period of Eligibility (EPE) and Grace Period	• The EPE begins the first month after the 9th TWP month. • After the start of the EPE, benefits are paid through the first month of SGA and for the subsequent two months ("grace period"). SSA calls the first month of SGA the month of "disability cessation." • After the grace period, benefits can be reinstated, with no need for reapplication, if the beneficiary stops engaging in SGA within the 36 month re-entitlement period.	
Expedited Reinstatement (EXR)	• In the 60 months after termination for work, the beneficiary may ask SSA to restart his or her benefits immediately if he or she is no longer able to continue working because of his or her condition. • Former beneficiaries who successfully reapply for benefits under EXR enter an initial reinstatement period (IRP). As with the EPE, during the IRP benefits are paid only when the beneficiary does not engage in SGA. Beneficiaries in IRP also receive Medicare coverage regardless of earnings. The IRP ends only after 24 months of benefits are paid, and it can last indefinitely if the beneficiary continues to engage in SGA. • EXR medical determination uses the medical improvement standard.	• There would be no EXR because there is no longer an SGA termination. • EXR would be available to individuals who have terminated due to work activity during the 60 months prior to the effective date of WISP.
Administrative Changes		
Evaluation of Earnings	• Earnings counted "when earned" for both wages and self-employment earnings.	• Earnings counted "when paid."
Continuing Disability Reviews(CDRs)	• Periodic medical CDRs scheduled according to diary maturation. Model determines whether to do a Full Medical Review or to send a mailer. • Work CDRs also performed at the end of a TWP as a condition for providing the EPE benefit.	• Periodic medical CDRs would continue. • Since TWP is eliminated under WISP, there is no associated work CDR at the end of the TWP.
Changes to Other Related Programs		
Ticket to Work Program	• TTW providers provide employment services to beneficiaries from whom they have received a "ticket" based on milestone or outcome payment system.	• Providers would use the "when paid" concept. Additionally, milestone and outcome terms might be modified once WISP intervention is fully specified.
Medicare Eligibility	• If cash benefits terminate due to work activity, eligibility to Medicare continues for at least 78 months after the 1st month of SGA occurring after the 15th month of the EPE. • Afterward, Medicare coverage is available only by paying a monthly premium.	• Lifetime eligibility unless benefits are terminated due to medical improvement.

[16] A self-employed person is also charged with a TWP month for a month in which he or she spends more than 80 hours in self-employment activities.

under a provision called Section 1619(b), provided that their medical condition does not improve and their income does not exceed a fairly high threshold (which varies from state to state).[17]

Some SSDI beneficiaries receive benefits from other public and private programs as well, further complicating the effects of changes in earnings on their economic well-being. Examples of such programs include Medicaid Buy-in (MBI), workers' compensation, private disability insurance, veterans' compensation or pensions, the Supplemental Nutrition Assistance Program (SNAP), housing assistance, and transportation assistance. In 2006, an estimated 6.0 percent of SSDI beneficiaries received veterans' benefits, 19.1 percent received SNAP, 2.3 percent received workers compensation, and 5.1 percent received private disability benefits (Livermore et al. 2009).

[17] For more information on 1619(b) eligibility rules, see http://www.ssa.gov/disabilityresearch/wi/1619b.htm.

III. KEY EVALUATION CONSIDERATIONS FOR TAP RECOMMENDATIONS

This section reviews information on evaluation issues presented to the TAP in briefing materials prior to their in-person deliberations. Before the start of this project, SSA formulated five key research questions that the agency would like the WISP evaluation to address. These research questions were included in the briefing materials and used as a starting point for the TAP's deliberations on the major features for the evaluation design. To provide some context on how many SSDI beneficiaries might be affected by WISP's provisions, the briefing materials also included information on the work incentive use and employment outcomes of current beneficiaries. Finally, the briefing materials included a summary of findings from SSA's Office of the Actuary (OACT) initial projections on WISP's potential effects, particularly on subgroups. The research questions highlight the major outcomes of interest, and the summary of work outcomes and actuarial projections provide a glimpse of WISP's potential, particularly on beneficiary subgroups. Each of these issues were relevant to the TAP's deliberations to the evaluation design and hence, provide additional context for the discussion in next section. The remainder of this section describes the original five SSA's research questions and summarizes the actuarial projections.

A. Five Key Research Questions Related to WISP's Potential Outcomes

SSA expects WISP to increase the number of beneficiaries reentering the workforce, reduce post-entitlement workloads for SSA staff, improve payment accuracy, and create new avenues for beneficiaries to report earnings through an automated system. These outcomes might well be realized, but there is always risk that actual outcomes might be different. SSA anticipates that the WISP evaluation would provide evidence to answer these general questions:[18]

1. **How do the various WISP provisions affect the work behavior and benefit payments of SSDI beneficiaries?** Is it possible to differentiate the effects of WISP individual provisions (for example, changes to the EPE vs. indefinite Medicare eligibility) on outcomes? How many beneficiaries do not make a work attempt due to the loss of the TWP? How many beneficiaries are encouraged to return to work due to the provision of permanent disability entitlement and permanent Medicare eligibility? How many additional months are spent on the SSDI rolls and what is the cost of the additional cash benefits?
2. **How do the various WISP provisions affect Medicare costs, utilization rates, and use of private insurance?** How many additional months of Medicare coverage, costs, and so on result from WISP? Do more individuals accept jobs that do not offer health coverage because of WISP? Under the Affordable Care Act, other health care options may be more advantageous for beneficiaries; therefore, should WISP consider an Opt Out clause to detach from Medicare? Should potential changes to WISP be identified that reduce Medicare costs and maintain the principle of administrative simplification?
3. **How much does an automated system of earnings reporting and benefit determination used by beneficiaries affect SSDI workloads?**

[18] The bolded five questions were listed in SSA's solicitation for the TAP contract. Additional questions were added to provide context on the five original SSA research questions by SSA and Mathematica staff by to the February 10th meeting.

4. **What is the impact of WISP on administrative costs and improper payments?** How does WISP impact administrative costs and savings (for example, processing medical and work CDRs)? Does it reduce overpayments and underpayments to beneficiaries? What effect does it have on the time frame for processing benefits?
5. **What is the potential for induced entry and how would SSA measure it?** Is it feasible to measure induced entry? If so, how should it be estimated?

For each research question, it is critical to understand what adverse outcomes could develop. For question 1, a potential adverse outcome concerns the level of SSDI benefit payments. WISP would not necessarily be viewed as a more desirable benefit design by all beneficiaries, because the elimination of the TWP would reduce benefits for some when they initially returned to work. That disincentive could offset the value of indefinite extension of eligibility for those engaged in SGA, creating indefinite Medicare attachment, and generally simplifying the design and administration of benefits. If the simplifications have no effect on or increase beneficiary earnings, months with no payments would increase because of the TWP's elimination. However, because beneficiaries would presumably understand the simplified rules better and be better able to avoid unintentional loss of benefits because of work, some might reduce their earnings earlier to avoid benefit loss. Further, for those whose benefits would be terminated under current law due to work, it would be easier under WISP to return to beneficiary status when earnings fall, which might mean that more would do so, or do so more quickly. At the extreme, some might use WISP as a temporary unemployment insurance program during times of economic downturn.

For question 2, the WISP evaluation design should consider the potential conflicting effects on Medicare expenditures. SSA would like to use WISP to better understand Medicare utilization rate differences between those who return to SSA disability beneficiary rolls just prior to the termination of extended Medicare through EXR and those who remain off SSA rolls several years after termination of Medicare. Additionally, if more beneficiaries secure good jobs and obtain private employer coverage, Medicare expenditures might fall. Consequently, SSA would also like to know how many beneficiaries obtain employer-based health insurance while maintaining Medicare as secondary payer. This would be especially true if implementation of health care reform induces more employers to offer coverage and/or induces employees to enroll in any coverage offered. But extension of the Medicare continuation period seems almost certain to increase Medicare expenditures.[19]

For question 3, the WISP evaluation design should consider whether the simplifications in WISP are well understood by WISP treatment subjects relative to the information provided to current beneficiaries. The key issues are whether such a system can be implemented to replicate what might happen in a national program and if the automated earnings report system would encourage timely reporting to reduce overpayments. Answers to both of these questions would hinge on SSA's

[19] Under the Affordable Care Act, other health care options may become more advantageous for beneficiaries to consider. Individuals covered by Medicare cannot receive subsidies for insurance purchased through the exchanges. The goal of WISP is to help eliminate the fear of the loss of insurance, not to disadvantage a beneficiary in terms of future options that could become available under the Affordable Care Act. A beneficiary might want to elect an option to opt out of his or her eligibility for Medicare in order to use other options that are currently designed to prevent utilization if an individual is already Medicare eligible. For example, if an SSDI beneficiary engaging in SGA believes that a private health care plan available on a state health insurance exchange is preferable to Medicare, then the beneficiary might opt out of Medicare so that he or she can receive a subsidy to purchase the private health care plan.

ability to effectively use automation and integrate the WISP intervention within its current infrastructure.

For question 4, it seems likely that the WISP intervention would reduce administrative costs for SSA. These cost reductions would likely occur because post-entitlement work for work incentive users would be much simpler. Elimination of work CDRs after the TWP would be especially important in reducing costs, as would the change in earnings measurement. Automated processing of earnings reports, elimination of terminations for work, and elimination of termination of Medicare eligibility also appear likely to produce significant administrative savings. However, there could also be increases in costs, particularly for those whose benefits are terminated under current rules. For example, this group might require more post-entitlement work because of their continued attachment to the program, including medical CDRs. It also is possible that WISP would increase the number of beneficiaries who require post-entitlement work simply because more beneficiaries have substantial earnings, although this effect would probably have to be surprisingly large to offset the expected savings from reducing other administrative costs.

For question 5, it is possible that WISP might encourage some people to apply for SSDI benefits who otherwise would not under current program rules. For instance, those with significant impairments who can engage in SGA but have short-term earnings prospects near or below the SGA amount might be induced to apply for SSDI benefits under WISP because the benefits would supplement their income until they can find a job that pays well above the SGA amount. Induced entry has been a major concern for the proposed $1-for-$2 benefit offset for earnings above the SGA level currently being tested under the Benefit Offset National Demonstration (BOND), and also has been a major concern in debates about elimination of the Medicare waiting period.

B. Work Patterns Under Existing Rules

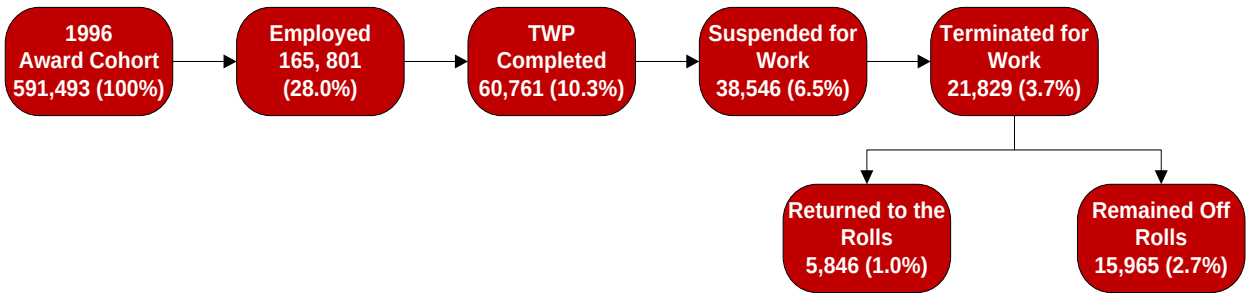
In Exhibit III.1, we summarize the work patterns of SSDI beneficiaries from a cohort of new awardees in 1996 at work incentive milestones, such as the TWP, to illustrate how WISP's provisions might immediately affect beneficiaries. We observe the work behavior of SSDI beneficiaries over multiple periods, which is important beneficiaries can move through different work stages during WISP (for example, TWP and EPE). While the economy and some relevant policies have changed since 1996, the outcomes of this cohort nonetheless provide some insights on the potential experiences of WISP, particularly of control subjects, that are helpful in thinking through options for the evaluation.

As Exhibit III.1 illustrates for the 1996 SSDI cohort, a substantial portion of SSDI beneficiaries return to work after entering the program, but far fewer complete the TWP and engage in SGA long enough to have their benefits suspended for work. Specifically, 28 percent of the 1996 SSDI cohort had posted earnings above the TWP level for at least one month, though less than half of those workers (10.3 percent) completed the TWP by working above the TWP level for nine months over a five-year period. Only 6.5 percent of the cohort had their benefits suspended for SGA. Furthermore, over a quarter of those whose benefits were terminated for work eventually returned to SSDI.

While not shown in the exhibit, Liu and Stapleton (2010) also showed that return to work rates are substantially higher over multiple periods than they are in a single cross section. For example, they found that less that approximately 15 percent of the 1996 cohort worked in a single year (1998), but by 2006, the number of workers in that cohort nearly doubled (to 28 percent).

These findings have implications for both the target population and the potential timing of impacts. Given that most SSDI beneficiaries do not complete their TWP, it is likely that the immediate impacts on a broad caseload would be small. However, given that work effort of beneficiaries will likely increase over time, it is important that the demonstration operate for a sufficient period to observe WISP's impacts.

Exhibit III.1. Use of Work Incentives by the 1996 SSDI Award Cohort Through 2006



Source: Statistics are from Liu and Stapleton (2010).

C. Projections of WISP's Costs and Benefits Using Key Sample Assumptions

In the planning stages of WISP, SSA's Office of the Actuary (OACT) projected the impact on expenditures if WISP's provisions were to become part of the national SSDI program.[20] These projections were meant to provide an initial estimate of WISP's potential cost. However, OACT's assumptions, estimates, and projections were not meant to influence or constrain the TAP's recommendations.

Cost estimates for the first six fiscal years after enactment project that WISP would result in a net savings in SSDI benefit payments, but those savings would eventually be overshadowed by new entrants and beneficiaries not being terminated due to work, resulting in a net increase in costs to the DI trust fund. Additionally, Medicare costs would increase throughout WISP.

OACT's assumptions about the sampling of certain subgroups are relevant to the evaluation design. The projections consider the following SSDI beneficiary subgroups:

- Beneficiaries who have never entered a TWP

[20] CMS's actuaries also conducted WISP cost projections to estimate the effects of Medicare expenditures. We were only given general details about the CMS projections, so have no information on effects for specific groups. One notable aspect of the CMS projection is the exclusion of beneficiaries age 55 and over. This exclusion sparked interest at SSA regarding whether an age restriction should be placed on WISP subjects, which is outlined in Section IV. This restriction was considered by CMS due to the fact that SSDI beneficiaries over age 55 already essentially have indefinite Medicare eligibility unless they recover medically—93 months of extended coverage under SSDI, followed by lifetime coverage starting at age 65—and therefore would not be affected by WISP's Medicare provisions.

- Beneficiaries who completed a TWP, were in EPE, but had not engaged in SGA
- Beneficiaries who completed a TWP, were in EPE, and had engaged in SGA
- Terminated beneficiaries in EXR
- Terminated beneficiaries in EXR and might reapply for benefits

Each of these groups would affect WISP's eventual net cost or savings. The savings come from three sources: (1) reduction in benefits paid during the former TWP to beneficiaries earning in excess of SGA, (2) reduction in benefits paid during the former EPE due to increased work effort, and (3) the reduction in benefits paid after the former EPE due to increased work effort among beneficiaries (include those with planned and long-term efforts). The costs also come from three sources: (1) increase in benefits due to continued entitlement of beneficiaries whose benefits would have been terminated in the absence of WISP, (2) increase in benefits resulting from induced SSDI applications, and (3) increase in benefits resulting from reduction in work effort during and after the former EPE due to elimination of the TWP incentive.

The SSA OACT projections provide context on WISP's potential impacts and were helpful in considering potential evaluation outcomes and impacts of subgroups of interest described in the next section. It is possible that WISP could have effects in each of these areas on the broad beneficiary pool, though its largest effects are most likely on beneficiaries who have already earned above the SGA amount, particularly those who are in or nearing the completion of the TWP, EPE, or EXR.

IV. TAP RECOMMENDATIONS

This section summarizes the topic areas discussed by the TAP members and their recommendations for the evaluation. The summary covers the following seven domains:

1. Research questions and key outcomes
2. Information dissemination/policy relevance
3. Evaluation design
4. Sampling and subgroups
5. Data sources
6. Timely medical CDRs
7. Outreach

The domains are generally organized according to topics outlined in the briefing materials and TAP meeting agenda.[21] Within each of these domains, we provide a summary of the TAP members' input, highlight their key recommendations, and provide additional discussion.

A. Research Questions and Key Outcomes

As described in Section III, SSA developed five primary research questions that capture the WISP outcomes of greatest interest to SSA. These questions and their related outcomes relate to administrative efficiencies, such as administrative burden and overpayments, and beneficiary outcomes, such as employment, earnings, health care, and benefit applications. The TAP was asked to consider whether these five research questions covered the most pressing issues for the evaluation. They also were asked whether additional information on other outcomes should be added as key questions and/or outcomes for a comprehensive evaluation.

The TAP members strongly agreed that the five research questions covered important outcomes for a basic evaluation. They noted significant challenges in measuring induced entry into SSDI because its effects were likely to be small, though some TAP members suggested potential methodological approaches to measuring this outcome. . They also thought a more comprehensive evaluation would include information on several outcomes not covered in the five research questions. Many, but not all, of the additional outcomes recommended would require a follow-up survey. Some of these outcomes, such as impacts on health, would be especially important for conducting a full cost-benefit analysis of WISP.

[21] In Appendix A, the TAP summarizes its responses according to nine categories. However, this section includes a summary for seven categories as we combined three categories from the briefing materials (external validity and site selection were incorporated into the summary of the evaluation design) because the TAP's comments in these areas overlapped. We also changed the ordering of sections from Appendix A to improve the flow of the discussion.

1. Outcomes Related to SSA's Five Research Questions

The TAP members emphasized the need for the evaluation to produce information on outcomes related to SSA's primary research questions. For each question, they identified several potentially important outcomes:

- **Employment, earnings, and benefit changes (research question 1).** To address research question 1, the evaluation would need to produce information on any employment, as well as specific SSA program milestones, such as earnings above SGA, entry into the TWP, entry into the EPE, and termination status. It would also need to include information on benefit outcomes, such as changes in benefit amounts, annual cessation rates, and change in benefit payment months (that is, number of months ineligible for income benefit) for SSDI and, if applicable, SSI.
- **Medicare/Medicaid (research question 2).** To address research question 2, the evaluation would need to produce information on Medicare and Medicaid program eligibility and payments. These outcomes include duration of eligibility, co-payments, overall payments, and types of services provided to beneficiaries.
- **Automated earnings reports and benefit adjustment (research question 3).** To address research question 3, the evaluation would need to assess whether beneficiaries understand WISP, and whether this understanding affects the number of overpayments and underpayments.
- **Administrative costs and improper payments (research question 4).** To address research question 4, the evaluation would need to produce information on payments and processing time changes under WISP, including changes in staffing levels and hours. Other outcomes include changes in specific administrative procedures, such as changes in work and medical CDRs and changes in beneficiary reinstatement.
- **Induced entry (research question 5).** To address research question 5, the evaluation would need to produce information on the number of new SSDI applications and awardees following the implementation of WISP.

The TAP most strongly emphasized measuring beneficiary outcomes related to employment, earnings, and Medicare. Almost every TAP member mentioned that measuring benefit usage, employment, and earnings outcomes would be critical for addressing SSA's most basic questions regarding how WISP affects beneficiary behavior. Measuring Medicare and Medicaid outcomes would help reveal how health care coverage interacts with SSDI employment incentives. Impacts on some of these primary outcomes would be particularly important to policymakers and other stakeholders.

Several TAP members also stressed the importance of measuring WISP's administrative outcomes. For example, if WISP had no impacts on employment and benefit outcomes, it could be cost beneficial to SSA if it substantially reduced the amount of resources needed to administer the SSDI program. Additionally, WISP's earnings reporting simplifications could encourage beneficiaries to return to work by reducing payment errors and expediting benefit reinstatement for those whose earnings fall back below the SGA amount. For example, SSA, beneficiaries, and advocates all hope that WISP's automated earnings reporting system would decrease overpayments and increase the speed of benefit adjustment. The TAP members stressed that payment errors and reinstatement delays, which often occur in the current system, create strong return-to-work disincentives because

they add uncertainty to the future benefit payments of beneficiaries, especially for beneficiaries who fear work might jeopardize their benefit status.

When asked to provide more detailed input on administrative outcomes, the TAP stated that it was difficult to do so because SSA had not yet specified how benefit adjustments would occur under WISP. SSA anticipates some automated procedure, though it is not clear where and how this procedure would be implemented. Without this information, it is difficult to predict the potential impact of WISP on key administrative procedures, such as field office staff interactions with beneficiaries. These administrative procedures could have important implications for staffing at local offices, particularly given SSA's large initial claims workload. Additionally, the administrative setup has implications for how beneficiaries interact with WISP and potentially perceive WISP services, and might influence their interest in participating in the demonstration or related services, such as WIPA or TTW.

2. Challenges in Measuring Induced Entry

The topic of induced entry received special attention from several TAP members because of the methodological challenges associated in detecting a small impact that could have major policy implications. In their follow-up comments, two TAP members wrote that measuring induced entry, though difficult, would be critical to any evaluation of WISP—especially a program cost analysis—and consequently should be attempted. As described in more detail below, individual members did mention possible options for measuring induced entry, using alternative methodological approaches and subgroup analyses. However, one member questioned whether any methodological approach could adequately capture induced entry effects, reflecting the intensively-studied methodological difficulties in measuring similar induced entry effects for BOND.

3. Additional Outcomes for Comprehensive Evaluation

The general inclination of the TAP was to cover a fuller set of outcomes for a more comprehensive evaluation. One TAP member stated that SSA's five research questions covered the necessary evaluation topics but did not constitute a comprehensive evaluation. All of the TAP members suggested at least one addition to the list of measured outcomes, including the following:[22]

- **Use of work supports, such as WIPA, TTW, VR, and One-Stops.** Several TAP members noted that some WISP subjects would use work supports from WIPA, TTW, VR, and One-Stops as an intermediate step to increasing employment. Measuring work support usage would provide information on how WISP subjects are progressing to the ultimate goal of employment. Additionally, if use of these supports changes under WISP, it would be important to capture this for WISP's benefit-cost evaluation.
- **General health, including Activities of Daily Living (ADL).** Several members noted that there is a relationship between employment and general health, with some even suggesting that employment improves general health. To better capture this relationship, a few members suggested monitoring the general health of WISP subjects, including

[22] These outcomes are generally summarized in order of the frequency they were mentioned by TAP members during the meeting and/or in their written input. We use the term "several" to denote that at least three or more TAP members noted a key issue.

their ability to perform ADL, as an exploratory outcome. One member also suggested "it would be useful to try and estimate the impact of WISP on Quality Adjusted Life Years."

- **Work attempts/job search.** At least two TAP members expressed interest in tracking work attempts and job search during the demonstration. Similar to the use of work supports, information on work attempts would provide information on the intermediate steps taken by beneficiaries toward employment.
- **Participant satisfaction.** One TAP member strongly suggested gathering information on participant satisfaction to gain the beneficiary perspective on the intervention. A beneficiary's satisfaction with services would provide SSA with important information on whether beneficiaries preferred the new work incentives and earnings reporting system to the existing system.
- **Other health insurance coverage.** One TAP member emphasized the importance of including private health insurance coverage as an outcome. Some WISP participants who find employment may obtain private health insurance, making Medicare their secondary health insurance payer. Measuring this outcome would reveal how many WISP subjects obtain private health insurance as well as how much Medicare expenditures decrease when Medicare is the secondary payer.
- **Income/consumption.** The TAP debated the addition of income as an outcome variable, given that the SSDI program is an income support program. Because of the challenges in obtaining total family or household income from a limited set of measures, there was not a uniform suggestion to add income as a key outcome. One TAP member suggested that consumption be measured as a direct measure of economic well-being. This TAP member pointed out that survey respondents generally report consumption more accurately than they report income.

4. Discussion

The TAP recommendations for the WISP evaluation to extend beyond the original five research questions reflects the members' desire for an in-depth evaluation of outcomes. The additional outcomes proposed by the TAP would allow for a more complete accounting of the benefits and costs in WISP. A follow-up survey would be required to measure most of the additional outcomes; the exception is the work support outcomes, which are captured in administrative data. The value of the additional information would need to be weighed against the value of using research resources in other ways.

One option to consider is to have the WISP evaluation track the same administrative, and possibly survey, outcomes as the BOND evaluation. Such an arrangement would help facilitate comparisons across the interventions, both of which alter work incentives and administrative processes to promote employment for beneficiaries. If SSA decides to conduct a survey for WISP, it could follow the approach used for BOND and field a year three WISP follow-up survey. The timing of this survey would be consistent with the WISP TAP's recommendations for when several key outcomes might emerge for treatment group members, which is described in Section IV.B below.

The TAP's recommendations indicate that finding a credible methodological approach for measuring induced entry into SSDI will be a difficult task. It is unlikely that an effect on induced entry can be measured using a fixed sample of randomized subjects unless that sample is especially

large. For this reason, an alternative methodological approach, such as those offered by two TAP members using a non-experimental design or subgroups, is a possibility (and described in more detail below). Additionally, if SSA develops a credible approach for estimating induced entry effects in BOND, which is still under development, there is a potential for this approach to be applied to WISP.

B. Information Dissemination/Policy Relevance

After discussing what questions and outcomes the WISP evaluation should consider, the TAP contemplated when and how WISP impact information should be reported. On the one hand, policymakers need timely information to quickly incorporate knowledge obtained from WISP into future policy decisions, and they might question WISP's value if they must wait too long. Conversely, their need for information on the progress of WISP must be balanced against the concern of producing misleading early findings on outcomes that might take longer to emerge, such as employment impacts of WISP for beneficiaries who are not yet in the TWP. In addition to informing policymakers, the evaluation team would need to consider how to best engage and inform other stakeholder groups, such as beneficiaries, advocates, and employment support providers.

When recommending reporting timetables and procedures, the TAP made three substantive recommendations. First, the TAP stated that different outcomes should be reported at intervals based on anticipated impacts. For example, because WISP is expected to have an immediate impact on administrative issues, impacts on SSA administrative burden could be reported after the first two years of implementation. Other outcomes, such as WISP's impact on employment and health care outcomes could take longer to develop and should be reported in later years. Second, the TAP agreed that the WISP evaluation should last at least five years. Third, the TAP stressed the importance of informing advocates about WISP's impacts, which could be done in shorter documents written in plain language.

1. Timeline for Impacts Varies by Outcome

The TAP encouraged periodic reporting of impacts on outcomes that could be readily measured in the administrative data, throughout the demonstration period. Such reporting would help maintain policymaker interest in the WISP demonstration.

However, the impact estimates should be interpreted relative to a timeline for impacts that is specified prior to the evaluation. As one TAP member noted, WISP's impacts on employment behavior "will likely evolve relatively slow over time. While electoral imperatives necessarily drive politicians to have short time horizons, given the importance of the SSDI program both in terms of budget and in terms of the number of relatively vulnerable individuals it serves, getting the correct answer, rather than getting the quick answer, surely is the social optimum."

Though the TAP members did not formally define a timeline of expected impacts, they did have general expectations about how long it would take some impacts to emerge. Specifically, in the short term (two years after implementation) WISP should reduce administrative burdens and overpayments to beneficiaries. In the longer term (beyond two years), WISP should affect key beneficiary outcomes, including employment, earnings, and benefit payments. The TAP members did not, however, incorporate administrative reporting lags into these time estimates. SSA usually cannot identify that a beneficiary is engaged in SGA for at least 18 months after SGA starts unless the beneficiary notifies SSA on the actual reporting of these outcomes (i.e., they report this information to local field offices).

The TAP members were also briefly asked to consider what findings would be considered "convincing" to policymakers. Three members stated that overall SSDI program savings—whether from reducing administrative costs or overall benefit payments—would strongly suggest to policymakers that WISP should be implemented nationally. Two members also thought that higher beneficiary employment and less "parking" by beneficiaries would show WISP's long-term policy value.

2. Long-term Outcomes Should Be Tracked for at Least Five Years

There was a consensus that the evaluation should track outcomes for at least five years. The five-year timeline underscored the TAP's recognition that several beneficiary outcomes, particularly impacts on earnings and benefit receipt, could take several years to develop. As one TAP member noted, SSA has tremendous flexibility in measuring these outcomes indefinitely because impacts on benefits and earnings can readily be tracked inexpensively using SSA administrative data. Two members suggested an indefinite follow-up period, and another TAP member suggested measuring outcomes at 5, 7, and 10 years after implementation.

One TAP member provided a particularly detailed description of how short- and long-term impacts could be measured and reported in a future WISP evaluation. The member suggested that the following outcomes be measured annually: work behavior, benefit receipt, improper payments, health insurance utilization, health status, and ability to meet basic needs. Some of these outcomes could be measured to a degree in administrative records (work behavior, benefit receipt, and improper payments), but a special survey would be needed to track more detailed outcomes in these areas (for example, on hours worked) and health-related outcomes that are not measured in administrative records. For some other outcomes, such as subjects' understanding of WISP features and incentives, automated wage reporting system use, SSDI workloads, administrative costs, and induced entry, the member recommended assessing them once early after implementation (for example, after 6 months to one year) and then again later (for example, two to three years) during the demonstration to see how the outcomes have changed over time. However, as noted above, it may take longer to report on some of the SSA administrative outcomes, such as overpayments, due to delays in the ways this information is reported. This TAP member advised disseminating findings at 12 months, 24 or 30 months, and 60 months after implementation. The first findings report could provide SSA with an early assessment of how the program was functioning and offer a glimpse of early implementation experiences. The second report could provide an interim snapshot that would presumably include full administrative and initial beneficiary impacts. Finally, the last report would focus on beneficiary outcomes, especially those that the evaluation team expected would take years to develop.

3. Dissemination Roles for Consumers and Disability Advocates

Finally, there was a consensus that WISP's findings would generate strong interest outside of SSA, particularly from people with disabilities, consumer advocates, and other program officials. These groups would be interested both in WISP's eventual impact on the national SSDI program as well as how the demonstration affects treatment group members in the short run. For example, one member suggested that these stakeholder groups be told how random assignment was conducted for the evaluation, which could be outlined in the early assessment reports for WISP. Two other TAP members suggested that these stakeholder groups would be particularly interested in outcomes related to earnings reporting, including program effects on overpayments, the ease of earnings reporting, and the promptness of benefit reinstatement after earnings fall back below SGA.

4. Discussion

The consensus of the TAP is that the evaluation design should include a strong timeline of outcomes, which could be specified in a logic model of anticipated impacts.[23] For example, a lack of impacts in year one on employment outcomes should not be interpreted to mean that WISP does not affect employment, because we would expect any impacts to emerge later. Hence, SSA might want to avoid reporting impacts for employment in early evaluation reports. If this approach is not acceptable, then any report on employment impacts should prominently point out that no employment impacts are not expected in the first year and should point to pre-demonstration documents that state that expectation.

The TAP members also strongly emphasized informing key stakeholders, such as people with disabilities and state policy officials, which might most efficiently be done through policy briefs. Specifically, that summarize results in a short, concise fashion might be particularly useful to informing organizations that might be affected by WISP, such as state vocational rehabilitation agencies, Employment Networks, and Centers for Independent Living, about the demonstration's impact. Tailoring policy briefs to specific groups would help keep outside entities engaged in WISP.

Ultimately, the evaluation's time frame and the number of reports and policy briefs would depend on the WISP evaluation's budget constraints. Given the time frame for expected impacts, a minimalist evaluation would produce reports at annual intervals. A more comprehensive model is to follow the reporting structure of the BOND evaluation (see Bell et al. 2011). The BOND evaluation includes assessment report on start-up activities, annual "letter" reports on key impacts, and two comprehensive participation, process, and impact reports that document both qualitative and quantitative findings at later intervals in the demonstration.

C. Evaluation Design

The TAP discussed the evaluation's general design, including the use of experimental and non-experimental approaches, and how best to evaluate outcomes using an individual, site or "hybrid" (i.e., combination of individual and site) level approach. This discussion included an extended debate over the merits of each approach for WISP and the possible definition of what might constitute a "site."

The TAP members strongly agreed that the evaluation should employ an experimental design, and some expressed an interest in using a non-experimental approach to supplement the experimental findings. A non-experimental approach could be especially useful in examining

[23] As in previous SSA evaluation designs, a strong logic model would also address the "multiple comparisons" problem. The multiple comparisons problem arises when there are several statistical tests performed on an evaluation's outcomes. Specifically, as more statistical tests are performed, the probability that at least one test will produce a statistically significant result by chance increases. Thus, as the number of tests increases, simply comparing each test's p-value to the 5 percent significance standard could lead an evaluation to find spurious impacts instead of underlying true effects. To address this issue in the BOND, the Youth Transition Demonstration, and the Accelerated Benefits demonstration, a small number of outcomes were included in the "core" impact findings and several other outcomes were included as "supplemental" or "exploratory" (see Rangarajan et al. 2009 and Michalopoulos et al. 2011). This approach could be adopted for WISP. For example, WISP's core outcomes could include those outlined in SSA's primary research questions. A supplemental or exploratory analysis could focus on outcomes that might experience an indirect effect of WISP, such as general health.

outcomes where the samples were limited under an experimental approach (e.g., effects on a subgroup or approaches to estimating induced entry).

The TAP could not reach a consensus on how to define the unit of random assignment or how to define a site because critical aspects of the WISP intervention still need to be defined. Nonetheless, the TAP's deliberations in this area should be especially useful as SSA considers operational options for WISP in the future.

1. Importance of an Experimental Design

The TAP members strongly argued for an experimental design to provide convincing evidence of WISP's efficacy. As one TAP member noted, "because an experimental design assigns treatment randomly to either individuals or sites, it removes the possibility of non-random selection into treatment as a function of observed or unobserved covariates that also affect outcomes. For this reason, experimental designs provide, in general, clear and compelling causal evidence." The emphasis on experimental design reflected the TAP's strong desire to produce credible results that were internally valid. Additionally, one TAP member noted that SSA's history in using experimental methods further motivates its use for WISP: "SSA has already used experimental methods to test some important concepts, so it appears feasible to use randomized control trials… Secondly, the assumptions required for non-experimental approaches such as propensity score matching and instrumental variables are quite strong and generally not testable. Thus, stakeholders who do not like the findings could always challenge them."

2. Non-Experimental Methods for Secondary Analyses

While there was a clear emphasis on experimental methods, three members noted there were benefits to supplementing the experimental analysis with quasi-experimental analyses. These quasi-experimental methodologies would be selected after determining what subpopulations and outcomes would be best measured using non-experimental techniques. One member suggested using propensity score matching to construct internal comparison groups and multilevel modeling to investigate how field office variation affects implementation and outcomes. Another TAP member strongly encouraged SSA to create two non-experimental evaluations that parallel the primary experimental evaluation. One evaluation approach could use reduced form methodologies including difference-in-difference and matching techniques to measure program impacts. The other evaluation approach could specify and estimate a structural model (for example, see Todd and Wolpin 2006) that would then be used to conduct various counterfactual policy experiments. The results from the experimental evaluation would be used to evaluate the forecasts from the structural model.

3. Individual vs. Site Level Random Assignment Considerations

The TAP considered three options for an experimental design for WISP:

- Individual level random assignment only
- Site level random assignment only
- Hybrid design of individual and site level random assignment

The TAP identified two relative advantages for site level random assignment. First, SSA could implement this design for the treatment sites in a manner that is quite close to the manner in which WISP would be implemented under a permanent program. All relevant staff at the site field offices

would be trained in WISP, all beneficiaries in the site would be subject to WISP, and community outreach efforts could mimic those that would be used in a national rollout. Second, successful random assignment of sites would provide SSA with a nationally representative sample at significantly less cost than only using individual level random assignment. One TAP member supported site level random assignment with the following comment:

> "Having sites as the unit of treatment means that both the site SSDI staff as well as local support organizations will have a strong incentive to develop real expertise regarding the WISP treatment. In contrast, with individuals as the unit of treatment most sites will not have all that many claimants under WISP and so will have a limited incentive to develop such expertise. Thus, with sites as the unit of treatment the environment experienced by treated individuals will better approximate what SSDI would be like with WISP as the ongoing policy for all claimants; put differently, using sites as the unit of treatment should enhance the external validity of the evaluation."

The TAP's primary concern with site level random assignment was that, relative to individual level assignment, it would considerably reduce the evaluation's ability to detect effects large enough to be of interest (that is, it would reduce statistical power), holding the number of sample beneficiaries constant. The number of sites needed for adequate statistical power would largely depend on the degree of heterogeneity between the evaluation sites. There was also a concern that site level random assignment would compromise the intervention for treatment subjects who move to a nontreatment site because the treatment subjects would then receive services from SSA staff and organizations that are unlikely to understand the WISP intervention. Another concern was that beneficiaries in a treatment site, defined by an area, might prefer to use a field office that was in a control area, or vice versa. These last two issues would not be relevant to a national program, but special provisions would be required to address them for the demonstration.

The primary advantage of individual level random assignment was that this design has the most power to detect effects for any given number of treatment beneficiaries. The issue of power is especially important in WISP, given that the effects on some groups, especially new beneficiaries, might be very small and take time to develop.

There were three concerns with individual level random assignment. First, because the current SSDI program and WISP would be administered side by side throughout the country, individual level random assignment would likely increase WISP's implementation and administrative costs. Second, there is potential for intervention spillover effects. For instance, treatment or control group members might receive incorrect information about the work incentives they face, which would compromise the internal validity of the demonstration. Third, lessons for national implementation about the operation of administrative processes might be more limited.

The hybrid option combines site and individual random assignment designs. Under this design, sites would first be randomly selected and then individuals within those sites would be randomized to treatment or control. This approach would allow the demonstration to address some site conditions that might influence outcomes, such as the local program environment, but also take advantage of individual random assignment to increase power. Three of the TAP members noted the promise of the hybrid design for WISP, which is also the design currently being used in BOND. However, this option has many of the administrative drawbacks noted above for individual level assignment, as both treatment and control rules would have to be administered at each site.

4. Site Definition and Selection Depends on Operational Considerations

The definition of a site—and by extension the decision of whether to use site level random assignment—is heavily influenced by operational considerations. One reason the TAP failed to make an unambiguous recommendation about the evaluation design is that SSA has not yet decided how to administer WISP for the demonstration. The choice of operational design and the choice of evaluation design are linked, as described below.

A site level random assignment strategy would generally be preferred if SSA envisions field offices having primary responsibility for administration of WISP under a national program and, during the demonstration, wants to maximize learning about how the field offices will perform. During the demonstration, all relevant staff in the treatment field offices would be trained together and would operate under a common set of procedures. This approach would also facilitate demonstration interactions with other support organizations in the community, on a par with interactions under a national program.

At the opposite extreme, if SSA envisions WISP being primarily administered by centralized, automated processes, with relatively little support from field offices and engagement with other local entities (e.g., WIPAs), then an individual random assignment or hybrid strategy could be preferred. In this scenario, individual random assignment might be preferred, both to maximize power for any given sample size and to test the viability of centralized administration.

Should SSA prefer to use field offices to administer WISP, one other consideration is the readiness of field offices to implement WISP. If some field offices are poor candidates for WISP administration, then SSA might be well advised to avoid use of field offices for the test, as it would be problematic to choose a nationally representative set of field offices. Instead, it might be more attractive to administer WISP outside of field offices and use individual level random assignment within a small number of sites, defined in some other way.

Three TAP members stated that site selection should occur at the field office level, with one suggesting that it would be easiest to have field offices administer WISP. However, some TAP members doubted that beneficiaries would rely on local field offices for support, diminishing the importance of random assignment of field office areas. Even these members, however, thought it would be useful to define sites as geographic areas, because they thought that WISP would likely interact with state and local programs. Random assignment of beneficiaries to treatment and control within these sites would be attractive for power reasons, although it would also complicate outreach and increase the risk of crossover effects.

However the sites are defined, the TAP members thought that some consideration should be given to site characteristics that might affect outcomes in the site selection process. The TAP identified several programmatic differences across sites that could influence WISP outcomes and create potential differences in WISP impacts, particularly across states. These programs include WIPA; Medicaid/Medicaid Buy-in; VR; TTW; Housing/Independent Living; Mental Retardation and Developmental Disability programs; and One-Stops.[24] The TAP discussed whether sites should

[24] These programs are listed in descending order with the first program being mentioned by the greatest number of members

be stratified based on some of these characteristics given the potential for differences in site impacts. One member strongly opposed stratification because these interactions were part of the policy environment, and stratification might influence the external validity of the results. Two other members, however, saw the potential in using this variation to better understand WISP's interaction with other programs, and one even noted that the idea could be incorporated into the sampling design of the sites. Stratified random assignment of sites can potentially improve power, because it can potentially control for site factors that affect impacts. Further, weights can be used to help ensure that impact estimates based on stratified samples are externally valid. None of the TAP members made a strong recommendation on this issue, however, at least in part because operational plans for WISP were unknown.

5. Discussion

The TAP strongly recommended that WISP use an experimental design for the evaluation, which is consistent with SSA's general strategy of using random assignment when practical in other demonstrations. For example, SSA's recent demonstrations, such as Accelerated Benefits, BOND, Mental Health Treatment Study, State Partnership Initiative, and the Youth Transition Demonstration, have all used random assignment.[25]

The TAP's suggestion of using non-experimental methods to supplement an experimental evaluation holds promise for evaluating outcomes not included in the core design, and potentially for evaluating rule changes that differ from those incorporated in WISP. This suggestion mirrored the approach adopted for the State Partnership Initiative evaluation; however, we must note that the non-experimental methods used for the initiative proved to be problematic.[26] One TAP member also stated that evidence from the experimental analysis could be used to assess a structural model of applicant and beneficiary behavior, which is a methodological approach that has not yet been used by SSA but offers strong value because it would create a platform for testing several counterfactual policy scenarios.

The TAP's deliberations about the unit of random assignment are informative to SSA's ongoing deliberations about how to operationalize and evaluate WISP. The TAP's feedback indicates how the choice of evaluation design is linked to operational choices. If field offices would have primary responsibility for administering WISP and other local entities will play significant roles in supporting WISP, then a design that uses all beneficiaries in a field office area as treatment subjects becomes attractive.[27] If field offices would play a minor role, WISP can be administered centrally. In that case,

[25] SSA has used non-experimental methods in other major evaluations, such as TTW. However, this program was initially rolled out nationwide, per the authorizing legislation, so it was not feasible to test using random assignment.

[26] In the State Partnership Initiative, the experimental evaluation results showed the limitations of non-experimental methods in generating impacts (see Peikes et al. 2005). This was an important finding because the only non-experimental methods were available to evaluate impacts in sites that did not use random assignment. Hence, the findings about the limitations of non-experimental methods were critical to interpretation of demonstration outcomes.

[27] If a site level design is preferred, there are several potential alternative options to consider, though these options were not explicitly discussed by the TAP. For example, the site selection design that would minimize the number of field offices needed to achieve a given level of precision for impact estimates would be stratified random selection of field offices within strata defined by factors expected to affect outcomes. Another design that would be somewhat less efficient, but perhaps preferred administratively, would be a two-stage process for selecting field offices: (1) stratified random selection of a small set of area offices (excluding BOND area offices, which were chosen to be nationally representative), then (2) stratified random selection of a larger set of field offices within these area offices. This approach

informing local agencies and organizations would not be a key operational issue, an individual or hybrid design would be preferred, because smaller beneficiary samples could be used to detect impacts of any given size.[28]

D. Sampling and Subgroups

The fourth issue that the TAP considered was the size and composition of the WISP demonstration sample. As a starting point for this discussion, the TAP was referred to the actuarial assumptions from CMS and SSA outlined in Section III, which used some preliminary assumptions for identifying subgroups (for example, those in the TWP).

The TAP recommended the sample be composed of a representative sample of SSDI beneficiaries that was sufficiently large to detect policy relevant impacts. They suggested that SSA obtain further estimates by statisticians to identify necessary sample sizes under various designs. The TAP also identified several subgroups for which impact estimates would be useful, including beneficiaries who have already achieved work incentive milestones, such as the TWP. However, some TAP members more heavily emphasized the long-term benefits of focusing on new SSDI beneficiaries, for purposes of making projections of WISP program impacts.

1. Sample Size and Statistical Power Considerations

Given that an evaluation design has not yet been selected and that SSA has not quantified what effect sizes it would like the evaluation to be able to detect, most TAP members did not make specific sample size recommendations. One member stated simply that the sample size "should be sufficient to obtain estimates of the desired power for the smallest subgroup of interest." One TAP member noted that a conventional statistical power level should be used, such as 80 percent. Another suggested that SSA consult subject area experts in the academic and consulting fields to gather views regarding likely program effect sizes. Two TAP members recommended using the sample sizes used in the actuarial projections—80,000 subjects. This happens to be the same size as one of the treatment groups in BOND.

2. Subgroups Considerations

The TAP then considered what subgroups the sample should target and whether any subgroups should be oversampled. Although one member opposed targeting any subgroups, most TAP members agreed that certain subgroups should be overrepresented in the sample. As stated simply by one member, if stakeholders are particularly interested in a certain subgroup, then that subgroup

(continued)
might have administrative efficiencies for SSA because fewer area offices would need to be involved in the demonstration. It would also reduce the precision of impact estimates relative to a single-stage design, but the difference in precision might be modest.

[28] For example, if SSA implemented WISP from a central location, it could choose an individual level or hybrid design (similar to BOND). For a given number of treatment beneficiaries, the hybrid design would have more statistical power than a design involving random assignment of sites, alone, but less than one based on random assignment of individual alone. Compared to a random site design, the hybrid design would perform less well at creating a local environment comparable to the environment under a national program, but compared to a random individual design it would provide SSA with some ability to involve local organizations in potentially important ways.

should be well represented. For current SSDI beneficiaries, the subgroup sampling discussion centered on measuring WISP's potentially differing impacts for those who have achieved certain earnings milestones outlined in Section III (for example, those in the TWP).

Despite the lack of consensus on a particular subgroup to oversample, most TAP members agreed that it would be beneficial in detecting WISP's potential differential impacts on beneficiary subgroups to oversample those who have achieved certain earnings milestones. As noted in Section III, WISP could have minimal impacts on the overall beneficiary population, but very strong impacts on subgroups, particularly those in or near the TWP. Holding the total sample size constant, oversampling from groups likely to use WISP would increase the precision of overall estimates as well as support more-precise estimates for subgroups of likely users.

The TAP also considered sampling decisions around other subgroups, such as by age, though there were no definitive recommendations. Most leaned toward undersampling or even potentially excluding those over age 55, because they expected the impacts on this subgroup to be small relative to younger beneficiaries. The main point of contention was whether SSDI beneficiaries near the SSA retirement age would likely respond to the WISP intervention, especially considering that under current law SSDI beneficiaries near the retirement age essentially are guaranteed Medicare benefits for life, regardless of their work behavior, unless SSDI benefits are terminated for medical improvement. Three members did not see a problem with undersampling this subgroup, though one member conditioned this recommendation on whether the data suggest that members of this subgroup are less likely to return to work than younger beneficiaries. However, one TAP member thought that any exclusion would harm external validity and that the idea relied on the untested "assumption that those over 55 will not return to work under WISP."

One TAP member's suggestion for subgroups was particularly noteworthy because it provided SSA an option for using a non-experimental design to evaluate outcomes, presumably at relatively limited cost.[29] This member noted "there may be potential for a regression-discontinuity analysis of impacts using those who are just below age 55 compared to those just at/above this cutoff." This suggestion is important because it illustrates that SSA has other options for identifying subgroup effects beyond using sample stratification. The advantage of a non-experimental approach is that it reduces the sample that might be needed for the evaluation (that is, it reduces the need to oversample a particular group). However, the downside to a non-experimental approach for a subgroup is the same as the downside of this approach more generally; namely, non-experimental findings would not be as convincing as those from an experimental design.

3. Oversampling New SSDI Beneficiaries

The remainder of the subgroup sampling discussion focused on the relative merits of sampling the "stock" of current beneficiaries versus the "flow" of new applicants and beneficiaries. Although studying WISP's effect on the stock would provide the quickest impact estimates and best reveal how an immediate transition from SSDI's current work incentives to the WISP intervention would affect short-run costs and other outcomes, WISP's long-term effects are best measured by focusing on the flow. This is the response to WISP by new beneficiaries would likely be differently than those

[29] This recommendation is a practical example of the non-experimental option noted in Section IV.C.2.

who have been on the rolls for many years, even after the new beneficiaries have themselves been on the rolls for many years.

One member forcefully stated the importance of new beneficiaries to the evaluation given their potential long-term value to the evaluation. Specifically, this TAP member noted that examining differences in outcomes between these two subgroups, as was done for the evaluation of the Self-Sufficiency Project in Canada, could shed light on induced-entry effects in addition to providing externally valid impact estimates. This member also advised that due to sample size and statistical power considerations, SSA should focus its attention on two or three subgroups of particular interest, with at least one subgroup being new SSDI beneficiaries or applicants.

4. Discussion

The TAP did not reach a consensus on a specific sample size or which subgroups to target. There was a lengthy discussion over the merits of oversampling new beneficiaries, with one TAP member heavily emphasizing this approach. The one drawback of solely focusing on new beneficiaries, however, is that impacts on this group could take longer to develop. One option SSA might want to consider is to use a sampling strategy similar to that used in BOND, which included targeted samples of new "short" duration beneficiaries, who had been on the program for two years, and "long" duration beneficiaries, who had been on the program for over two years. This approach is generally consistent with TAP members who had a focus on new beneficiary subgroups and has the advantage of allowing comparison of WISP estimates to those from BOND.

If SSA could include large subgroup oversamples in WISP, then the TAP's multiple subgroup recommendations are very important, given that the size and time frame for WISP's impacts would likely vary significantly across these subpopulations. Once SSA defines WISP's sample target size, it could follow the TAP's recommendation to calculate minimum detectable effects to assess whether the sample size is sufficient to identify meaningful policy impacts.

An evaluation design that uses random assignment of sites but not of beneficiaries within site would not lend itself to oversampling for any subgroup; all beneficiaries in treatment sites would be subject to WISP. The required sample size for the smallest subgroup(s) of interest to SSA would in effect dictate the total sample size.

E. Data Sources

The TAP considered the data sources needed for the evaluation. The discussion built on the earlier recommendations about key research questions and outcomes (see Section IV.A).

The TAP recommended using a mixture of quantitative and qualitative measures for evaluating WISP impacts and examining whether WISP was implemented as designed. The TAP strongly recommended using all available administrative data—including those from SSA, CMS, and Rehabilitation Services Administration—to measure outcomes when possible, and noted that surveys would be needed for some outcomes, especially those outcomes identified in Section IV.B. The TAP also recommended that the fidelity of WISP program services be rigorously tracked using a combination of quantitative and qualitative data to determine whether WISP was implemented as envisioned.

1. Importance of Administrative and Survey Data

The TAP identified several potential administrative data sources that could be directly integrated into the evaluation. All TAP members noted that SSA's administrative data include demographic, earnings, benefit, and work incentive milestone information that could be leveraged at relatively limited cost for an evaluation. The SSA administrative data also include information from the SSA Master Earnings File, which contains annual earnings information from the Internal Revenue Service. Some TAP members also debated the merits of adding additional employment records for quarterly unemployment insurance wages, from the New Hires database, but expressed concerns about how or even if the evaluation team could access those data. Several TAP members noted that WIPA, EN, and VR administrative data would provide information on a variety of work supports. Finally, TAP members noted that Medicare and Medicaid claims would provide records of beneficiaries' Medicare and Medicaid eligibility, health care utilization and covered health care expenditures.

Most TAP members also acknowledged that the evaluation would require survey data to answer some questions they considered to be important, such as general health status and participant satisfaction. Four of the five TAP members that addressed the issue in their written input stated that the WISP evaluation should include a survey of WISP subjects. The fifth member did not strongly recommend a survey, acknowledging a bias against relying on survey data to measure benefit and work behavior, but did note a survey's value for measuring certain outcomes, such as subjects' understanding of the rules.

One suggestion for follow-up was to assess whether Medicare and Medicaid data provide any useful information on private insurance coverage. This issue was especially important because the CMS actuaries project that WISP cost impacts would heavily depend on Medicare and Medicaid expenditure impacts. Some TAP members suggested that Medicare administrative data could be used to assess whether a person left Medicare or Medicaid for private coverage, which might be identified by a variable indicating the presence of an outside primary payer. However, there was uncertainty over whether the Medicare administrative data include this information. One TAP member suggested consulting with CMS actuaries to design a measurement scheme, given the uncertainty of information on private insurance in administrative records. In written input, two TAP members suggested using administrative data on Medicare matched to survey data on private insurance to explore the issue.

2. Need for Program Fidelity Measures

The TAP members noted the importance of tracking how well WISP is implemented (that is, program fidelity), and they recommended that a combination of quantitative and qualitative data sources be used for that purpose. Measuring program fidelity is critical to an evaluation because poor fidelity can compromise the intervention's ability to influence outcomes. If detected early, poor fidelity can potentially be mitigated by altering program implementation.

The TAP members all agreed that measuring program fidelity was important, though the methods they suggested varied. One TAP member stated that, given the complexity of SSA work rules, it was essential to verify if the treatment group understands the WISP intervention and also desirable to measure how well the control group understands current SSDI work incentives. This TAP member recommended two ways of measuring program fidelity, including to audit "cases with work experience and see if the cases are treated appropriately" and to audit "written material and counseling sessions where participants receive an explanation of WISP." The remaining TAP

members suggested a variety of methods for measuring program fidelity in their written comments. The variation in their comments at least partly reflects that SSA had not specified its operational approach for WISP.[30]

3. Discussion

The TAP's recommendation to adopt a rigorous data collection effort that includes administrative, survey, and qualitative data is consistent with SSA practice in completed and ongoing demonstration projects. Because WISP and BOND are similar in that they both change SSDI beneficiary work incentives, there may be some efficiencies for the WISP demonstration to follow a data collection approach similar to that used for Stage 1 of BOND. In BOND, there are several rounds of qualitative site visits, extensive use of administrative data, and a follow-up survey data collection effort (three years after Stage 1 outreach). In considering the TAP's suggestions for data collection, SSA will need to weigh the advantages of adding data (that is, bringing new information to the evaluation) against the costs of those additional efforts—most notably for a survey. If SSA wants to conduct a full benefit-cost analysis for WISP, at least one follow-up survey would be needed.

F. Timing of Medical CDRs

In addition to the evaluation design topics above, SSA asked the TAP about the timing of the medical CDRs, which relates to the intervention and evaluation design of WISP. The purpose of this discussion was to provide SSA with external perspectives on aspects of WISP intervention components that had not yet been specified. The TAP's recommendations on this topic were meant to be exploratory as currently unknown design and budgetary considerations would ultimately play a major role in the final design of WISP.

SSA asked the TAP for input on SSA's internal policies concerning medical CDRs for SSDI beneficiaries whose benefits are suspended for work under WISP. SSA often defers medical CDRs for those whose benefits are in suspense for work to save administrative resources; if they continue to engage in SGA, their benefits will be terminated anyway. WISP, however, calls for timely medical CDRs for WISP treatment subjects because their benefits cannot be terminated for work. This change is important because if timely medical CDRs are implemented for the treatment group in WISP but not the control group, the impact estimates from the evaluation would not be able to separate out the effects of the WISP work incentive change from the changes in medical CDR policy under WISP.

The TAP was asked to consider three viable medical CDR options for WISP. Option 1 was to have to the treatment and control groups both receive timely medical CDRs. Both groups would

[30] One TAP member made three recommendations concerning program fidelity: (1) design data collection, evaluation, and sample design in ways that least interfere with the way the intervention is intended to be; (2) monitor fidelity for important subgroups; and (3) include data sources for measuring fidelity—administrative data, survey data, and (possibly) SSA staff interview data. A second member stated that the timeliness of benefit suspension and reinstatement is a key fidelity measure. A third member believed that the simplicity of WISP relative to current rules will make fidelity easier to maintain. This member also advised that the fidelity study should attempt to answer three questions: (1) Do claimants understand WISP's implicit incentive structure? (2) Are SSA staff delivering the intervention as intended? and (3) Are WIPA and other support organizations interacting with WISP as intended.

have similar exposure to medical CDRs, which would ensure that the WISP evaluation captures impacts unrelated to the change in medical CDR policy. Option 2 was to have the WISP treatment group only receive timely medical CDRs. This option would require changing current operational policy for medical CDRs for the WISP treatment group, but the control group would continue to have deferred medical CDRs in cases where benefits are suspended for work. The likely result would be that the treatment group would have more terminations for medical recovery than the control group. Under this option, Medicare costs for the control group would be higher than what they would be if SSA did not defer medical CDRs. Option 3 was to proceed with no change in medical CDR policy for the treatment or control group. Operationally, this would be the simplest solution because it does not require changing current operations or increasing resources dedicated to medical CDRs. This option would also likely increase WISP cost estimates because WISP treatment subjects who are deferred would rarely have their Medicare benefits terminated. This option conflicts with the current WISP design, which states that the treatment group should receive timely medical CDRs.

1. Medical CDR Timing Considerations

There was substantial variation in input around the three options. Two TAP members believed that both the treatment and control groups should receive timely medical CDRs (Option 1). One TAP member advised timely medical CDRs for the treatment group only (Option 2), because it maximized the evaluation's internal and external validity. Two other TAP members, however, did not recommend timely medical CDRs for either experimental group (Option 3). Proponents of this approach were concerned that increasing the medical CDR rate of any group would be perceived negatively by advocates and beneficiaries who, in response, might work to actively suppress beneficiaries' work behavior under WISP. The sixth member did not make a specific recommendation but stated that "CDRs should be conducted as they would be conducted for an ongoing program."

2. Discussion

Despite the lack of consensus, the TAP members did mostly agree on one point: SSA should time WISP's medical CDRs in the same way as it would if WISP were a national program. Timing medical CDRs as they would be planned in a national program would ensure that WISP's impact estimates are externally valid. We recommend that SSA decide what medical CDR timing policy it would likely use for a national version of WISP and then use that timing in the demonstration. Any policy on increasing medical CDRs should be made part of a broader operational decision about the SSDI program rather than a specific decision that pertains only to WISP.

G. Outreach

The final area considered by the TAP is how WISP treatment subjects and employment support providers should be informed about WISP and how its features affect benefits. As with the timing of medical CDRs, outreach is a WISP intervention design topic that also has implications for the evaluation. As above, these recommendations should be considered more exploratory given that outreach could ultimately be specified based on currently unknown considerations that could influence the final design of WISP.

For WISP treatment subjects, the TAP considered several outreach issues, including (1) how should treatment subjects be informed about their assignment into WISP, (2) what media should be used for contact with WISP subjects, and (3) whom should treatment subjects contact to ask questions about the intervention. The TAP was also asked whether SSA should provide special

training to employment support providers so that WISP treatment subjects could obtain accurate information and advice from these organizations. Most notably, these include WIPA organizations and TTW providers.

Overall, the TAP recommended that SSA lead proactive outreach to both providers and treatment subjects. This strategy would ensure that the information provided during the demonstration would match that which would be provided in a national program.

1. SSA Should Lead Proactive Outreach to WISP Subjects

Four TAP members provided several recommendations regarding how WISP subjects should receive information about their participation. Though their recommendations varied, the members largely agreed that SSA should lead the outreach effort, that there should be checks to determine whether outreach materials were received and understood, and that treatment subjects should be contacted through multiple media, including mail and phone. Most members believed that SSA should lead the initial outreach because it would help beneficiaries perceive the demonstration as legitimate. Ensuring that outreach materials are received and understood is essential to external validity because treatment subjects who do not learn about or understand the WISP rules are unlikely to alter their behavior in response to it, and may make potentially costly mistakes such as believing they are in their TWP when, in fact, they no longer have a TWP. Using multiple means of contact to inform beneficiaries provides more opportunities for beneficiaries to receive information in a form that is convenient and understandable.

Some members had more recommendations regarding initial outreach. One member stated without elaboration that the outreach effort's design should depend upon the level of randomization that is chosen. A second member outlined a two-step outreach process that could be modified if too costly. In the first step, an outreach letter written in accessible language would be sent to each WISP treatment subject. The letter would be followed by a call from a local employment support service provider, such as a WIPA counselor. The member believes that this level of outreach is necessary if WISP is to have any significant impact on work behavior. However, the member suggests that service providers could alternatively reach out to WISP subjects using a letter, presumably if calling each WISP subject is cost prohibitive. A third TAP member outlined another outreach approach, which includes some features not found in the other members' recommendations. The member advised that SSA send an initial outreach letter to each subject, but the initial letter would be followed by reminder letters that should "be sent semi-annually to those not participating." This member also suggested using webcasts and teleconferences to inform WISP subjects about the intervention and to provide a platform for answering frequently asked questions.

Some TAP members suggested options that would allow WISP treatment subjects to ask questions about ongoing services. Specifically, some members recommended that WISP treatment subjects should be provided with a toll-free number that they could call to ask questions. Opinions varied, however, regarding what training the phone bank staff should receive. One member suggested that the phone bank could be staffed by personnel exclusively devoted to answering questions from WISP treatment subjects, whereas another member suggested that the phone bank staff could potentially be trained to inform WISP subjects as well as other SSI and SSDI beneficiaries. There was also a recommendation that WISP subjects should be able to contact WIPA counselors to ask questions (if one exists after the planned WIPA changes on June 30, 2012). However, the final member who commented seemed concerned that too many opportunities to ask questions might compromise the study's external validity. Consequently, the member suggested that

WISP subjects contact whomever SSDI beneficiaries currently contact to ask benefit and work incentive questions.

2. Outreach to Other Employment Support Providers

Because many non-SSA service providers such as WIPA and TTW providers have the potential to interact with WISP subjects and, per some TAP recommendations, might provide additional outreach, these service providers would need training on WISP. Two TAP members made conventional (and cost-effective) training recommendations. Because the intervention is relatively easy to understand, one member believed a webcast explaining the intervention is all the training the service providers would need. The other member suggested an approach similar to treatment subject outreach in which the service providers are sent outreach materials and provided with a phone number in case they have questions. To avoid internal validity issues, the member also raised the possibility of asking the service providers to refer all treatment subjects to better WISP information sources, such as SSA staff or a WISP hotline.

One TAP member with extensive experience in the local program administration recommended a three-step outreach/training approach to local employment support service providers. This member believed that buy-in from these local providers is essential if the intervention is going to have any effect because organizations who do not feel that they were consulted might actively attempt to undermine the demonstration. In the first step, the member recommends that the organizations' state and local leadership receive personalized outreach and be provided with an opportunity to provide feedback. The second step entails conducting webinars or video conferences capable of informing most ser

vice provider staff about the intervention. The final step would be to send WISP trainers to local offices, where they would conduct training and answer questions in person.

Two members commented that it is very important that non-SSA service providers have the ability to identify WISP treatment subjects who do not self-identify. One member thought that site level randomization would assist service providers in determining whether most of their clients are WISP treatment subjects. This member also hoped that WISP subjects would be able to self-identify their status. The other TAP member suggested that WISP subjects should be identified either by the subject himself or herself, a benefits request form, or a claims representative contact.

3. Discussion

The TAP largely agreed that SSA should lead a WISP subject outreach effort that includes multiple forms of contact and attempts to ascertain whether subjects have received and understood outreach materials. The TAP is advocating a more proactive outreach approach for WISP than that used for BOND, for both consumers and providers, though offered multiple ways for which that outreach to occur (e.g., letters versus phone calls and periodicity of follow-up). When considering the size and extent of WISP outreach to both groups, SSA should also weigh whether the information available in the demonstration matches that which would be available in a national program.

V. CONCLUSION

The TAP provided several recommendations for a rigorous evaluation of WISP that would provide an evidentiary basis for fundamental policy reforms. The need for a rigorous evaluation is especially important because WISP makes several fundamental changes to the SSDI program that could be viewed with skepticism by various groups who are concerned about its potential effects. The most notable of the TAP's recommendations is to use an experimental design.

SSA's five original research questions provide a framework for the evaluation of WISP, but TAP members had several additional recommendations that would make the evaluation more comprehensive. The original research questions cover the basic components of an evaluation, most of which can be measured using existing administrative data sources. The TAP identified several outcomes that are not in any administrative data source but could be measured using a follow-up survey. Some possible survey outcomes include job search, health, participant satisfaction with WISP intervention services, other health insurance usage, and consumption.

There was also a consensus that the evaluation track outcomes for at least five years and present findings in a series of reports to policymakers and key stakeholders. While it would be important to produce timely information about WISP's impacts, the TAP cautioned that producing findings too early in the demonstration might provide misleading information. The TAP advised that WISP's impacts on SSA administrative procedures should be reported after the first two years of the project. Other beneficiary outcomes, especially employment and benefit receipt, should not be reported before the third year, because impacts on these key outcomes are likely to take that long to emerge.

In several other areas, including the unit of random assignment, site selection, and sample selection, the TAP did not achieve consensus or make definitive recommendations, partly because of limited information, especially on operational aspects of WISP. Nonetheless, the TAP's deliberations on these issues should be useful to SSA as the agency continues to develop the specific operational components of WISP. For example, the TAP's feedback on evaluation design shows how operational considerations affect the strengths and limitations of various design options. Most notable, if field offices would have primary responsibility for WISP administration and support from other local agencies and organizations would be critical, then random assignment of field office areas might be the best option. Conversely, if field offices and other local entities would play less significant roles, then individual random assignment or a hybrid involving random assignment of areas (not necessarily field office areas) and random assignment of individuals within field office areas would be preferred for reasons of statistical power and costs.

The TAP also weighed in on some components of WISP that have implications for the intervention design, including the timing of medical CDRs and outreach to beneficiaries and providers. The TAP members think that beneficiary understanding of WISP is essential to the external validity of the evaluation, and most also argued that SSA should lead the outreach effort. Though the TAP had several suggestions how to address these issues, a guiding principle behind their recommendations is that WISP should be designed to match SSA's expectations for a national program as closely as is feasible. This design strategy would enhance the external validity of the demonstration.

In summary, the TAP made multiple recommendations about the evaluation design but did not make recommendations in some important areas where the optimal evaluation design is dependent on operational design. The TAP did, however, provide information that will help SSA understand the consequences of various operational design options for the evaluation and will help guide development of the evaluation design as SSA makes decisions about the operational design.

REFERENCES

Andrews, Kristin L., Robert R. Weathers II, and Su Liu. "How do Medicaid Buy-in Participants Who Collect Social Security Disability Insurance Benefits use SSA Work Incentive Programs." Work and Insurance Issue Brief Report No. 7. Washington, DC: Mathematica Policy Research, December 2007.

Bell, Stephen H., Daniel Gubits, David Stapleton, David Wittenburg, Michelle Derr, Arkadipta Ghosh, and Sara Ansell. "BOND Implementation and Evaluation. Evaluation Analysis Plan. Final Report Submitted to Social Security Administration." Cambridge, MA: Abt Associates, March 2011.

Liu, Su and David C. Stapleton. "How Many SSDI Beneficiaries Leave the Rolls for Work? More than You might Think." Disability Policy Research Brief Report No. 10-01. Washington, DC: Center for Studying Disability Policy, April 2010.

Livermore, Gina A. "Wage Reporting and Earnings-Related Overpayments in the Social Security Disability Programs: Status, Implications, and Suggestions for Improvement." Washington, DC: Ticket to Work and Work Incentives Advisory Panel, May 2003.

Livermore, Gina, David Stapleton, and Allison Roche. "Work Activity and use of Employment Supports Under the Original Ticket to Work Regulations: Characteristics, Employment, and Sources of Support among Working-Age SSI and DI Beneficiaries." Washington, DC: Mathematica Policy Research, 2009.

Michalopoulos, Charles, David C. Wittenburg, Dina A. R. Israel, Jennifer L. Schore, Anne W. Warren, Aparajita Zutshi, Stephen Freedman, and Lisa K. Schwartz. "The Accelerated Benefits Demonstration and Evaluation Project: Impacts on Health and Employment at Twelve Months. Volume 1." New York, NY: MDRC, February 2011.

Peikes, Deborah N., Sean M. Orzol, Lorenzo Moreno, and Nora A. Paxton. "State Partnership Initiative: Selection of Comparison Groups for the Evaluation and Selected Impact Estimates." Princeton, NJ: Mathematica Policy Research, October 2005.

Porter, A., J. Smith, A. Payette, T. Tremblay, and P. Burt. "SSDI $1 for $2 Benefit Offset Pilot Demonstration: Vermont Pilot Final Report." *Montpelier: Vermont Division of Vocational Rehabilitation,* 2009.

Rangarajan, Anu, Thomas M. Fraker, Todd C. Honeycutt, Arif A. Mamun, John Martinez, Bonnie L. O'Day, and David C. Wittenburg. "The Social Security Administration's Youth Transition Demonstration Project: Evaluation Design Report. Final Report." Submitted to the Social Security Administration, Office of Program Development and Research. Washington, DC: Mathematica Policy Research, Inc., January 2009.

Schimmel, Jody; David C. Stapleton; and Jae G. Song. How Common is "Parking" among Social Security Disability Insurance Beneficiaries? Evidence from the 1999 Change in the Earnings Level of Substantial Gainful Activity. Social Security Bulletin, Vol. 71 No. 4, 2011.

Social Security Administration. "Annual Statistical Report of the Social Security Disability Insurance Program." Baltimore, MD: Social Security Administration, 2010.

Stapleton, David C., Bonnie L. O'Day, Gina A. Livermore, and Andrew J. Imparato. "Dismantling the Poverty Trap: Disability Policy for the 21st Century." *The Milbank Quarterly,* vol. 84, no. 4, 2006, pp. 701-732.

Todd, P. E. and K. I. Wolpin. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *The American Economic Review,* vol. 96, no. 5, 2006, pp. 1384-1417.

APPENDIX A: TAP MEMBER INPUT FORMS [1]

[1] Mathematica sent these forms to the TAP members to facilitate their written feedback. Six of the TAP members returned the form. The form includes a general comments section and input for each of the topics covered in the briefing materials.

TAP Member #1

General Comments

Which items are most important to the WISP evaluation?

The most important things are as follows:

- Make certain that we obtain unbiased estimates of the key treatment effects, and that the sample sizes are large enough that we are likely to get statistically significant estimates if the treatment has the desired effect
- Make sure that the process study documents if the treatment group understands the rules under WISP

Were there any key topics not covered in the brief materials and the questions below that you would like to raise here?

No.

Other comments?

No.

Research Questions

1.1 Do the five questions cover all the outcomes of interest to evaluators and SSA?

No in the sense that they are too high level. It would be useful to know the effects for subgroups, such as new v. prior beneficiaries, concurrent v. SSDI only, and by age. Also, although the background document deals with the issue of how WISP affects well being of the participants, that issue is not covered in the five questions.

1.2 Which additional outcomes should be incorporated into the research questions? Should any of those questions be given more weight than the five core questions in the original solicitation?

As noted in the background document, WISP will interact with several other programs that could result in additional costs or benefits to the government. To the extent possible, the evaluation should also look at usage and costs of these programs such as Medicaid, supplemental security income, and the ticket to work program.

Dissemination/Policy Relevance Guidelines

2.1 When does the evaluation need to produce key findings to policymakers?

I am puzzled by this question. Without intending to be flippant, the evaluation should provide findings to policymakers when they are available. There are several dangers in presenting findings too early. First, it is possible that longer-term findings may reverse earlier findings, so caution should be exercised in presenting short-term findings. Two recent Mathematica Policy Research projects for the U.S. Department of Labor had important policy findings reversed when a longer-term follow-up evaluation was completed (The Job Corps evaluation and the ITA experiment).

Other than the caveat above, results should be made available as they become available.

2.2 What should the follow-up period be on those findings?

The rules being changed by WISP take place over a lengthy period. For example, eligibility provisions vary over many years under current rules, with the trial work period, extended period of eligibility, and expedited reinstatement running for up to five years. Moreover, it may take time for the system under WISP to run smoothly and for beneficiaries to understand the rules and trust the government. Unfortunately, I do not know about the timing of when beneficiaries leave the program, but I imagine that a follow-up period of at least five years would be useful.

2.3 Will other stakeholders demand findings and information that policymakers would consider less important?

No expertise to answer.

Dissemination/Policy Relevance Guidelines

2.4 What information would other stakeholders finding interesting and how should it be disseminated to them?

No information to answer.

2.5 What short run and long run findings would be considered "convincing"?

Convincing to whom? Congress would be most interested in saving benefit and administrative funds. I imagine that savings in a ten-year period or less would be most important as that is how budget scenarios are sometimes presented.

Evaluation Design

3.1 Should WISP use an experimental or non-experimental design?

It is very important that an experimental design be used to test the primary hypotheses. The research community is divided on how well non-experimental methods perform, but even those who have some faith in non-experimental methods would agree that experimental designs provide stronger and less contestable evidence. For recent relatively positive assessments of the ability of nonexperimental assessments to replicate the findings from RCTs, see Cook, Shadish and Wong (2008); Dehejia and Wahba (2002); and Heckman, Ichimura, and Todd (1998). Studies and reviews that are more cautionary include Bloom, Michalopoulos, Hill, and Lei (2002; Glazerman, Levy, and Myers (2003); Smith and Todd (2005); and Wilde and Hollister (2007).

There are several problems with using non-experimental methods here. First, SSA has already used experimental methods to test some important concepts, so it appears feasible to use RCTs. The burden would be on SSA to explain why this approach would not be used to test WISP. Second, the assumptions required for approaches such as propensity score matching and instrumental variables are quite strong and generally not testable. Thus, stakeholders who do not like the findings could always challenge them. Third, the non-experimental design that is preferred, regression discontinuity design, does not appear feasible in this situation.

Howard Bloom, Charles Michalopoulos, Carolyn Hill, Ying Lei (2002). *Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Program?* New York: Manpower Demonstration Research Corporation.

Thomas D. Cook, William R. Shadish, and Vivian C. Wong (2008). "Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management*. 27(4): 724-750.

Rajeev H. Dehejia and Sadek Wahba(2002). "Propensity Score-Matching Methods For Nonexperimental Causal Studies." *The Review of Economics and Statistics*. 84(1): 151-161.

Steven Glazerman, Dan M. Levy, and David Myers (2003). "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *The ANNALS of the American Academy of Political and Social Science. 589(1): 63-93.*

James J. Heckman, Hidehiko Ichimura, and Petra Todd (1998). " Matching As An Econometric Evaluation Estimator." *Review of Economic Studies*. 65(2): 261-294.

Jeffrey A. Smith and Petra E. Todd (2005). "Does Matching overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125: 305-353.

Elizabeth Ty Wilde and Robinson Hollister (2007). "How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment." *Journal of Policy Analysis and Management*. 26(3): 455-477.

3.2 If a non-experimental design should be used, then what type of design offers most promise for WISP design specifically?

I do not see any that I would advocate. The large number of beneficiaries makes propensity score matching potentially useful if the evaluation would have access to all the data on beneficiaries.

Evaluation Design

3.3 What are the implications for the choice of design on both internal and external validity?

The only way to assure internal validity is with an experimental design. External validity is more of a problem, even with an experimental design. One threat to external validity is if the study population is not representative of all types of beneficiaries. Another possible threat is if the evaluation takes place in a limited number of local offices, they may not be typical of all offices in the country. Perhaps the greatest threat to external validity is that unless the experiment operates for a long time, there will not be a large range in national economic conditions. To some extent this can be overcome by including areas with a range of unemployment rates.

Evaluation Design

3.4 If random assignment is used, should it be individual or site level?

I have limited background on this issue, so I would defer to other panel members. My impression is that the evaluation would have the most power if random assignment is at the individual level to reduce clustering problems. However, because staff in participating local offices must be trained, it might be more economical to perform random assignment at the local office level. Basically, there are tradeoffs between cost and statistical efficiency.

3.5 Given the selected design, which evaluation components should be emphasized over what time frames?

Both the impact and process studies are important. Given the complexity of social security programs and potential mistrust of beneficiaries, it is crucial that the process study address the issue of whether participants understand the treatment under WISP and the implications of the new provisions for their benefits.

The impact evaluation is essential to the evaluation. The key issues are the impacts of WISP on benefits, work, and administrative costs. It is important to learn whether increased work effort and reduced benefits are a long-term or short-term phenomenon. I do not know the precise period that must be studied, but my conjecture is that the participants should be tracked for at least five years. (This need not be expensive as SSA has access to administrative earnings and benefits data.)

Sample

4.1 What should the recommended sample size be (relates to the demonstration design)?

I do not have sufficient information to calculate this.

4.2 Should subgroups be targeted along the work path? If so, are the subgroups originally suggested by OMB and SSA appropriately selected?

The background paper provided to the TWG described five subpopulations of special interest. I do not have enough background to select the optimal representation of these groups, but if groups with certain characteristics, e.g., have engaged in some work activity subsequent to qualifying for SSDI, are most likely to respond to the WISP incentives, then it could be useful to oversample them. On the other hand, an important goal is to estimate the steady state costs and benefits of WISP, and the proportion of people who fall into various categories might change as a result of WISP. Thus, the safest strategy for estimating the national total savings would be to track new beneficiaries for an extended period. It could be argued that one could oversample certain groups and then use weighting to get national estimates, but such a procedure might miss changes in the mix of subgroups.

4.3 What should the minimum detectable effects be set for producing key policy impacts? What levels of statistical power would be recommended?

I do not have sufficient information to answer this question at this time.

Sample

4.4 What are the implications of SSA subsample division on the demonstration's external validity?

In theory, one can achieve external validity so long as all groups are included even if groups of particular interest have a higher probability of being sampled. As I noted above, my concern with this approach is that it is possible that the appropriate weights in a pre-WISP situation may not be appropriate once WISP is available, i.e., the proportion of beneficiaries with various post-qualification work experiences might change so that the prior weights are no longer valid.

4.5 What are the implications of SSA's age restriction for the evaluation?

The background paper notes that SSA is considering restricting WISP to those under age 55. This is presumably because those over 55 have a shorter time period to gain from going back to work and the incentives for them to return to work are not very strong in any event. The main thing sacrificed by not including people 55 and over is external validity--the evaluation will not provide any information about how older beneficiaries respond to the WISP incentives. If SSA strongly believes that such beneficiaries are unlikely to benefit from WISP, then this would not be a serious loss.

4.6 What other subgroups should be targeted (e.g., SSI recipients and those in the Medicaid buy-in)

I do not know enough about the interactions of these programs to say much. However, it is clear that those who receive both SSDI and SSI do face different incentives than those on SSI alone, so it would likely be useful to target such individuals for the demonstration. The Meidcaid buy-in group also might be of interest. A group that sounds interesting to me is individuals who have worked with the Ticket to Work program as they have demonstrated an interest in returning to work.

Data Collection

5.1 What data sources are most important for addressing research questions?

WISP is fortunate in that the sponsoring agency controls the most important data bases. The key outcomes—earnings, employment, SSDI and SSI benefits—and key background variables on impairments and personal characteristics are all available from SSA. Data on Medicare and Medicaid usage, vocational rehabilitation services, and are available from other federal agencies. Thus, the evaluation could be done without a survey. One or more surveys would be useful for the following purposes:

- To obtain additional background and demographic information not available from SSA
- To obtain additional information about the disability and impairment status
- To obtain an alternative source of earnings information
- To obtain family income information
- To ask questions about work plans and capabilities
- To ask about how well the treatment group understands WISP
- To ask about rehabilitation efforts and job search

I think it is essential to verify if the treatment group understands the WISP rules and the control group understands about current law, and desirable to learn how the control group understands current law. Earnings information varies when using surveys instead of administrative data, and there have been occasions where, surprisingly, the impact estimates are markedly different when another source is used.

5.2 Which administrative data sources can be leveraged and what are their advantages?

As noted above, SSA program and earnings data are essential, as are Medicare and Medicaid data.

5.3 Will survey data be needed?

I think it would be very unwise not to collect survey data at baseline, and it would be useful as well for follow-up. There is no substitute to find out if the participants understand the program, and the data will be useful for other purposes as well, as noted above.

Data Collection

5.4 When should outcomes in the data be measured?

I am not an expert in this area, but it is important to track outcomes for at least five years. It may take time before WISP participants risk going into the labor market under the new rules, and it is critical to learn if any increases in work and decreases in payments persist.

5.5 How should savings to Medicare and private health insurance enrollment be measured?

This is tricky, and the actuaries might know how to do this best. Using reductions in actual expenditures is risky because a few chance outliers could drive the results. On the other hand, average cost is not a good measure because people with higher expected costs are likely to be more risk averse about the possibility of losing their Medicare. Although I do not have the answer here, I can see that it will not be simple to address.

5.6 How should demonstration fidelity be measured?

There are several approaches that could be used. One approach is to audit cases with work experience and see if the cases are treated appropriately. A second important aspect is to audit written material and counseling sessions where participants receive an explanation of WISP. All of these approaches should be used.

Site Selection

6.1 What constitutes a site?

A site would be defined as a local office where people are randomly assigned. In the extreme case, where random sampling is done throughout the nation without primary sampling units (PSUs), then all local offices would be "sites."

6.2 How many sites should be selected?

There is a tradeoff here. As the number of sites increases, costs increase and precision increases. Holding costs constant, means that as the number of sites increases, the total number of participants must decrease. The right decision depends on fixed cost per sit, cost per participant, and the effect of clustering on precision.

6.3 What programs are most likely to interact with WISP?

I am not an expert on this, but relevant programs (in addition to SSDI) include Medicaid, Ticket to Work, SNAP and other welfare programs, and vocational rehabilitation.

Site Selection

6.4 Should/Could interstate program variation be exploited to minimize these interactions?

I am not sure what the question means. It would be useful to include states with program variations to better obtain national estimates for the impact of WISP. If results vary by state, we would need to compute a weighted average for a national estimate.

Timely Medical CDRs

7.1 Should WISP conduct timely medical CDRs for the treatment and/or control/comparison groups?

As opposed to untimely CDRs? I think the CDRs should be conducted as they would be conducted for an ongoing program and the participants should be told what they would hear for an ongoing program. To make the CDRs timely for the demonstration if they are not done on a timely basis ordinarily might tell us the impact of a program package not being implemented.

Information Dissemination and Training

8.1 What should the initial and primary medium(s) of contact be with WISP subjects?

This is not an area where I have any knowledge.

8.2 Who should treatment subjects contact to ask questions about the intervention?

I would imagine they should be contacting whoever they would contact for any program issues.

8.3 Should there be checks on whether WISP subjects received or understood the outreach materials that they were sent?

It is essential that we learn if participants received and understand the material on WISP. This is so that you know if you are evaluating the intended program.

Information Dissemination and Training

8.4 Does the TAP have any recommendations regarding training for WIPA organizations, TTW, providers, or any other entities that will interact with WISP subjects?

I do not.

8.5 How will these entities identify WISP subjects?

NA

External Validity

9.1 What would be the key features of a national version of WISP?

Coverage should include the entire nation, and eligibility should be specified however SSA wants it.

9.2 How will the TAP's other recommendations affect the evaluation's external validity, especially in relation to a national version of WISP?

This can only be answered if one knows all the TAP responses. I have noted a few places where there is a tradeoff for external validity, and these primarily involve oversampling of groups expected to be most responsive to WISP.

TAP Member #2

General Comments

Which items are most important to the WISP evaluation?

Did the work behaviors of the participants a VW ISP program change?

Did that participation in private health insurance increase for the WISP program participants?

Did the elimination of the trial work period TWP discourage participants in the WISP program?

Were there any key topics not covered in the brief materials and the questions below that you would like to raise here?

Other comments?

Research Questions

1.1 Do the five questions cover all the outcomes of interest to evaluators and SSA?

It is critical that the evaluation determine changes in benefits and in earnings for the participants in the WISP program as this is the whole point of establishing the program.

It will also be important to measure the effect on Health Insurance and general health of the participants. As there is a general assumption that the health of working people is better than people who are not employed.

It will also be critical to evaluate whether or not participation in the WISP program affects the number of overpayments.

1.2 Which additional outcomes should be incorporated into the research questions? Should any of those questions be given more weight than the five core questions in the original solicitation?

General satisfaction from WISP participants.

Dissemination/Policy Relevance Guidelines

2.1 When does the evaluation need to produce key findings to policymakers?

It will be important for the pilot to produce findings for policymakers within the first year. Such things as reduced overpayments and administrative burdens should be able to be measured during the first year and presented to policy makers. Outcomes related to benefits and earnings will likely not be available until later in the project.

2.2 What should the follow-up period be on those findings?

The follow-up period should be indefinite.

2.3 Will other stakeholders demand findings and information that policymakers would consider less important?

Other stakeholders will demand information on benefit interruptions and benefit reductions.

Dissemination/Policy Relevance Guidelines

2.4 What information would other stakeholders finding interesting and how should it be disseminated to them?

How many people were using the TWP before the pilot. How many people would have used the TWP If they were not in the pilot.

What was the reduction in overpayments?

2.5 What short run and long run findings would be considered "convincing."

The short-term findings that would be considered convincing would include,

1. **A reduction in overpayments**
2. **A reduction in administrative burdens**
3. **More people working.**

The long-term findings that would be considered missing would include,

1. **More people working**
2. **A reduction in SSDI payments**

Evaluation Design

3.1 Should WISP use an experimental or non-experimental design?

A combination of the two.

3.2 If a non-experimental design should be used, then what type of design offers most promise for WISP design specifically?

A number of field offices should be selected randomly. The consumers of those field offices should be selected randomly.

3.3 What are the implications for the choice of design on both internal and external validity?

Evaluation Design

3.4 If random assignment is used, should it be individual or site level?

3.5 Given the selected design, which evaluation components should be emphasized over what time frames?

Sample

4.1 What should the recommended sample size be (relates to the demonstration design)?

4.2 Should subgroups be targeted along the work path? If so, are the subgroups originally suggested by OMB and SSA appropriately selected?

Yes, I believe subgroups should be targeted along the path of work. I also believe that the subgroups originally OMB and SSA are appropriate.

4.3 What should the minimum detectable effects be set for producing key policy impacts? What levels of statistical power would be recommended?

Sample

4.4 What are the implications of SSA subsample division on the demonstration's external validity?

4.5 What are the implications of SSA's age restriction for the evaluation?

4.6 What other subgroups should be targeted (e.g., SSI recipients and those in the Medicaid buy-in)

A definitely think that the subgroup of those people who are in the Medicaid Buy-In should be targeted.

Data Collection

5.1 What data sources are most important for addressing research questions?

5.2 Which administrative data sources can be leveraged and what are their advantages?

5.3 Will survey data be needed?

Data Collection

5.4 When should outcomes in the data be measured?

5.5 How should savings to Medicare and private health insurance enrollment be measured?

5.6 How should demonstration fidelity be measured?

Site Selection

6.1 What constitutes a site?

6.2 How many sites should be selected?

6.3 What programs are most likely to interact with WISP?

Medicaid, SSI, Medicare, WIPA and housing.

Site Selection
6.4 Should/Could interstate program variation be exploited to minimize these interactions?

Timely Medical CDRs
7.1 Should WISP conduct timely medical CDRs for the treatment and/or control/comparison groups? **Participants in the WISP program should have CDRs performed at the same rate and times as people are not participating and WISP program.**

Information Dissemination and Training

8.1 What should the initial and primary medium(s) of contact be with WISP subjects?

Participants in the program should have their first contact come from SSA.

8.2 Who should treatment subjects contact to ask questions about the intervention?

8.3 Should there be checks on whether WISP subjects received or understood the outreach materials that they were sent?

Yes

Information Dissemination and Training

8.4 Does the TAP have any recommendations regarding training for WIPA organizations, TTW, providers, or any other entities that will interact with WISP subjects?

8.5 How will these entities identify WISP subjects?

External Validity

9.1 What would be the key features of a national version of WISP?

9.2 How will the TAP's other recommendations affect the evaluation's external validity, especially in relation to a national version of WISP?

TAP Member #3

General Comments

Which items are most important to the WISP evaluation?

I will interpret "important" here as referring to the amount of policy-relevant knowledge that the evaluation generates. The most important aspects of the evaluation (in my perceived order of importance from most to least) are then:

1. Identification strategy – i.e. random assignment if possible.

2. External validity – e.g. focusing on the flow rather than the stock, choosing the experimental sites in a reasonable way, and making sure that treatment group members understand the treatment at a level similar to what would be the case under a full scale permanent implementation of WISP.

3. Long-term follow-up over many years after random assignment.

4. Collection of data on outcomes not present in the administrative data.

5. Undertaking parallel non-experimental research.

Were there any key topics not covered in the brief materials and the questions below that you would like to raise here?

The briefing document did not indicate the duration of the WISP treatment – i.e. how long the WISP treatment group members would be governed by the WISP rules. Note that the duration of the intervention is distinct from the duration of the evaluation. The duration of the treatment is an important aspect of the treatment as it affects the return to making investments related to a return to work. The duration of the treatment is also important in ensuring that any impacts represent long-term employment effects rather than just movements in labor market activity over time to respond to short-run changes in incentives. There is a nice literature on this issue associated with the design of the U.S. negative income tax experiments back in the 1970s. See e.g. the book on social experimentation by Hausman and Wise.

If an experimental design is selected, there is likely great value in using some sort of stratified randomization scheme in the site selection. There are two issues here. First, randomizing within matched pairs or matched groups of sites increases the power of the analysis without having to increase the sample size by improving covariate balance between the treatment and control groups. Second, there may be interest in over-sampling sites with particular characteristics, such as high staff-to-claimant ratios, quality and quantity of state VR services, being in states with particular forms of ACA implementation or states with particular economic conditions. In regard to the latter, a potential issue with external validity concerns the fact that the evaluation will likely start during the middle of a rather hesitant aggregate economic recovery. We know from the work of David Autor and Mark Duggan that the business cycle affects the nature and magnitude of flows onto SSDI. Stratifying on where the states containing sites are relative to the business cycle would facilitate a subgroup analysis based on local economic conditions. Any sort of stratification can, of course, be undone with weights at the end of the day.

A full cost-benefit analysis, from both individual and social perspectives, is a key component of an evaluation of WISP.

I strongly encourage SSA to have two parallel non-experimental evaluations alongside an experimental evaluation of WISP, should an experimental design be adopted for the evaluation. Learning about what non-experimental methods (if any) work given the data generally available in the SSDI program context is a potentially very useful side benefit to doing an experimental evaluation. The first of the non-experimental evaluations would look at standard partial equilibrium reduced-form methods such as difference-in-differences and matching at the site level. The second would follow along the lines of the Todd and Wolpin (2005) American Economic Review paper that applied structural methods to the control group data from the Mexican PROGRESA experiment and then used the estimated structural model to forecast the impacts of that intervention. The experimental treatment group outcomes were then used to

evaluate the forecasts from the structural model.

Provision should be made in the evaluation contract for the production, documentation and dissemination of data from the evaluation for use by other researchers, both to allow replication of the official evaluation and, more importantly, for use in additional analyses. These additional analyses are likely to have policy relevance for SSA and so represent additional value to be obtained from the money spent on the evaluation.

Other comments?

How to estimate the induced entry effects (if any) that result from WISP is quite a challenge. One strategy, inspired by the design of the Self-Sufficiency Project (SSP) in Canada, would conduct a separate randomization of new UI claimants with the treatment group subject to WISP should they apply for SSDI and the control group not eligible. The SSP evaluation did a similar sub-analysis in which new social assistance claimants were randomized to be eligible or not for the SSP treatment, which began only after 12 months of continuous social assistance receipt. That study found modest but non-zero effects of SSP eligibility on the probability of staying on social assistance for at least 12 months. A second way to get at (some of the) induced entry effects would be to focus the flow sample (if there is one – see my other comments) on new SSDI applicants rather than new SSDI beneficiaries. Evidence of higher or lower rates of transiting from application to benefit receipt in the treatment group relative to the control group would be evidence of induced entry effects.

It would be really useful to have someone, ideally one or more economists closely familiar with the relevant literatures (e.g. John Bound or John Rust) sit down and think about all of the incentives for particular types of behavior that are implicit in WISP, both in an absolute sense and relative to the existing program rules. The results of that thinking would make a very useful paper that could be one product of the evaluation. It could also help to inform SSA's and the research community's priors about the likely effects of WISP on labor market and program-related behavior.

As noted elsewhere in my responses, it is very important to the external validity and general usefulness of the evaluation that WISP treatment group members understand the intervention. This concern suggests the possibility, in the context of an evaluation based on the random assignment of sites to WISP, of doing a cross-cutting random assignment of an intensive information provision treatment.

WISP embodies a 100 percent implicit tax rate on earnings in a given month that exceed the SGA level. This is a high tax rate relative to a fairly low SGA level. This suggests the potential value of cross-cutting randomization of the BOND treatment within the WISP evaluation.

Research Questions

1.1 Do the five questions cover all the outcomes of interest to evaluators and SSA?

No. Among the other outcomes of potential interest are claimant health and claimant consumption. But the five questions do capture the major outcomes of policy interest.

1.2 Which additional outcomes should be incorporated into the research questions? Should any of those questions be given more weight than the five core questions in the original solicitation?

I think there are a number of additional outcomes that would be of interest in an evaluation of WISP:

a. Self-reported health. This is of interest both in its own right, and because it would be useful to try and estimate the impact of WISP on Quality Adjusted Life Years (QALYs). Collecting self-reported health data would also help with the fact that Medicare and Medicaid administrative data will miss some claimants in both the treatment and control groups who leave the program.

b. Human capital related activities such as use of a ticket to work, participation in vocational rehabilitation and participation in other sorts of employment and training programs. Some of these could be measured using administrative data, e.g. ticket use, others would require survey data.

c. Consumption. A number of studies have found that individuals report consumption differently than income. For example, the book "Making Ends Meet" by sociologists Kathryn Edin and Laura Lein plays off the fact that AFDC recipients report consumption more accurately than income. Perhaps SSDI claimants do the same? If so, impacts on consumption may capture impacts on components of earnings missed by administrative data sources.

d. WISP provides an incentive for claimants to have employers pay them infrequently. For example, a claimant who earns SGA+1 every month for a year will receive zero SSDI payments under WISP while a worker who earns (12 x SGA + 1) in one month of a year will receive 11 months of SSDI payments. It would be useful for the evaluation to be able to try and measure possible strategic responses by claimants to this incentive structure. Doing so requires earnings information at a finer level of temporal detail than the annual data available in SSA administrative records.

f. Measures of job search intensity (e.g. time spent in job search activities) and mode (e.g. on-line, sending resumes to employers known to hire the disabled, newspaper want ads, etc.).

g. Amount of benefit counseling received from SSA and from non-SSA sources. WISP may reduce the counseling burden on both SSA staff and on others. These reductions are a potentially important part of a full social cost-benefit calculation.

h. Claimants' opinions about the design and effectiveness of the WISP treatment. Staff opinions in this regard would be useful as well.

I also have a comment on one of the other outcomes suggested in the briefing document: Bullet 4 on page 18 suggests using SSDI re-application and re-entry as outcomes. There are important dynamic selection issues with these outcomes. See e.g. the Eberwein, Ham and LaLonde (1997) Review of Economic Studies paper which considers the conceptually identical issue of estimating impacts on employment and non-employment spell durations in experimental evaluations of active labor market programs.

Dissemination/Policy Relevance Guidelines

2.1 When does the evaluation need to produce key findings to policymakers?

Because the behavior of interest in this context evolves relatively slowly over time, and because, as noted elsewhere, the flow sample is of particular policy interest, the most important results from the evaluation will take time to arrive. While electoral imperatives necessarily drive politicians to have short time horizons, given the importance of the SSDI program both in terms of budget and in terms of the number of relatively vulnerable individuals it serves, getting the correct answer, rather than getting the quick answer, surely is the social optimum.

2.2 What should the follow-up period be on those findings?

The follow-up period should be indefinite or until WISP (or some similar reform) is implemented throughout the program, whichever is shorter. As noted in my response to question 2.1, the outcomes of interest in this context take time to evolve. Expectations (about how long the program will really last) and complementary investments (in e.g. vocational training) loom large in this context as well, which further increases the value of long-term follow-up. Of course, interim impact estimates should be produced and disseminated at regular intervals.

2.3 Will other stakeholders demand findings and information that policymakers would consider less important?

I am not qualified to remark on the preferences of stakeholders other than policymakers, the research community and the taxpayer; my assumption is that this question has in mind the advocacy community, about which I know little. My other comments cover the findings and information relevant to policymakers, the research community and taxpayers.

2.4 What information would other stakeholders finding interesting and how should it be disseminated to them?

I am not qualified to remark on the preferences of stakeholders other than policymakers, the research community and the taxpayer; my assumption is that this question has in mind the advocacy community, about which I know little. My other comments cover the findings and information relevant to policymakers, the research community and taxpayers. To the extent that the support of advocacy groups for serious, policy-relevant research can be obtained at relatively low cost by providing them with information of interest that should surely be done.

2.5 What short run and long run findings would be considered "convincing."?

Findings are convincing when they emerge from an evaluation that has a strong design (e.g. random assignment), precise estimates, high internal and external validity, and high implementation fidelity.

Evaluation Design

3.1 Should WISP use an experimental or non-experimental design?

My recommendation is that WISP should use an experimental design. Experimental designs are not without issues – see e.g. the discussion in Heckman and Smith (1995) Journal of Economic Perspectives – but they provide the clearest and most compelling evidence of policy impact in most circumstance and would do so, in my view, in this circumstance.

3.2 If a non-experimental design should be used, then what type of design offers most promise for WISP design specifically?

Assuming that sites would remain the unit of treatment in a non-experimental evaluation, there are two natural non-experimental designs.

The first non-experimental design would be a differences-in-differences design in which the before-after changes in outcomes of treated and untreated sites are compared. This could be done within a standard panel data framework that would allow the analysis to control for time varying covariates such as local economic conditions and claimant characteristics.

The second design would compare the selected treated sites to matched untreated sites. The matching would take place on site-level characteristics such as staffing levels, local economic conditions, state vocational training quality and quantity, state ACA implementation and site claimant characteristics. An important consideration in such a design is making sure that not all sites with a particular relevant characteristic (e.g. high staffing levels relative to claimant volume) are assigned to treatment, so that some remain for potential inclusion in the comparison group.

3.3 What are the implications for the choice of design on both internal and external validity?

Using an experimental design is first and foremost about internal validity. Because an experimental design assigns treatment randomly to either individuals or sites, it removes the possibility of non-random selection into treatment as a function of observed or unobserved covariates that also affect outcomes. For this reason, experimental designs provide, in general, clear and compelling causal evidence.

The most common disruptors of internal validity in experimental evaluations, namely treatment group dropout and control group substitution – see e.g. Heckman, Hohmann, Smith and Khoo (2000) Quarterly Journal of Economics – are not relevant here because the treatment (and the control) states are both budget constraints (rather than, say, job training).

The most common disruptor of external validity in experimental evaluation is non-random site selection due to a requirement of using only volunteer sites. This issue is common to many of the experimental evaluations performed by the Institute of Education Sciences and also affected the experimental evaluation of the Job Training Partnership Act funded by the Department of Labor. Apparently, SSA is not constrained to use only volunteer sites, so this concern with experimental designs is not an issue in this context.

Evaluation Design

3.4 If random assignment is used, should it be individual or site level?

I recommend that the evaluation use site-level random assignment (and, indeed, that it use sites as the units of treatment even in a non-experimental evaluation). The key argument, in my view, for preferring sites to individuals as the unit of treatment is that having sites as the unit of treatment means that both the site SSDI staff as well as local support organizations will have a strong incentive to develop real expertise regarding the WISP treatment. In contrast, with individuals as the unit of treatment most sites will not have all that many claimants under WISP and so will have a limited incentive to develop such expertise. Thus, with sites as the unit of treatment the environment experienced by treated individuals will better approximate what SSDI would be like with WISP as the ongoing policy for all claimants; put differently, using sites as the unit of treatment should enhance the external validity of the evaluation.

It should be noted that all sites will have to have at least some WISP expertise even when using sites as the unit of treatment in order to handle WISP treatment group members who move.

3.5 Given the selected design, which evaluation components should be emphasized over what time frames?

See my responses to the other questions.

Sample

4.1 What should the recommended sample size be (relates to the demonstration design)?

The sample size should be sufficient to obtain estimates of the desired power for the smallest subgroup.

4.2 Should subgroups be targeted along the work path? If so, are the subgroups originally suggested by OMB and SSA appropriately selected?

The subgroup question is an interesting and important one. The discussion in the briefing document focuses on subgroups of existing claimants defined by having reached or not various milestones within the program. Such subgroups are of undeniable policy relevance and have the (political) virtue that meaningful impact estimates will be obtained relatively soon.

An alternative view, to which I am very sympathetic, is that the key subgroup of interest in terms of both external validity and longer-term policy relevance consists of individuals with no experience of the existing SSDI program. This could mean either new SSDI applicants or new SSDI beneficiaries. The choice between the two relates to concerns about estimating induced entry effects as well concerns about the speed with which meaningful impact estimates are obtained. For the purposes of this question, consider a choice of new beneficiaries as a subgroup. These individuals can learn the WISP incentives from the very start of their participation, as would be the case were WISP implemented as policy throughout the program. They will not have to unlearn the (extremely complicated – hence the desire to examine WISP) current system and then learn WISP in its place. This means that understanding of the WISP incentives among treatment group members will likely better approximate what it would be in a world where WISP was policy; put differently, this means more credible external validity.

In addition to recommending that new claimants be the main subgroup of interest, or at least one of the main subgroups of interest, it is important to keep in mind that the number of subgroups for which separate impacts will be produced has important implications for the sample size and for the power of the resulting estimates. I encourage SSA to focus on a very limited (i.e. two or three) number of subgroups of interest so that relatively precise impact estimates can be obtained with reasonable sample sizes. I strongly recommend that one of the selected subgroups be a flow sample of new applicants or new beneficiaries. A second, and less important, advantage of keeping the number of subgroups very small is that the evaluation avoids having to deal with issues of multiple comparisons.

4.3 What should the minimum detectable effects be set for producing key policy impacts? What levels of statistical power would be recommended?

I am not enough of an expert on the substantive literature on SSDI to have precise views regarding likely effect sizes. I encourage SSA to canvass the views of subject area experts such as the researchers in the disability group at Mathematica Policy Research, as well as others in the consulting (e.g. the relevant people at Urban and Abt) and academic (e.g. my Michigan colleague John Bound) worlds for this purpose.

The literature on the evaluation of social and educational programs has evolved norms regarding levels of statistical power. I recommend that the design of the evaluation adhere to those norms.

Sample

4.4 What are the implications of SSA subsample division on the demonstration's external validity?

See my response to Question 4.2

4.5 What are the implications of SSA's age restriction for the evaluation?

SSA's proposed age restriction for the evaluation seems very reasonable to me on two grounds. First, current claimants age 55 and older will age off of the program relatively soon and so are less relevant in regard to the longer-term effects of WISP. Second, standard human capital theory predicts, quite reasonably, that older individuals are, all else equal, less likely to undertake costly investments (such as job search or vocational training) than younger workers due to the shorter time horizon over which they can reap the returns to such investments. Thus, impacts are likely to be lower among older claimants.

4.6 What other subgroups should be targeted (e.g., SSI recipients and those in the Medicaid buy-in)

See my response to Question 4.2.

Data Collection

5.1 What data sources are most important for addressing research questions?

The most important data sources are administrative data on labor market outcomes, SSDI benefit receipt, and Medicare utilization.

The most obvious source of data on labor market outcomes is the administrative earnings data housed at SSA. This data has the advantage of being quite accurate and having no issues of survey non-response. It does miss informal work that might be captured in a survey and, because it is annual data, it lacks the temporal detail to study short-term employment behavior. Such short-term behavior is interesting both to determine how steadily current and former SSDI claimants are working and to investigate the incentives around the timing of earnings receipt provided by WISP.

A number of other research questions would require either alternative administrative data, such as UI earnings records, or survey data. I discuss these other research questions and the related outcomes in my responses to other questions, especially Question 1.2 on outcomes.

5.2 Which administrative data sources can be leveraged and what are their advantages?

See my response to Question 5.1.

5.3 Will survey data be needed?

Survey data are not needed to estimate impacts on the most important outcomes: employment, earnings, benefit receipt and Medicare utilization.

At the same time, survey data are likely required to assess how well treatment group claimants understand the work incentives provided by WISP.

Survey data are also required to examine many other outcomes of interest, such as many of the outcomes listed in my response to Question 1.2.

Data Collection

5.4 When should outcomes in the data be measured?

The outcomes in the data should be measured as often as possible, starting at the time of random assignment if not before. Pre-random-assignment outcomes may be valuable conditioning variables in obtaining experimental impacts as they can reduce the residual variance and so increase the precision of the estimated average treatment effects. In terms of outcomes after random assignment, having a complete time-series of outcomes is critical for doing cost-benefit calculations.

5.5 How should savings to Medicare and private health insurance enrollment be measured?

I am not an expert on health-insurance-related data sets. My understanding is that savings to Medicare can be estimated (with some effort) using matched Medicare administrative data. The discussion at the meeting suggested that Medicare have some information on private insurance but not enough to do a good job of estimating impacts on this outcome. Thus, such information would have to be collected using surveys.

5.6 How should demonstration fidelity be measured?

In my view, there are three key aspects to demonstration fidelity. The first is that claimants themselves understand the incentive structure implicit in WISP. It is hard to see how this can be examined without doing surveys. Designing a survey instrument (or part of a survey instrument) for this purpose will require some thought as individuals may be broadly aware of what WISP is up to without necessarily being able to consistently answer questions about particular details.

The second aspect to demonstration fidelity concerns SSA staff. SSA staff knowledge of WISP could be tested via surveys, in-person interviews or by using "mystery shoppers" (actors posing as claimants who report their experiences back to the evaluation) or by some combination of these.

The third aspect to demonstration fidelity concerns staff at WIPAs and other support organizations. The alternatives for measuring fidelity are the same here as with SSA staff.

Site Selection

6.1 What constitutes a site?

I think it makes sense to consider field offices as sites for the purposes of this evaluation. There are simply too few regional offices or areas to allow effective random assignment.

The one alternative to using field offices would be to use claimant zip code or some other geographic division The reason why using claimant zip code would make sense is that randomly assigning field offices in cities with multiple field offices could lead to a situation in which potential claimants choose their field office based on whether they want to end up with WISP or with the current system. This type of selection, should it occur, would (partially) undermine both internal and external validity. Even if a decision is made to go with field office random assignment for simplicity or cost reasons, care should be taken to estimate the empirical extent of such "field office shopping" in the experimental treatment and control groups.

6.2 How many sites should be selected?

Enough sites should be selected to get the desired level of power for the smallest subgroup of interest.

6.3 What programs are most likely to interact with WISP?

I am not enough of a subject area expert to add much here. There was discussion at the meeting about avoiding areas where the BOND demonstration is underway. This seems reasonable to me.

Site Selection

6.4 Should/Could interstate program variation be exploited to minimize these interactions?

I am not exactly sure what this question has in mind.

Timely Medical CDRs

7.1 Should WISP conduct timely medical CDRs for the treatment and/or control/comparison groups?

I recommend Option 3 on page 22 of the briefing document. My concern here is mainly with internal and external validity. The control group should certainly experience the current norm, which is no timely CDRs. This dictum rules out Options 1 and 2. As CDRs increase in importance to the successful operation of the SSDI system under WISP, my expectation is that were WISP implemented as ongoing policy, timely CDRs would be funded. Thus, I think Option 3 maximizes both internal and external validity.

Information Dissemination and Training

8.1 What should the initial and primary medium(s) of contact be with WISP subjects?

I defer to others with more relevant knowledge on the details. The key is to make sure that treatment group members understand the nature and duration of the WISP treatment. If they don't know about it, they cannot respond to it.

8.2 Who should treatment subjects contact to ask questions about the intervention?

I defer to others with more relevant knowledge.

8.3 Should there be checks on whether WISP subjects received or understood the outreach materials that they were sent?

There should definitely be checks on whether WISP subjects understand the WISP treatment. My prior is that effectively communicating the WISP treatment, which is a quite substantial change from the existing system, will require more than just sending a brochure or two. As noted in the response to an earlier question, the ideal is to bring the treatment group members to a level of understanding of the WISP rules that equals that which would be present if WISP were in place as a permanent policy. Failure to do so weakens the external validity of the evaluation and increase the risk of a "false zero" wherein the evaluation shows little or no impact even though WISP as policy would have non-trivial impacts. One or more surveys is likely the best way to measure claimant understanding of the WISP program rules.

Information Dissemination and Training

8.4 Does the TAP have any recommendations regarding training for WIPA organizations, TTW, providers, or any other entities that will interact with WISP subjects?

Other than reinforcing that such training is a good idea, I have nothing to add here.

8.5 How will these entities identify WISP subjects?

I have little to add here other than noting that an important motivation for a design that varies the WISP treatment at the site level, rather than the individual level, is so that these entities can be fully informed about the presence of WISP in their area and thus know to ask about it when dealing with claimants.

External Validity

9.1 What would be the key features of a national version of WISP?

I am not clear on what this question is asking. My thoughts on external validity more generally appear in my responses to several other questions.

9.2 How will the TAP's other recommendations affect the evaluation's external validity, especially in relation to a national version of WISP?

See my responses to the other questions.

TAP Member #4

General Comments

Which items are most important to the WISP evaluation?

1. Tracking the employment outcomes of beneficiaries while accounting for other influences that led to their decision to RTW. For example, did they receive other services, such as WIPA, TTW, VR, One-stop services that influenced their success at RTW. Was WISP the real reason for RTW or is it some other reason(s)?

2. Comparing WISP outcomes to other programs such as BOND, TTW, WIPA, VR to determine whether the services received there by non-WISP participants were more or less successful than WISP RTW outcomes.

3. Evaluating the effectiveness of the field offices' responsiveness to beneficiaries, especially related to reinstating benefits when a beneficiary stops working.

Were there any key topics not covered in the brief materials and the questions below that you would like to raise here?

I am not certain what the logic is to incentivize beneficiaries to RTW with this proposal. As complicated as the current work incentives are, when beneficiaries learn about them, they are attractive. The BOND program allows the safety net of 2:1 to avoid the earnings cliff that encourages people to park. WISP does not address the earnings cliff. It seems logical that it will not reduce or relieve parking any more than the current TWP.

Looking at the bigger picture of WISP vs. BOND, it seems more logical to simplify the work incentives by beginning the 2:1 offset from the first earned dollar. That is simple and avoids the earnings cliff. If we need Congressional demonstration authorization, why not ask to test that option too?

Other comments?

Research Questions

1.1 Do the five questions cover all the outcomes of interest to evaluators and SSA?

See 1.2 for additional comments. I think #5 should be deleted. Measuring induced entry, especially with the 2-year waiting period for Medicare eligibility, would require a huge sample size. Just as with the ticket to work, induced entry can be approximated by the actuaries at CBO.

1.2 Which additional outcomes should be incorporated into the research questions? Should any of those questions be given more weight than the five core questions in the original solicitation?

I recommend determining how the WISP provisions affect work behavior and benefits payments when part of the intervention includes services, such as WIPA, TTW, VR or One-stops? The current key questions, especially #1, evaluate how a change in the work incentives impact RTW behavior, but does not look at whether additional services along with WISP could lead to more RTW. This outcome should be given more weight than #3, #4, or #5.

Dissemination/Policy Relevance Guidelines

2.1 When does the evaluation need to produce key findings to policymakers?

Annually.

2.2 What should the follow-up period be on those findings?

Throughout the duration of the project.

2.3 Will other stakeholders demand findings and information that policymakers would consider less important?

Other stakeholders will want to know:

1. Manner in which information is disseminated and participants are "randomly" selected.

2. Ease of reporting earnings.

3. Are beneficiaries confident they will immediately get their benefits back if they fall below SGA? Can SSA guarantee they will?

Dissemination/Policy Relevance Guidelines

2.4 What information would other stakeholders finding interesting and how should it be disseminated to them?

Other stakeholders would be interested in all the same finding data. Dissemination should be via SSA's website, listservs, e-letters for updates and webcasts with archives.

2.5 What short run and long run findings would be considered "convincing."

Short-term: Beneficiaries going to work above SGA. Long-term: Beneficiaries staying at work and not parking. These will only be convincing, however, when these outcomes can unequivocally be associated with WISP and not contaminated by other variables.

Evaluation Design

3.1 Should WISP use an experimental or non-experimental design?

Experimental

3.2 If a non-experimental design should be used, then what type of design offers most promise for WISP design specifically?

3.3 What are the implications for the choice of design on both internal and external validity?

1. There is always the potential of harming participants by requiring them to use a RTW option that is not as attractive as the current work incentives, especially by eliminating the TWP and EPE. Even though WISP is presumably random assignment, those randomly chosen cannot choose to participate. If a WISPer returns to work without benefit of the TWP, conceivably (s)he will not have as much earned and unearned net income as a non-WISPer.

2. Implementing WISP to those already in the TWP or EXR would likely be harmed. However, if those subgroups are not included in the demo project, it will not demonstrate what will happen when the program is operationalized.

3. To achieve the least contaminated outcomes, the project could target only beneficiaries who are not engaged in TWP or EPE and do not even know they exist. However, new beneficiaries are not likely to RTW right away.

Evaluation Design

3.4 If random assignment is used, should it be individual or site level?

Site level. Fewer field offices and community providers would need to be trained on the program. Carefully selecting representative and geographically dispersed sites would avoid contamination (e.g. from BOND).

3.5 Given the selected design, which evaluation components should be emphasized over what time frames?

The entire project should attempt to complete as close as possible to completion of BOND so the outcomes can be compared with regard to the 2:1 offset. The field office and administrative efficiency will depend largely on SSA's ability to implement the automated earnings reporting system and policies with regard to immediately resuming benefits when a beneficiary stops working.

The number of beneficiaries who RTW should be measured and reported in the first year and each year thereafter. That is the most important element in this project. The administrative efficiencies and Medicare costs and utilization would likely take longer to get meaningful data. Those components should probably be measured and reported beginning in year two and each year thereafter.

Sample

4.1 What should the recommended sample size be (relates to the demonstration design)?

I thought this was already decided at 80,000. If the typical RTW rate now is .5%, a RTW rate in excess of 400 participants would show some positive impact. Given the various subgroups to be measured, however, that might be a small sample size. Depending upon the number of subgroups, I recommend a sample size of working individuals of 100 for each group. I have no scientific basis for that recommendation because I am not a researcher. As a practitioner in the real world, however, I know that significant data can be collected from 100 people.

4.2 Should subgroups be targeted along the work path? If so, are the subgroups originally suggested by OMB and SSA appropriately selected?

I don't recommend there should be subgroups. Studying WISP behavior by subgroups is not consistent with external validity. If eventually the WISP becomes operationalized, it will not be offered to some groups and not others. In the real world, when operationalized, there will no longer be an option. The TWP, EPE and EXR will cease to exist. If anything, it might make sense to grandfather those already using the TWP. Even testing only those under 55 because of the lifetime Medicare inducement makes no sense in terms of external validity. For purposes of WISP, only those not using the TWP, EPE or EXR should be included in the sample.

Not testing subgroups will also necessitate a smaller sample size, thus costing less to conduct the project.

4.3 What should the minimum detectable effects be set for producing key policy impacts? What levels of statistical power would be recommended?

I don't know what this question means.

Sample

4.4 What are the implications of SSA subsample division on the demonstration's external validity?

Already discussed in 4.2 above.

4.5 What are the implications of SSA's age restriction for the evaluation?

I think this sets up a false assumption that those over 55 will not RTW under WISP. The argument is that the lifetime Medicare provision of WISP would not induce those over 55 to RTW because they already have the Medicare extension available to them, which would take them up to their usual retirement age anyway. As already discussed, however, once operationalized, it will not matter how old or which subgroup an individual belongs to. He or she will go to work or not.

4.6 What other subgroups should be targeted (e.g., SSI recipients and those in the Medicaid buy-in)

None.

Data Collection

5.1 What data sources are most important for addressing research questions?

1. Beneficiary earnings reporting

2. Field office monitoring of how long it takes to suspend and reinstate benefits.

3. How many WISPers end up with overpayments?

5.2 Which administrative data sources can be leveraged and what are their advantages?

5.3 Will survey data be needed?

Yes. Beneficiaries need to be asked whether their decision to RTW is because of the WISP or some other reason. They also need to be asked what other programs they are participating in, such as WIPA, TTW, VR, MH, DD, One-stops, that could encourage RTW.

Data Collection

5.4 When should outcomes in the data be measured?

RTW should be measured in the first year and each year thereafter. Administrative outcomes should be measure after year two, assuming the automated reporting system is in place and operating.

5.5 How should savings to Medicare and private health insurance enrollment be measured?

Savings in Medicare can be measured by savings from beneficiaries who RTW and employer-provided health insurance becomes primary.

5.6 How should demonstration fidelity be measured?

This is another reason to keep the project simple. Not including subgroups will eliminate complexity in design and implementation, and as such will likely preserve project fidelity. The difficulty will be in selecting the sites. The WISP project sounds so much simpler than other programs like BOND that it should be easy to implement, thus making fidelity that much easier to maintain among sites.

Site Selection

6.1 What constitutes a site?

A site should be a group of geographically clustered field offices covering a specific area (a state?). I tend to lean toward the BOND design on this because I think the two projects are related to each other in terms of likely outcomes. The WISP sites should be separate from the BOND sites, but should be similarly selected.

6.2 How many sites should be selected?

10, again using BOND as a model.

6.3 What programs are most likely to interact with WISP?

WIPA, TTW, VR and One-stops. To a lesser extent, DDD and Mental Health (MH).

Site Selection

6.4 Should/Could interstate program variation be exploited to minimize these interactions?

I don't understand this question.

Timely Medical CDRs

7.1 Should WISP conduct timely medical CDRs for the treatment and/or control/comparison groups?

Yes for both groups, unless the beneficiary is consistently working above SGA. Apply the same CDR protections for Timely Progress used in the TTW program.

Information Dissemination and Training

8.1 What should the initial and primary medium(s) of contact be with WISP subjects?

First, a letter from SSA informing them of being selected for WISP. Invite them to participate in a webcast/teleconference to learn about the project and what it means. Include an information line where a beneficiary can call and speak to an agent one-on-one to answer questions. Make sure this is a WISP line with staff specifically versed on the WISP project as opposed to just loading this responsibility onto site field offices, or worse, the TSC. Reminder letters should be sent semi-annually to those not participating.

8.2 Who should treatment subjects contact to ask questions about the intervention?

A specially assigned WISP line and staff persons versed on WISP.

8.3 Should there be checks on whether WISP subjects received or understood the outreach materials that they were sent?

Additional attempts at contact should be made if the originally sent letters are sent back undeliverable.

Information Dissemination and Training

8.4 Does the TAP have any recommendations regarding training for WIPA organizations, TTW, providers, or any other entities that will interact with WISP subjects?

This is such a simple project requiring so little non-SSA field intervention that a webcast explaining the provisions should be all that is necessary.

8.5 How will these entities identify WISP subjects?

This could be a big problem. Unless a participating beneficiary mentions it to a provider, there is little likelihood that the provider will know. This will lead to confusion and misunderstanding. Without a signed release from the beneficiary, community providers cannot get information from SSA about a beneficiary's status in any way. Even if the beneficiary is willing to sign a release of information, it still takes time to get the information from the field office. A beneficiary might call the WIPA for some general information about the work incentives. The beneficiary would possibly think (s)he is eligible for the TWP, not realizing they are in a special project that makes them ineligible.

By using a site-specific design, providers in the specific geographical areas could know to make it a habit to ask callers or visitors if they are participating in the WISP before giving out information. Hopefully, beneficiaries will remember whether they are a participant. I think we can count on some confusion along the line.

External Validity

9.1 What would be the key features of a national version of WISP?

I honestly doubt that WISP's primary feature (all or nothing) will be attractive to beneficiaries. It does not prevent parking. There will likely be significant doubt that benefits will be quickly reinstated when needed. This will be such a small part of the field office's overall workload that WISP will be low on their priority list. I have serious doubts that SSA can get a reliable telephone/internet reporting system in place quickly enough. Ultimately, I think what will work is a straight 2:1 offset from BOND (without waiting for SGA earnings before the offset begins) and lifetime Medicare. A combination of WISP and BOND has the most potential to succeed.

9.2 How will the TAP's other recommendations affect the evaluation's external validity, especially in relation to a national version of WISP?

I think external validity will be seriously affected negatively if we get too involved with subgroups in the demonstration. Keep the demo as "real world" as possible; that is, use a truly random sample to achieve natural diversity among subjects. But use carefully selected sites to avoid contamination from BOND. Do not include those in TWP, EPE or EXR in the demo because those would no longer exist in a national version rollout.

**********THANKS FOR THE OPPORTUNITY TO PARTICIPATE ON THE TAP**********

TAP Member #5

General Comments

Which items are most important to the WISP evaluation?

There are a number of key decisions that need to be made about the sampling design for the evaluation and the recommended strategy for implementing a randomized experimental evaluation. I have recommended a nationally representative sample for the evaluation, stratified by some key criteria that still need to be finalized to ensure that important subgroups are adequately represented in the sample (see the discussion that follows). If the recommendation for an experimental design with randomization at the individual level is accepted, it will be important to determine if there will be any randomization in subgroups as well, or which subgroup impact estimates would be recovered through nonexperimental analyses of the data collected. These decisions will also affect required sample sizes and power calculations. To the extent possible, key issues and tradeoffs have been identified in my comments. In some cases, it is simply not possible to make definitive recommendations without additional information or in advance of key decisions that will affect subsequent evaluation design features.

Were there any key topics not covered in the brief materials and the questions below that you would like to raise here?

No, the briefing materials and the discussion of them on 2-10-12 with the TAP were very thorough.

Other comments?

None; see the many that follow below.

Research Questions

1.1 Do the five questions cover all the outcomes of interest to evaluators and SSA?

The five key research questions cover the following important dimensions/domains of potential outcomes: (1) work behavior and SSDI benefit payments, (2) Medicare costs and utilization rates and use of private insurance, (3) beneficiary knowledge/understanding of benefits, use of the automated system of earnings reporting and SSDI workloads, (4) administrative costs and improper payments, and (5) induced entry. Given the limitations of available administrative data to adequately address these five key questions, I recommend that a survey of WISP-eligible individuals be conducted in the evaluation. This would allow for the assessment of several additional outcomes that I think would be valuable for more fully understanding program impacts and informing stakeholders of outcomes likely to be of interest to them.

As discussed in the 2/10/12 WISP TAP meeting, health status is an important outcome for beneficiaries, program administrators and stakeholder groups such as the Consortium of Citizens with Disabilities (CCD). There might be concern about individuals disconnecting from SSDI benefits and then suffering health setbacks, or alternatively, increased access to Medicare through WISP and/or engagement in work activities could contribute to mental and/or physical health improvements. Although it might be possible to get information from individuals about the most recent health care visits, it might be easier to get comparable data for all surveyed individuals by asking a set of questions on daily living activities/functioning. In addition, one of the TAP members suggested that beneficiaries are trying to get their most basic needs met through participation in SSDI, thus, I also recommend a set of questions about how well they are able to meet their basic needs (food, housing, etc.), with or without SSDI benefits.

A sub-question of the first research question listed in the briefing materials asks how many beneficiaries do not make a work attempt due to the loss of the TWP. Although it should be possible to empirically explore (with either experimental or nonexperimental methods) the differential of effects of WISP on those who lose access to TWP, a causal inference about the relationship between work attempts and loss of TWP may be difficult to ascertain given the way that work activities are reported. This would be another question to ask on a survey of individuals eligible for WISP, i.e., how the loss of TWP affects their work efforts (and/or related outcomes). The same point generally holds for understanding more fully the affects of permanent disability entitlement and permanent Medicare eligibility on work behavior and outcomes. Because there is an interest in knowing *attempts* to work (or the incentives they create) as well as success in securing work, it will be important to gather data from a survey.

Medicare is mentioned in question 3 but effects on Medicaid utilization and costs should also be included. Among the sub-questions listed in the briefing materials in this section, it asks if more individuals accept jobs that do offer health coverage under WISP. I agree that this is a potentially important outcome/question of interest, but I think this could only be assessed with supplementary survey data collection, and I recommend that this be collected. It ties in closely with questions related to work behavior and individuals' understanding of the program incentives.

For the questions about reductions in administrative burdens and associated cost savings, it will be important to examine available outcome measures by subgroups who may be differentially affected by the WISP changes (such as those involved in TWP).

The fifth question about induced entry is important but may be among the most challenging to measure well. Even small effects, which may be more difficulty to precisely estimate, could have substantial implications for program costs, and these would likely be known later than those related to reducing administrative costs.

1.2 Which additional outcomes should be incorporated into the research questions? Should any of those questions be given more weight than the five core questions in the original solicitation?

I recommend adding a sixth core question that asks about the impacts of WISP on eligible individuals' health status/functioning in daily living activities and their ability to meet their basic needs. The effects of WISP on these dimensions are likely to be related to work behavior, access to health insurance/services and SSDI payments, which are addressed in the first two evaluation questions, but these outcomes are sufficiently distinct that I think they merit a separate core research question in the evaluation. While it may not be critical to SSA to understand this outcome, I do think that it may be critical to some stakeholders, and this information could be important later in determining whether a full rollout of the policy carries minimal or some serious risks to those affected.

Dissemination/Policy Relevance Guidelines

2.1 When does the evaluation need to produce key findings to policymakers?

There are some impacts which will be more readily measured with accuracy in the short-run, and others that will require a longer time to fully measure or understand their trajectories over time. For example, it should be possible to measure and report administrative cost savings within the first and second years of WISP operations (e.g., such as savings due to the change in recording earnings from work when paid rather than earned). However, as one of the participants in the 2-10-12 meeting noted, it would be important to forestall final conclusions about the program's ultimate impacts on cost burdens/savings until potentially important incentive effects of induced entry or changes in work behavior could be more fully assessed.

Another area of outcomes that could be assessed earlier in the WISP program operations is individuals' understanding of the program rules and their effects on their incentives to work. For any whose participation in TWP is affected, this impact should be assessed close to the time that these changes come into effect. Again, however, the longer-term affects on work behavior and individual well-being should also be assessed, as participants' initial response to program features might change over time as they come to better understand the implications of WISP provisions for their choices and well-being.

In addition, some individuals affected by WISP will be current SSI/SSDI beneficiaries, while others will be newly taking up the program. It is difficult to say in advance whether current beneficiaries, who are already familiar with the SSI/SSDI systems and benefits, will find it easier to revise their understanding of program benefits and requirements as they transition to WISP than it will be for new entrants to understand the WISP program provisions and requirements. This could be assessed earlier in the program operations, but the implications may be that it will take longer to fully observe the program impacts on the core outcomes defined for the evaluation.

A ballpark estimate is that a 5-year evaluation period should be anticipated after the start of the evaluation (after the first random assignment is made) to fully measure the program's impacts, as impacts might also vary over time with the implementation of the program and if program take-up by those eligible for WISP changes.

2.2 What should the follow-up period be on those findings?

See above discussion—a 5-year follow-up period after the start of treatment and control assignment and monitoring should be planned. A related question concerns how often measurement of different outcomes would happen over the follow-up period. Some outcomes, such as those on work behavior, benefit receipt, improper payments, health insurance utilization, health status and ability to meet basic needs should probably be assessed annually. Others such as individuals' understanding of the program features and incentives, automated work reporting system use, SSDI workloads and administrative costs might be assessed early in the evaluation period and then possibly again later in the evaluation period to assess if these outcomes change over time. The induced entry effect might be assessed at one or two time points after the start of WISP.

2.3 Will other stakeholders demand findings and information that policymakers would consider less important?

Other stakeholders should be similarly interested in the outcomes associated with the six research questions discussed above (the five core questions plus the additional recommended question). Although policymakers are also likely to see these as important, policymakers and stakeholder groups (which are diverse) might view tradeoffs between different outcomes in differing ways. In times of budgetary duress, for example, policymakers might see administrative and operational cost savings as more important than stakeholder groups such as CCD who might be more concerned about program changes on beneficiaries' health and employment statuses and their ability to meet their basic needs.

Dissemination/Policy Relevance Guidelines

Dissemination/Policy Relevance Guidelines

2.4 What information would other stakeholders finding interesting and how should it be disseminated to them?

Other stakeholders may be particularly interested in the interactions of the WISP with other interventions (e.g., WIPA) and policy changes (e.g., implementation of the Affordable Care Act). These would be difficult to fully explore in an experimental evaluation framework, but supplemental nonexperimental analyses could be undertaken to understand interaction effects. In addition, policy changes such as those coming under ACA, or other environmental/economic changes that might influence interactions, may likewise change/evolve over time, so interactions effects would likely not be well understood in the short-term.

2.5 What short run and long run findings would be considered "convincing."

This has likewise been addressed in the discussion above. Program impacts that are most likely to be accurately observed in the short-run include those with clear/direct effects on administrative burdens (e.g., the change to recording earnings in the month paid) or with immediate potential implications for SSI/SSDI benefits, work and health insurance (such as changes in work effort after the ending of TWP or the permanent extension of Medicare eligibility). I would recommend that the evaluators plan for dissemination of findings at three time points: (1) early findings at the end of the first year, (2) intermediate findings at 24 or 30 months, and (3) long-run findings at 5 years (60 months), with the possibility of later releases if there are delays in data access or problems in program implementation or important policy changes that are found to interact with core program outcomes over time (e.g., ACA changes related to health insurance, utilization rates, work behavior, etc.).

Evaluation Design

3.1 Should WISP use an experimental or non-experimental design?

I definitively recommend that the WISP evaluation incorporate an experimental design. There are many complexities in SSDI and changes coming with the new WISP program, and the experimental design will be critical to the identification of impacts. At the same time, it would require a fairly complex experimental design to be able to experimentally estimate all desired subgroup impacts and potential interactions, and thus, it will also be important to plan for the use of nonexperimental methods in estimating some impacts.

3.2 If a non-experimental design should be used, then what type of design offers most promise for WISP design specifically?

The decision of what types of nonexperimental methods to employ will likely have to come later, after subgroup sample sizes and characteristics are ascertained, interactions are observed, and it is clear which questions need to be addressed using nonexperimental methods. For those questions concerning possible heterogeneous effects of WISP on different SSI/SSDI subgroups, it should be possible to construct internal comparison groups, which should make methods such as propensity score matching more viable. It might also be possible to use multilevel modeling with individuals nested within areas or field offices to explore how differences in implementation might have affected individuals' understanding of WISP, their responses to program incentives, and program impacts.

3.3 What are the implications for the choice of design on both internal and external validity?

The experimental design will be important for establishing internal validity for the estimation of average impacts of WISP (at the level of randomization). External validity will be enhanced by drawing a national study sample for the evaluation. The tradeoff between these two is that it is more costly to conduct an experimental evaluation in a large number of randomly distributed sites. As discussed at the 2/10/12 meeting, more SSI/SSDI clients are accessing benefits and services in ways that are not tied to a specific site office, so this concern may be less important in implementation for an experimental design at the individual level. (The next question addresses this issue more specifically).

Evaluation Design

3.4 If random assignment is used, should it be individual or site level?

I recommend that randomization take place at the individual level. Sites (areas or field offices) should first be selected randomly within strata, and individual-level randomization should then take place within the sites. This is probably essential (individual-level randomization) to ensuring sufficient sample sizes and the opportunity to explore heterogeneity in program implementation and impacts using nonexperimental methods.

3.5 Given the selected design, which evaluation components should be emphasized over what time frames?

Randomization should take place over a period long enough to ensure a sufficient inflow of new SSI/SSDI/WISP beneficiaries (and persons assigned to the control group), as well as the migration of existing beneficiaries to the WISP program. Without additional information about the current "stock" of beneficiaries and the anticipated "flow" of new clients, it is difficult to say exactly for how long randomization should take place. It will also be important to monitor implementation of the new WISP features and to continue random assignment into a period when the program appears to be operating as intended and without glitches that would compromise the intended treatment. The nonexperimental methods can be employed after the random assignment phase and treatment and control groups are established/complete.

Sample

4.1 What should the recommended sample size be (relates to the demonstration design)?

The suggestion during the 2/10/12 meeting that a 10% national sample be chosen for the demonstration (likely involving approximately 130 field offices and 80,000 clients) seems reasonable, although ultimately, costs and power calculations would likely drive the final determination. In order to compute the expected power of a given sample size to detect true program impacts, specific information is needed on the variability of these outcomes in the target population (and this is not information that is currently available to the TAP members, and these calculations should likely be done by the evaluator in conjunction with SSA).

4.2 Should subgroups be targeted along the work path? If so, are the subgroups originally suggested by OMB and SSA appropriately selected?

OMB and SSA appear to have identified key subgroups that might be differentially affected by the WISP intervention, including those that have begun TWP and those in EPE and termination status. There maybe be additional subgroups that will be identified in the course of program implementation, for example, individuals whose WISP program incentives might be affected by other program participation or environmental factors (e.g., changes coming under the Affordable Care Act). It is valuable to identify these subgroups in advance of the evaluation and to anticipate potential differences in their responses to WISP. However, if a nonexperimental evaluation component is also included in the overall evaluation design, it is not necessary to factor all of these potential subgroups into the experimental design.

4.3 What should the minimum detectable effects be set for producing key policy impacts? What levels of statistical power would be recommended?

This relates to question 4.1, and more information would be needed to calculate the minimum detectable effects and associated power with a given sample size. One convention in the evaluation field is to plan for evaluations that will identify impacts with at least 85% power. This is a decision that should be made in consideration of the policy context and the resources required for both program implementation and evaluation.

Sample

4.4 What are the implications of SSA subsample division on the demonstration's external validity?

The decision of where to randomize at the individual level is an important question for the estimation of impacts for subgroups of SSI/SSDI participants who may be differentially affected by the new program. One possibility for taking into account subgroup impacts would be to include them as strata within which individuals are randomized within sites or areas. As this complicates the design somewhat, SSA should probably decide if there are any of the subgroups for which separate experimental impact estimates are particularly important so as to warrant random assignment within these groups. For example, it was noted in the 2-10-12 meeting that individuals who have not completed the TWP may be at risk for "potential harm" (i.e., may be worse off under WISP); specifying them as a subgroup in the evaluation design will draw more attention to this possibility, but it might also generate results that would suggest these risks are minimal. It might also be possible to get sufficient data for understanding impacts on these subgroups through oversampling of them. The implications for external validity of the results will depend on the extent to which the approach selected facilitates a subgroup sample that is representative of these subgroups in the national SSA population.

4.5 What are the implications of SSA's age restriction for the evaluation?

Estimates provided in the 2-10-12 meeting suggested that about one half of individuals potentially eligible for WISP would be age 55 years or over. The concern is that those age 55 years or older will be less likely to return to work and would already have access to Medicare benefits, and thus, their response to the WISP features and incentives would likely be different and would possibly be less important than for individuals under age 55. It appears that a decision is yet to be made as to whether SSA will include those 55 years or older in WISP. If there is no age restriction for participation in WISP, it will be valuable to understand if the program generates different work behavior responses for this group. This could be a potential subgroup analysis in a nonexperimental evaluation, depending on how many individuals are close to age 55. If a decision is made to exclude those age 55 years and older from WISP (this a program design decision, not an evaluation design decision), there may be potential for a regression-discontinuity analysis of impacts using those who are just below age 55 years compared to those just at/above this cutoff (again, depending on sample sizes and the age distribution within them).

4.6 What other subgroups should be targeted (e.g., SSI recipients and those in the Medicaid buy-in)

This question about what other subgroups should be targeted in the sampling strategy is difficult to answer definitively without additional information on their respective sizes and expected inflow in the future. The response to question 4.4 applies here, too. For example, the opportunity to be a part of the Medicaid buy-in is changing over time across states, so this might be easier to examine nonexperimentally than to attempt to target individuals in this subgroup in sampling. In the 2-10-12 discussion, it was noted that SSI/SSDI concurrents are expected to be about 25% of the WISP-eligible population, and given that WISP is expected to make SSDI more like SSI in some respects, it might be worthwhile to stratify and/or allow random assignment within this subgroup; however, this is probably not essential (i.e., nonexperimental analyses could be done to examine differential impacts for this subgroup as well).

Data Collection

5.1 What data sources are most important for addressing research questions?

All available administrative data should be examined for potential use in the evaluation. Based on the information included in Exhibit 1, which was circulated following the 2/10/12 TAP meeting to show the potential data sources for the WISP evaluation, it appears that there is very limited information on individual characteristics and employment and current activities. There appear to be important "holes" in terms of data sources available for constructing outcome measures as well as measures of WISP services. If new data fields can be incorporated into the SSA management information systems, particularly for WISP services, this would be advantageous for tracking these measures over time. It appears from Exhibit 1 that it will be essential to collect survey data (in the absence of significant additions to the SSA administrative data system).

5.2 Which administrative data sources can be leveraged and what are their advantages?

Unemployment Insurance (UI) records might be accessed to get more details on individual earnings, but this would not provide information on hours worked or any other aspects of the type of employment or benefits received with it, including private insurance. Relying on UI data may also miss some work activities, and it might also be accessed with a delay that would affect the timeliness of impact estimation. Administrative data might also be obtained to measure other public assistance benefits such as food stamps/SNAP and possibly individuals access to public health insurance benefits. If data sharing agreements with the states need to be established to obtain access to these types of administrative data, this could require considerable time and effort.

5.3 Will survey data be needed?

Survey data is likely to be important to capturing these key data that are listed in Exhibit 1(Potential Data Sources for WISP): additional personal characteristics; information on employment and current activities such as training; WISP services (unless new administrative data fields are created); detailed employment (e.g., hours worked) and current activities (e.g., training) following random assignment; private health insurance and other employment benefits; other public program participation; health status and functioning, and individuals' understanding of and attitudes towards SSDI work incentives. These data fields/variables are central to a comprehensive evaluation of the WISP, suggesting that it will be essential to collect survey data in the WISP evaluation.

Data Collection

5.4 When should outcomes in the data be measured?

As indicated in response to question 2.5, I recommend that some outcomes be measured at the end of the first year (such as administrative cost savings, individuals' understanding of the program provisions and incentives, SSDI benefit payments, use of the automated system of earnings reporting, and SSDI workloads); some should be measured at an intermediate time point of 24-30 months, (including those measured at the end of the first year, work behavior, Medicare costs and utilization rates and use of private insurance, induced entry and health status/functioning), and all outcomes should also be assessed in the long-run, at approximately at 5 years or 60 months. The precise timing should also be determined in consideration of the costs of obtaining these data from administrative data systems or separate surveys of individuals.

5.5 How should savings to Medicare and private health insurance enrollment be measured?

These outcomes should be measured with a combination of administrative and survey data, to the extent that administrative data are available for Medicare (and Medicaid) eligibility and claims, and from surveys (for private health insurance). The precise measures will depend on the administrative data fields that can be accessed, and in part on constraints on the size/length of a survey of beneficiaries (treatment and control group members).

5.6 How should demonstration fidelity be measured?

In designing the evaluation, care should be taken to ensure that random assignment and other evaluation design features and data collection do not interfere with how the program is intended to roll out. To the extent that WISP program activities and services to beneficiaries can be captured in an administrative data system, it will be easier to track demonstration implementation and its fidelity to the intended program model over time and across areas and offices. Demonstration fidelity should also be monitored for important subgroups, such as those for whom TWP is affected by WISP, SSI/SSDI concurrents, and others for whom there may be differential responses or interaction effects. Essentially, the "processing" of beneficiaries in the program, their understanding of the program features and incentives, their use of automated earnings reporting and SSDI payments will be important to monitor to assess demonstration fidelity. Administrative data, data from beneficiary surveys, and possibly from interviews with SSA staff involved in implementing WISP would be useful in measuring demonstration fidelity.

Site Selection

6.1 What constitutes a site?

I do not have a definitive recommendation on what should be considered a site in the WISP evaluation. Possibilities that were discussed in the 2-10-12 meeting include using the 52 areas or approximately 1,300 field offices or zip codes for site selection. The advantage of the office level is that there are a large number of field offices for selecting a stratified, random sample and it is also a distinct physical unit. However, it was noted that some clients are doing more reporting on the phone and internet and may not connect with a single field office. The automated system for reporting earnings will also likely affect beneficiary contact with field offices. From an administrative perspective for the evaluation, it might be easier to work with field offices, so I would suggest using field offices in selecting a stratified, random sample within which individuals would be randomly assigned to WISP.

6.2 How many sites should be selected?

Again, I do not have a definitive recommendation for the number of sites that should be selected, in part because this will depend on other decisions made in the evaluation about stratification of the sample and how impacts will be evaluated for particular subgroups of interest. It will also depend on estimates of the size of the beneficiary stock and flow at the time that WISP will roll out and during the period of random assignment. The length of the period of random assignment will also be a factor to consider. In the 2-10-12 meeting, it was suggested that a sample of 10% of sites (130 field offices) be considered.

6.3 What programs are most likely to interact with WISP?

The programs most likely to interact with WISP include SSI, Medicaid buy-in (and possibly other policy interactions with the Affordable Care Act), and components of the current intervention (TWP, terminated beneficiaries in EXR, and EPE). The potential for interactions with WIPA (Work Incentive Planning and Assistance) was also raised in the TAP meeting, although there was some debate as to how extensively the WIPA services are used.

Site Selection
6.4 Should/Could interstate program variation be exploited to minimize these interactions? Interstate program variations can be exploited to understand the interactions of WISP with other policy changes (such as those coming under ACA), although it is not clear how they would minimize the interactions. This is something to consider in the random, stratified selection of sites for the evaluation. This program variation and its relationship to program impacts could be explored in nonexperimental analyses, such as multilevel modeling.

Timely Medical CDRs

7.1 Should WISP conduct timely medical CDRs for the treatment and/or control/comparison groups?

In the 2-10-12 TAP meeting, the different options for conducting timely medical CDRs (listed as Options 1, 2, 3 and 4 in the briefing materials) were discussed. It is clear that Option 2 is not viable. The reasonable choices seem to be between Option 1 and Option 3, and in general, it would probably be best to go with the option that is most likely to be national policy in the future—this is Option 1, where medical CDRs are conducted for both treatment and control groups. Because it is not clear yet how resource constraints will limit this option, I do not have a definitive recommendation between Options 1 and 3, although Option 1 seems to be the most desirable from an implementation and policy perspective.

Information Dissemination and Training

8.1 What should the initial and primary medium(s) of contact be with WISP subjects?

The initial and primary mediums of contact with WISP subjects will depend in part on the evaluation design (e.g., stage of randomization). It would be best to ensure multiple mediums of contact—letters by mail, followed up by phone calls and/or invitations to schedule an appointment with the field office are possible strategies. Importantly, I think it will be important to try to track and/or get confirmation from WISP treatment group members that they received the communications and understand the changes that are coming for them with WISP. The processes will also likely be different for current beneficiaries vs. newly entering beneficiaries. It was mentioned in the 2-10-12 meeting that robocalls could be used (possibly with current SSI/SSDI beneficiaries?), and that local stakeholders could also play a role in informing individuals about WISP.

8.2 Who should treatment subjects contact to ask questions about the intervention?

If there a toll-free number that current SSI/SSDI beneficiaries already use to get questions answered about their applications and benefits, it would be good to use this number and allow for a phone option/box to speak with someone about WISP. In this case, it would be important to ensure that operators who respond to these calls have sufficient training about how WISP will work. Another possibility is to establish a designated phone line just for WISP that could also refer individuals to persons knowledgeable about WISP in the nearest field office.

8.3 Should there be checks on whether WISP subjects received or understood the outreach materials that they were sent?

Yes, as indicated above, efforts to document individuals' understanding of WISP should be a formal part of the evaluation by including questions on the survey of beneficiaries as recommended. In addition to including questions on the survey of beneficiaries (or in lieu of them if a survey is not conducted), operators answering phone calls or field office workers speaking to individuals about WISP could be asked to complete a question (or a few questions) rating their perceptions of these individuals understanding of WISP and the outreach materials. However, I do not see this as essential; this could also be gauged through interviews or focus groups with program staff that might be easier to arrange.

Information Dissemination and Training

8.4 Does the TAP have any recommendations regarding training for WIPA organizations, TTW, providers, or any other entities that will interact with WISP subjects?

To the extent possible, outreach materials should also be sent to organizations that interact with SSI/SSDI beneficiaries so that they are aware of the demonstration and the changes it implies for beneficiaries. Individuals in these organizations should also be given a phone number that they can call to get additional information and/or clarifications about WISP provisions. It will be important to minimize the amount of misinformation that may circulate once word about WISP gets out by mouth. For this reason, it may also be wise to ask these other organizations not to directly provide information about WISP to those inquiring but rather to refer them to the appropriate program staff or operators handling these inquiries. If training sessions on WISP are held for workers in field offices, it might be helpful to invite representatives of these organizations as well.

8.5 How will these entities identify WISP subjects?

See above; I recommend that they correspond directly with SSA and/or its field offices with their questions about WISP or about who is in the WISP demonstration.

External Validity

9.1 What would be the key features of a national version of WISP?

As currently planned, the WISP demonstration is designed to go forward with features that are expected to be the same provisions that would be implemented in a national version of WISP. Of course, part of the reason for conducting the evaluation is to ensure that these features are going to work as intended and to allow for the possibility of subsequent modifications that would improve the program's functioning and the results (impacts) that are achieved (as findings become available on the program's implementation and impacts).

9.2 How will the TAP's other recommendations affect the evaluation's external validity, especially in relation to a national version of WISP?

As noted in the general comments section, the sampling design and decisions about where to randomize (i.e., at what level/stage) will be important to establishing external (and internal) validity in the evaluation. A nationally representative sample for this study is being recommended to aim for the greatest degree of external validity relative to a national version of WISP.

TAP Member #6

General Comments

Which items are most important to the WISP evaluation?

The impact of the WISP on overpayments and incorrect suspension of benefits and the impact of the WISP on SSA administrative burden (cost) are the two most important items. Unfortunately I do not believe the WISP alone is a powerful enough intervention to change beneficiary work behavior. I would be very happy to be proved wrong.

Were there any key topics not covered in the brief materials and the questions below that you would like to raise here?

There was no discussion of conducting an implementation pilot for a limited period of time to work out any kinks prior to full roll out. Even a very time limited (6 months) implementation pilot can help work out crucial details with a small and manageable number of beneficiaries. Early implementation issues can undermine a demonstration. For example, a significant number of overpayments or incorrect benefit suspensions can taint the intervention in the minds of beneficiaries and other key stakeholders. It is very difficult to win back folks trust if there is an initial perception that the demo is not working as advertized. The four state pilots that laid the ground work for BOND identified numerous potential implementation issues. Because of the small numbers these issues were manageable and resolved overtime.

I am not suggesting the WISP needs multi-year pilots to work out the implementation issues. The WISP in some ways is a less complex demo. However a 6 to 12 month limited process pilot might be wise.

Historically with demos and programs, SSA and it's contractors have not done a good job engaging the state or local disability service providers and advocates. SSA and it's contractors have tended to perceive local systems as uniform and not recognized there maybe huge differences between states and counties. As a result they have sometimes been surprised by the response. Engaging local stakeholders through WIPA, the IL centers, VR or others can go along way helping SSA and contractors with the implementation of demos like the WISP.

Other comments?

From a policy perspective, evaluating the WISP as a potential platform for the BOND offset provisions makes the most sense to me. It has the potential to:

- Reduce administrative burden for SSA.
- Reduce overpayments and incorrect suspensions that harm beneficiaries.
- Combined with an offset would provide clear incentives for beneficiaries to increase employment.

In terms of marketing change to the advocates and the disability community this is a very compelling picture.

Research Questions

1.1 Do the five questions cover all the outcomes of interest to evaluators and SSA?

I believe the five core questions cover the most important outcomes. I have added some items in 1.2 that might enhance understanding of the impact of the WISP. Question 5 is clearly the most difficult to measure in a time limited demo. However, it is very important that the WISP attempt an assessment of induced entry, otherwise the SSA actuaries will fill the void with the most conservative estimates. I do not have the expertise to suggest a methodology. The approach suggested in the meeting of using data from when SGA was significantly increased to model the potential impact sounded promising.

1.2 Which additional outcomes should be incorporated into the research questions? Should any of those questions be given more weight than the five core questions in the original solicitation?

I would suggest the following additional outcome or perhaps sub-outcome.

- Did WISP employment outcomes vary for individuals participating in the following programs: WIPA, VR, Ticket to Work or the State Medicaid Buy In.

Dissemination/Policy Relevance Guidelines

2.1 When does the evaluation need to produce key findings to policymakers?

I recommend that the evaluation of potential administrative cost savings and impact on improper payments occur as soon as is practical. It could perhaps be possible to have good data after two years from implementation. This might allow policy makers to consider WISP administrative impacts along with BOND employment outcome data. For example, the BOND might demonstrate that SSDI beneficiaries are more likely to work with an offset but administering an offset plus a TWP was administratively burdensome. Data from the WISP regarding administrative cost savings might provide policy makers options for combining the two.

Regarding effects on work behavior, it might be expected that the impact of WISP may take some time to be apparent. The maximum benefit of the WISP does not occur till the beneficiary has used up their EPE. Therefore it might take four to five years for the impact on earnings to emerge.

If the WISP enrolls beneficiaries in their EPE or after their EPE there maybe a potential for impacts on earnings to emerge more rapidly.

2.2 What should the follow-up period be on those findings?

It seems administratively relatively easy to track beneficiaries long term. The results might be illuminating after five, seven or ten years, in particular on the impact of ongoing connection to the program.

2.3 Will other stakeholders demand findings and information that policymakers would consider less important?

Advocates maybe particularly concerned about the impact of the loss of the TWP for WISP participants. They may want data on any adverse effects or harm to individuals.

Dissemination/Policy Relevance Guidelines

2.4 What information would other stakeholders finding interesting and how should it be disseminated to them?

I strongly recommend that WISP engage the following stakeholders at the state or county level,. Stakeholders such as VR, community rehab providers, WIPA, ENs, Independent Living Centers, and local mental health or disability state agencies. These folks will likely feel aggrieved if the WISP is implemented without prior engagement. These stakeholders have the capacity to undermine the WISP if they do not believe it is in their clients interests, so you need them on board.

I strongly recommend in person outreach and engagement to key leadership in those groups.

Evaluation Design

3.1 Should WISP use an experimental or non-experimental design?

I recommend experimental design at the individual level with the following assumptions:

- Individual random assignment does not present SSA with significant administrative issues in tracking and managing beneficiaries within the WISP.

- The demo is limited to a number of sites (similar to BOND) where stakeholders and supporting professionals (such as WIPA , VR) can be made aware the demo is occurring for some of their beneficiaries.

3.2 If a non-experimental design should be used, then what type of design offers most promise for WISP design specifically?

I do not feel qualified to comment.

3.3 What are the implications for the choice of design on both internal and external validity?

Based on my field experience I believe the impact of the WISP on employment behavior will be modest. Therefore, if there is a significant but small impact it is important that the results not be open to question because of the lack of random assignment.

Evaluation Design

3.4 If random assignment is used, should it be individual or site level?

See above.

3.5 Given the selected design, which evaluation components should be emphasized over what time frames?

Administrative savings and impact on

Sample

4.1 What should the recommended sample size be (relates to the demonstration design)?

I do not feel qualified to respond.

4.2 Should subgroups be targeted along the work path? If so, are the subgroups originally suggested by OMB and SSA appropriately selected?

I strongly recommend that SSA select subgroups along the work path. It is reasonable to expect that the WISP will have different impacts on work behavior in the TWP, EPE and in termination. In particular WISP participants in the TWP period may have lower earnings than comparison group members with a TWP. On the other hand WISP participants who are beyond the EPE would be expected to be more likely to work above SGA than comparison group members.

4.3 What should the minimum detectable effects be set for producing key policy impacts? What levels of statistical power would be recommended?

I do not feel qualified to comment on this issue.

Sample

4.4 What are the implications of SSA subsample division on the demonstration's external validity?

I do not feel qualified to respond.

4.5 What are the implications of SSA's age restriction for the evaluation?

I conditionally agree that the demonstration should exclude DI beneficiaries who are 55 or older. This group seems the least likely to change their work behavior as a result of WISP. Given that the impact of WISP is likely to be small anyway including this group may reduce the possibility of a statistically significant finding.

My field experience in VR suggests beneficiaries 55 and above are less likely to work. However I am not familiar with the overall return to work data for beneficiaries ages 55 and above. If my assumption that DI beneficiaries aged 55 and above are less likely to return to work is not supported by SSA data then I would withdraw this comment.

4.6 What other subgroups should be targeted (e.g., SSI recipients and those in the Medicaid buy-in)

I would strongly recommend that con-current SSDI/SSI beneficiaries be targeted because of the implications of WISP for this sub-group. Furthermore concurrent beneficiaries tend to become DI only beneficiaries over time as they earn work quarters, so the WISP has significant policy implications for this group.

I believe the Medicaid Buy In group is less significant for policy makers for the following reasons. Only about 30 states have buy in programs. There is significant variation in the design and rules of the state buy ins and differences in levels of participation. Some buy ins are no longer active. It might be very difficult to draw and policy inferences from this group because of the state level variation.

Sample

Data Collection

5.1 What data sources are most important for addressing research questions?

Outcomes: SSA administrative data, federal DOL wage data, IRS wage data

Services: WIPA data, VR 911 reporting, Medicaid claims data, Ticket, EN reports

5.2 Which administrative data sources can be leveraged and what are their advantages?

Medicaid claims data may also allow you to determine if beneficiary received supported employment services through state mental health or DD services. This is important contextual information to understand outcomes. For example DD supported employment consumers rarely work above SGA and the WISP is highly unlikely to change that.

5.3 Will survey data be needed?

I do not have any strong recommendations in this area. My bias is against any significant reliance on survey data especially regarding benefits and work behavior.

I would suggest that the WISP conduct survey's to determine beneficiary understanding of the WISP. I expect beneficiary knowledge and understanding of the WISP to vary widely across participants especially those with cognitive disabilities.

It might also be useful to survey beneficiaries around aspects of their employment that cannot be captured through administrative data sources such as; employer provided health insurance, retirement, 401K, sick time etc. However, the value of this data should be balanced with the costs associated with collecting survey data.

Data Collection

5.4 When should outcomes in the data be measured?

Not sure I have enough information to respond.

5.5 How should savings to Medicare and private health insurance enrollment be measured?

I assume Medicaid and Medicare claims data is the best source of data for possible cost savings. I am not familiar enough with these data sets to know the potential and limitations of this data set.

I know of no way to measure enrollment in private health insurance other than directly surveying the beneficiary.

5.6 How should demonstration fidelity be measured?

The timeliness of the suspension and restarting of benefits should be a prime measure of fidelity across sites. Delayed suspensions resulting in overpayments and delays in restarting benefits will undermine beneficiary confidence in the WISP. Significant variations across sites because of SSA staffing issues would be a major issue.

Site Selection

6.1 What constitutes a site?

I would strongly recommend sites that conform to the county level or state level local government structures. This would make engagement of local much easier because there will likely be one VR, one IL center, one WIPA etc. It will also allow the evaluators to characterize sites around other variables that may interact with the WISP. For example:

Site A: Rural, few active ENs, limited WIPA access, limited VR presence and a very active IL center.

Site B: Urban, many active ENs, strong VR program, limited WIPA and limited IL presence.

Such site context maybe useful to policy makers especially if there are variations in outcomes across sites.

6.2 How many sites should be selected?

I do not feel qualified to respond.

6.3 What programs are most likely to interact with WISP?

High interaction (meaning a large number of beneficiaries): VR, Medicaid, Medicaid funded community MH and DD services

Medium interaction: WIPA, IL centers (in both cases because of limited resources)

Low or varied: ENs (there are still far too few ENs to result in significant interaction)

Site Selection

6.4 Should/Could interstate program variation be exploited to minimize these interactions?

No. These interactions are part of the policy environment and these interactions should be of interest to policy makers.

Timely Medical CDRs

7.1 Should WISP conduct timely medical CDRs for the treatment and/or control/comparison groups?

No for the following reasons:

- Implementation of timely CDRs in the sites could be perceived negatively by beneficiaries and the advocacy community. The mandatory nature of the WISP is already a hard sell for SSA. This may engender more fervent opposition to the demonstration.

- Implementation of timely CDRs is a significant intervention in it's self. It is possible implementation of timely CDRs could suppress beneficiary work behavior.

Trade off: If SSA determines it must implement timely medical CDRs I recommend they be implemented for both the treatment and comparison group, so any suppressive effect on work behavior would occur across groups.

Information Dissemination and Training

8.1 What should the initial and primary medium(s) of contact be with WISP subjects?

Ideal: A letter written in accessible language, followed by a phone contact from the local WIPA or a local entity with the expertise to fully explain the demo. The initial contact should allow for an in person follow up with the local WIPA or a contracted local entity to explain the demo. For the WISP to have any possibility of impacting beneficiary work behavior, I believe this level of engagement is necessary.

Next Best: A letter written in accessible language including contact information for the local WIPA or a local entity with the expertise to fully explain the demo.

8.2 Who should treatment subjects contact to ask questions about the intervention?

WIPA (SSA would have to provide the WIPA resources to be able to respond adequately in the sites)

8.3 Should there be checks on whether WISP subjects received or understood the outreach materials that they were sent?

Yes. This could be a huge issue. I fear the overwhelming majority of beneficiaries will not read or not fully understand even the best outreach materials. This is why I recommend a telephone contact with the possibility of an in person follow up. I recognize the costs associated with the above may be a factor.

Information Dissemination and Training

8.4 Does the TAP have any recommendations regarding training for WIPA organizations, TTW, providers, or any other entities that will interact with WISP subjects?

Training and outreach should be targeted to WIPA, VR, ENs, local independent living centers and local community rehabilitation providers. I would suggest a three tier approach. First outreach to the leadership of the organizations outlined above. Ask for their input on how to best outreach and train their staff and who needs to be trained. Second, provide webinar or video conference presentations to inform larger groups as identified by the local partners. Third as part of a small contract have the WIPA or another local qualified contractor be the local trainer on WISP.

8.5 How will these entities identify WISP subjects?

WIPA, ENs and VR will likely identify WISP participants in the course of their work, though contact with the beneficiary, through the BPQY provided by SSA or through contact with the claims rep. For example, typically a VR counselor will confirm a beneficiaries SSDI status through a BPQY. This then informs the counseling process and job goal.

External Validity

9.1 What would be the key features of a national version of WISP?

From a national policy perspective, the WISP as a potential platform for the BOND offset provisions makes the most sense to me. It has the potential to:

- Reduce administrative burden for SSA.
- Reduce overpayments and incorrect suspensions that harm beneficiaries.
- Combined with an offset would provide clear incentives for beneficiaries to increase employment.

9.2 How will the TAP's other recommendations affect the evaluation's external validity, especially in relation to a national version of WISP?

I don't understand the question.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC