Design Options for an Impact Evaluation of the Participant Assistance Program

Final Report

December 17, 2013

Annalisa Mastri
Jeanne Bellotti
Julie Bruch
Grace Roemer
Beenu Puri

This report was prepared for the U.S. Department of Labor (DOL), Office of the Assistant Secretary for Policy, Chief Evaluation Office by Mathematica Policy Research, Inc., under Contract No. DOLQ129633249 Task Order No. DOLU129634057. The views expressed are those of the authors and are not be attributable to DOL. They do not represent DOL, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

MATHEMATICA
Policy Research

This page has been left blank for double–sided copying.

Contract Number:
DOLQ129633249/
DOLU129634057

Mathematica Reference Number:
40172.210

Submitted to:
Chief Evaluation Office
U.S. Department of Labor
200 Constitution Ave., NW
Washington, DC 20210somewhere
Washington, DC 22222
Project Officer: Celeste Richie

Submitted by:
Mathematica Policy Research
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005
Project Director: Jeanne Bellotti

Design Options for an Impact Evaluation of the Participant Assistance Program

Final Report

December 17, 2013

Annalisa Mastri
Jeanne Bellotti
Julie Bruch
Grace Roemer
Beenu Puri

This report was prepared for the U.S. Department of Labor (DOL), Office of the Assistant Secretary for Policy, Chief Evaluation Office by Mathematica Policy Research, Inc., under Contract No. DOLQ129633249 Task Order No. DOLU129634057. The views expressed are those of the authors and are not be attributable to DOL. They do not represent DOL, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

MATHEMATICA
Policy Research

This page has been left blank for double–sided copying.

CONTENTS

TABLES

EXHIBITS

This page has been left blank for double–sided copying.

EXECUTIVE SUMMARY

The Chief Evaluation Office (CEO) at the U.S. Department of Labor (DOL) contracted with Mathematica Policy Research to conduct a study to explore potential research designs for determining the impact of the participant assistance program administered by the Office of Outreach, Education, and Assistance (OEA) within DOL's Employee Benefits Security Administration (EBSA). Through its participant assistance program, trained Benefits Advisors (BAs) work in the agency's field offices to provide outreach and direct assistance to employees, employers, benefit plan sponsors, service providers, and other stakeholders who contact EBSA with benefits-related issues. The study involved development of three key components: (1) a program logic model, (2) designs for an objective evaluation to study the impact of the program, and (3) suggestions for refinements to OEA's performance measurement process.

This report focuses on two potential impact evaluation designs that would test—in slightly different ways—the impact of receiving a referral to EBSA's website compared with receiving BA assistance through the telephone hotline, which is currently the predominant mode inquirers use to access services. OEA has indicated a desire to increase participants' use of its website to acquire information to resolve inquiries on their own and to submit web inquiries for further assistance. If this approach were successful at enabling participants to resolve relatively simple questions on their own without BA assistance, this could allow BAs to focus on inquiries that require specialized knowledge of benefits-related laws. In addition, inquiries submitted through the website require less data entry on the part of BAs, resulting in less time spent per inquiry.

An impact evaluation of these models would answer the research question "What is the impact of referring inquirers to the EBSA website on their customer satisfaction and knowledge of and access to their entitled pension/health benefits, compared with receiving BA assistance through the telephone hotline?" The results from this analysis could be used to inform management decisions about future operation of the program. The evaluation would also enhance the program's knowledge of its customer base and inform ways to target future outreach activities by providing information about the characteristics of the participants seeking BA assistance. In addition, the evaluation could examine how program-level outputs change when a portion of inquirers is diverted to the website. An implementation study could provide insights into the operational successes and challenges of a web referral service model, and a cost study could assess the cost of such a model relative to delivering telephone services as usual.

The report discusses implementation considerations for this type of evaluation, data collection needs, potential evaluation sample sizes, and the strengths and drawbacks of the two service delivery options. It also discusses other potential impact designs that were considered, including an evaluation of the net impact of BA activities on participants' knowledge of benefits rights and access to entitled benefits. Although CEO decided not to pursue a net impact study at this time, implementation factors for such a design are included in the report to document what was considered. Finally, the report presents recommendations for possible revisions to program performance measurement strategies, focusing on suggested revisions to the five priorities identified by OEA in fiscal year 2013 and additional priorities identified through information-gathering activities used to develop the logic model.

This page has been left blank for double–sided copying.

I. INTRODUCTION

The Office of Outreach, Education, and Assistance (OEA) within the Employee Benefits Security Administration (EBSA) of the U.S. Department of Labor (DOL) provides outreach and assistance related to pension and welfare benefits. Through its participant assistance program, trained Benefits Advisors (BAs) work in the agency's field offices to provide outreach and direct assistance to employees, employers, benefit plan sponsors, service providers, and other stakeholders who contact EBSA with benefits-related issues.

Given the importance and breadth of the program's reach, EBSA collects extensive data on the work conducted by BAs. These include data on the types of inquiries received, services provided, the ultimate outcomes of each inquiry, and customer satisfaction with the services they receive from BAs, among many others. These data collection efforts support the program's continued goals of monitoring program performance over time and continuously improving services. A rigorous impact evaluation of the work of BAs could supplement the important contributions of these data collection and monitoring efforts.

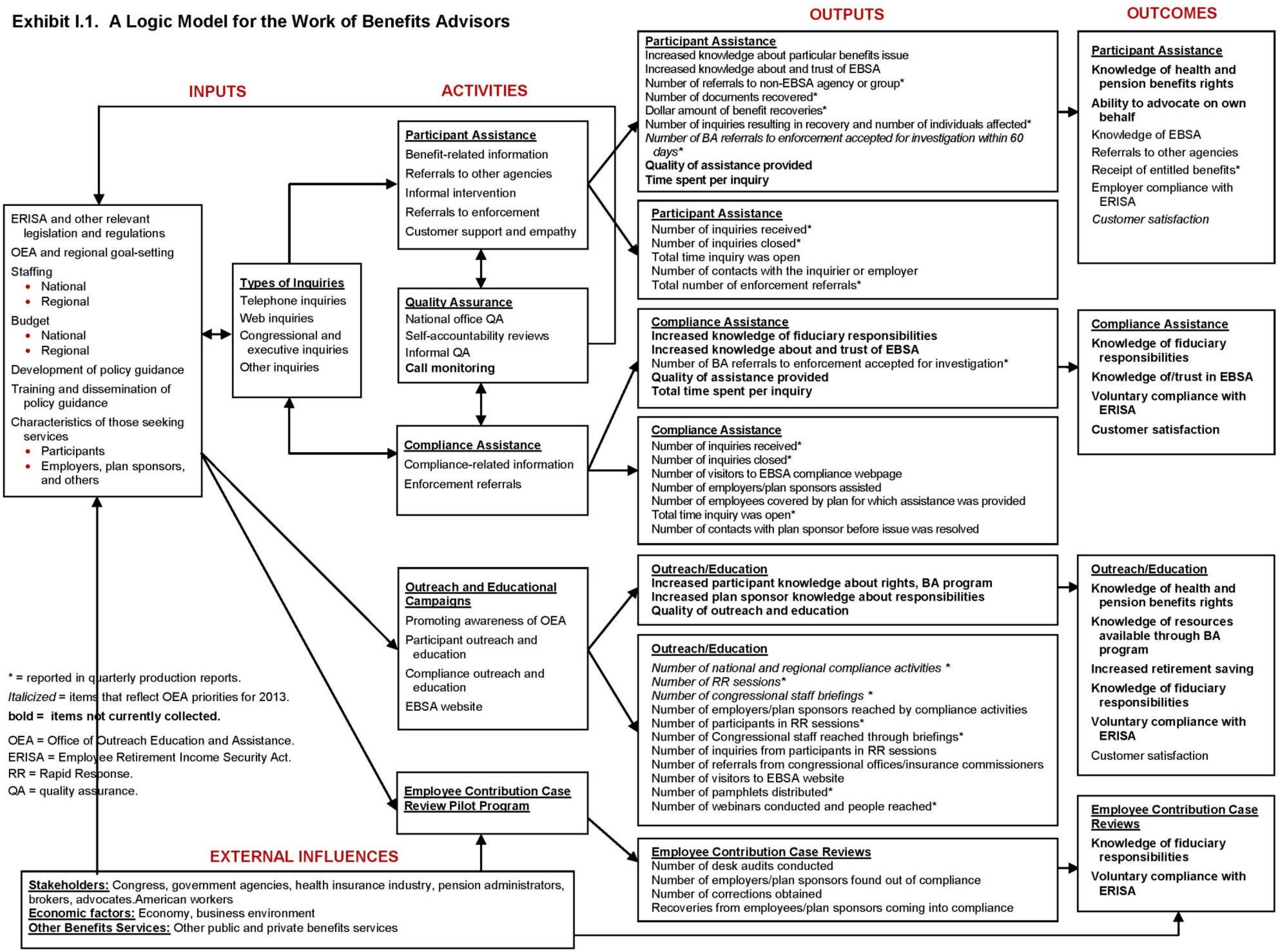
To this end, the Chief Evaluation Office (CEO) at DOL contracted with Mathematica Policy Research to conduct a study to explore potential research designs for learning about the impact of activities conducted by BAs. The first step in this process was to develop a program logic model. Using this model as a starting point, we then developed designs for an objective evaluation that could be used to study the impact of assistance provided by BAs. We also examined OEA's fiscal year (FY) 2013 performance measurement process and developed suggestions for refinements that could help the program better track its progress toward key goals over time. Throughout this study, we have conducted numerous discussions with CEO, OEA, and EBSA's Office of Policy Research (OPR), as well as a panel of four experts who were part of a technical working group (TWG). These discussions focused on the development of the logic model, possible impact evaluation designs, and performance measurement strategies. This report attempts to capture the information, preferences, and other feedback obtained during those discussions.

This chapter sets the stage for discussions of possible impact evaluation designs and suggested revisions to OEA performance measurement. We begin with an overview of the OEA program logic model. We then discuss the rationale for the recommendation that possible evaluation designs focus on the impacts of BAs' direct participant assistance activities and provide a description of the key outcomes of interest to an evaluation of those activities. The discussion then turns to exploring the range of evaluation options that might fit within the structure of the program as it currently operates. The chapter ends with a road map of the rest of the report.

A. The OEA Program Logic Model

Exhibit I.1 presents the logic model for the work of BAs. The model is important for understanding the underlying organizational structure of the OEA program, how services are provided by BAs, and how the structure and services combine to achieve the program's ultimate outcomes. The goal of the logic model is to describe how the program operates in theory. Thus it

Exhibit I.1. A Logic Model for the Work of Benefits Advisors



INPUTS

ERISA and other relevant legislation and regulations

OEA and regional goal-setting

Staffing
- National
- Regional

Budget
- National
- Regional

Development of policy guidance

Training and dissemination of policy guidance

Characteristics of those seeking services
- Participants
- Employers, plan sponsors, and others

ACTIVITIES

Types of Inquiries

Telephone inquiries

Web inquiries

Congressional and executive inquiries

Other inquiries

Participant Assistance

Benefit-related information

Referrals to other agencies

Informal intervention

Referrals to enforcement

Customer support and empathy

Quality Assurance

National office QA

Self-accountability reviews

Informal QA

Call monitoring

Compliance Assistance

Compliance-related information

Enforcement referrals

Outreach and Educational Campaigns

Promoting awareness of OEA

Participant outreach and education

Compliance outreach and education

EBSA website

Employee Contribution Case Review Pilot Program

OUTPUTS

Participant Assistance

Increased knowledge about particular benefits issue

Increased knowledge about and trust of EBSA

Number of referrals to non-EBSA agency or group*

Number of documents recovered*

Dollar amount of benefit recoveries*

Number of inquiries resulting in recovery and number of individuals affected*

*Number of BA referrals to enforcement accepted for investigation within 60 days**

Quality of assistance provided

Time spent per inquiry

Participant Assistance

Number of inquiries received*

Number of inquiries closed*

Total time inquiry was open

Number of contacts with the inquirier or employer

Total number of enforcement referrals*

Compliance Assistance

Increased knowledge of fiduciary responsibilities

Increased knowledge about and trust of EBSA

Number of BA referrals to enforcement accepted for investigation*

Quality of assistance provided

Total time spent per inquiry

Compliance Assistance

Number of inquiries received*

Number of inquiries closed*

Number of visitors to EBSA compliance webpage

Number of employers/plan sponsors assisted

Number of employees covered by plan for which assistance was provided

Total time inquiry was open*

Number of contacts with plan sponsor before issue was resolved

Outreach/Education

Increased participant knowledge about rights, BA program

Increased plan sponsor knowledge about responsibilities

Quality of outreach and education

Outreach/Education

*Number of national and regional compliance activities **

*Number of RR sessions**

*Number of congressional staff briefings **

Number of employers/plan sponsors reached by compliance activities

Number of participants in RR sessions*

Number of Congressional staff reached through briefings*

Number of inquiries from participants in RR sessions

Number of referrals from congressional offices/insurance commissioners

Number of visitors to EBSA website

Number of pamphlets distributed*

Number of webinars conducted and people reached*

Employee Contribution Case Reviews

Number of desk audits conducted

Number of employers/plan sponsors found out of compliance

Number of corrections obtained

Recoveries from employees/plan sponsors coming into compliance

OUTCOMES

Participant Assistance

Knowledge of health and pension benefits rights

Ability to advocate on own behalf

Knowledge of EBSA

Referrals to other agencies

Receipt of entitled benefits*

Employer compliance with ERISA

Customer satisfaction

Compliance Assistance

Knowledge of fiduciary responsibilities

Knowledge of/trust in EBSA

Voluntary compliance with ERISA

Customer satisfaction

Outreach/Education

Knowledge of health and pension benefits rights

Knowledge of resources available through BA program

Increased retirement saving

Knowledge of fiduciary responsibilities

Voluntary compliance with ERISA

Customer satisfaction

Employee Contribution Case Reviews

Knowledge of fiduciary responsibilities

Voluntary compliance with ERISA

EXTERNAL INFLUENCES

Stakeholders: Congress, government agencies, health insurance industry, pension administrators, brokers, advocates.American workers

Economic factors: Economy, business environment

Other Benefits Services: Other public and private benefits services

* = reported in quarterly production reports.

Italicized = items that reflect OEA priorities for 2013.

bold = items not currently collected.

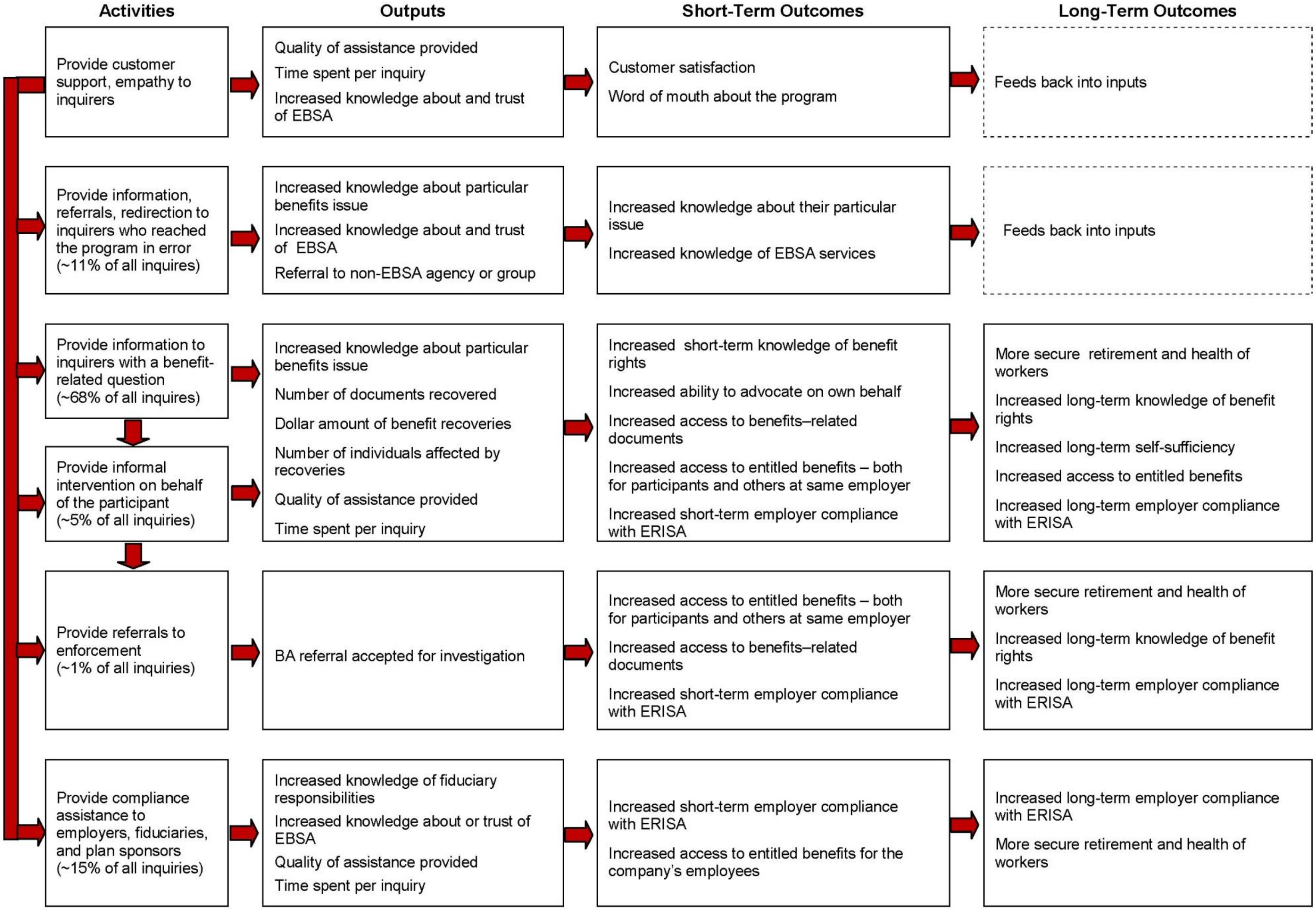
OEA = Office of Outreach Education and Assistance.

ERISA = Employee Retirement Income Security Act.

RR = Rapid Response.

QA = quality assurance.

Exhibit I.2. Participant Assistance Activities, Outputs, and Related Short- and Long-Term Outcomes



Activities	Outputs	Short-Term Outcomes	Long-Term Outcomes
Provide customer support, empathy to inquirers	Quality of assistance provided Time spent per inquiry Increased knowledge about and trust of EBSA	Customer satisfaction Word of mouth about the program	Feeds back into inputs
Provide information, referrals, redirection to inquirers who reached the program in error (~11% of all inquires)	Increased knowledge about particular benefits issue Increased knowledge about and trust of EBSA Referral to non-EBSA agency or group	Increased knowledge about their particular issue Increased knowledge of EBSA services	Feeds back into inputs
Provide information to inquirers with a benefit-related question (~68% of all inquires) Provide informal intervention on behalf of the participant (~5% of all inquiries)	Increased knowledge about particular benefits issue Number of documents recovered Dollar amount of benefit recoveries Number of individuals affected by recoveries Quality of assistance provided Time spent per inquiry	Increased short-term knowledge of benefit rights Increased ability to advocate on own behalf Increased access to benefits–related documents Increased access to entitled benefits – both for participants and others at same employer Increased short-term employer compliance with ERISA	More secure retirement and health of workers Increased long-term knowledge of benefit rights Increased long-term self-sufficiency Increased access to entitled benefits Increased long-term employer compliance with ERISA
Provide referrals to enforcement (~1% of all inquiries)	BA referral accepted for investigation	Increased access to entitled benefits – both for participants and others at same employer Increased access to benefits–related documents Increased short-term employer compliance with ERISA	More secure retirement and health of workers Increased long-term knowledge of benefit rights Increased long-term employer compliance with ERISA
Provide compliance assistance to employers, fiduciaries, and plan sponsors (~15% of all inquiries)	Increased knowledge of fiduciary responsibilities Increased knowledge about or trust of EBSA Quality of assistance provided Time spent per inquiry	Increased short-term employer compliance with ERISA Increased access to entitled benefits for the company's employees	Increased long-term employer compliance with ERISA More secure retirement and health of workers

includes factors that are currently measured, some that are not currently measured, and, in fact, some that are not measureable.

As the model shows, the BAs conduct a wide range of activities; however, their primary focus is providing PA to plan participants, employers, and plan sponsors, primarily by telephone. DOL was particularly interested in these activities, so the study team also developed Exhibit I.2 to provide further detail on the links between BA direct participant assistance activities and their intended outcomes.

These logic models form the basis for the impact evaluation designs discussed in this report. They reflect the types and nature of activities conducted by BAs, the outcomes those activities are intended to produce, and the factors that might facilitate or inhibit the program in its efforts to achieve those outcomes. We draw upon this critical information in our presentation of possible design options.

To inform the development of the logic model, Mathematica conducted a series of information-gathering activities from January to March 2013. These included reviewing program and policy documents; interviewing national office staff at OEA and the Office of Enforcement (OE); and interviewing regional directors, supervisory benefits advisors (SBAs), and groups of BAs in 10 field offices. A final logic model memo delivered to DOL in July 2013 provides an analysis of data from all of these sources and describes each component of the logic model in detail.

B. Focus on BAs' Direct Participant Assistance Activities

As reflected in Exhibit I.1, BAs conduct a wide range of activities, including providing direct assistance to plan participants, typically by telephone; providing compliance assistance to employers, plan sponsors, and other stakeholders; conducting outreach and education; and conducting employee contribution case reviews. Each of these activities could be the focus of a possible impact evaluation. For example, a study could be designed to assess the impact of compliance assistance on voluntary employer compliance with the Employee Retirement Income Security Act (ERISA). A study could also be designed to assess the impact of financial education campaigns on the retirement savings of pension plan participants.

During discussions with CEO, OPR, OEA, and members of the study's TWG, the consensus was that this impact evaluation should focus on designs related to the impact of BAs' direct and individual-level participant assistance activities on plan participants. The rationale was that these activities represent the largest proportion of BAs' efforts (BAs spend only about 5 percent of their time on education and outreach activities) and affect the largest number of individuals. Therefore, we confine the discussion of possible evaluation designs to those testing the impact of these activities.

As shown in Exhibit I.2, the extent of assistance that can be provided to inquirers ranges dramatically. Some individuals require referrals to other entities/agencies, others require informational assistance, and yet others require informal intervention by the BA on their behalf. When developing possible evaluation designs, we considered each of these types of activities. In collaboration with DOL, the study team determined that an evaluation should do the following:

- **Include inquirers who need benefits-related information.** More than two-thirds of inquiries require BAs to provide informational assistance on benefits-related questions or issues. These represent the heart of the BA program and should be included in an evaluation design. Although the study team considered focusing designs only on inquiries that require informal intervention by a BA on behalf of a participant (for instance, contacting an employer for a plan document), such a study would not capture a full and accurate picture of the impact of BAs' direct participant assistance activities.

- **Include inquirers who need informal intervention.** Inquiries requiring BAs to informally intervene on a participant's behalf represent less than 5 percent of the program's total inquiries. However, these cases often result in recoveries of plan documents and monetary benefits of substantial value to participants. Therefore, these types of inquirers should be considered a component of an impact evaluation design.

- **Include inquirers who need referral to enforcement.** Referrals to enforcement result from only about 1 percent of all inquiries. However, these represent instances in which BAs believe employers might be willfully noncompliant with ERISA, or an issue could affect more individuals than only the inquirer. BA referrals to the enforcement program within their regional office are the source of nearly 30 percent of enforcement cases. Therefore, these inquiries should be included in an evaluation design.

In addition to these inclusion criteria, EBSA, CEO, and TWG members thought certain types of inquirers should be excluded from a potential impact evaluation of BA services. In particular, they recommended that the evaluation should:

- ***Not* include inquirers who need compliance assistance.** Compliance assistance comprises about 15 percent of all inquiries nationwide. However, OEA determined and CEO agreed that its efforts to assist employers, plan sponsors, and other stakeholders should not be considered paramount in the impact evaluation design options. The program has collected data on its compliance assistance activities in the past that indicated a high level of success with these types of inquiries.

- ***Not* include inquirers who require a resource assistance resulting in a simple referral to another agency.** Although making referrals to other agencies is important, BAs do not typically have an opportunity to follow through on information and service provision until completion for these inquiries. Removing these from a study of the program would represent a slight change from normal program operations and, if the program wanted BAs to continue spending time to make these referrals, those types of inquirers could conceivably be included in an evaluation. However, because we would not expect BAs to influence the benefits-related outcomes of those individuals, the estimated impacts might be smaller than we would expect if the program were not serving them.[1] Appendix A contains further thoughts

[1] Chapter II discusses further considerations for including these types of inquiries in an evaluation testing the relative impact of services delivered via the website compared with services delivered by telephone.

on ways to potentially reduce the number of these calls by funneling them instead to the EBSA participant assistance website.

C. Outcomes of Interest to an Evaluation

Before discussing potential evaluation designs, it is important to consider the outcomes of interest for an evaluation. The range of outcomes and the ways they are measured have implications for the feasibility of different design options. For example, if data on outcomes are currently collected for only a portion of all inquirers, it might be necessary to expand the program's current data collection effort to support an evaluation.

The short- and long-term outcomes boxes of Exhibit I.2 include the key outcomes that might be of interest to a potential evaluation of BAs' participant assistance activities. They include outcomes measured at both the individual level—such as a person's ability to advocate on his or her own behalf—and the employer level, as follows:

- **Customer satisfaction.** OEA uses customer satisfaction as its primary measure of program success and EBSA would like this to be a key outcome for an impact evaluation. Therefore, this is a primary outcome of interest for an impact evaluation assessing the relative impact of different service delivery strategies.
- **Knowledge about benefits rights.** As mentioned in the OEA FY 2013 strategic plan and emphasized in interviews with regional office staff, educating inquirers about their benefits rights is one of the most important activities in which BAs engage and helps fulfill EBSA's mission. It would be difficult to assess actual knowledge of benefits rights without imposing substantial burdens on the inquirers—for instance, by asking them to answer several benefits rights-related knowledge questions—but an evaluation could capture customers' perceived knowledge of their rights.[2] Therefore, we recommend this as a key outcome for a potential evaluation.
- **Self-sufficiency or ability to advocate on one's own behalf.** Staff in regional offices emphasized during our information-gathering interviews the importance of helping participants learn how to advocate on their own behalf, both for the issue in question at the time and as issues arise in the future. Again, an evaluation could capture perceived ability to advocate for one's self. We recommend this as another key outcome for a potential evaluation.
- **Access to entitled benefits.** Beyond educating inquirers, BAs sometimes help participants recover benefits to which they are entitled, usually through informal intervention with the employer or plan sponsor on the participants' behalf; therefore, we recommend considering inquirers' access to entitled benefits or benefits-related

[2] Although in theory perceptions of knowledge of benefits rights should be related to actual knowledge of benefits rights, this has not been shown empirically as far as we know. Before any impact evaluation took place, a pilot effort could try to correlate perceptions of knowledge with actual knowledge, for instance by asking for perceived knowledge and then testing that with some objective questions. This would give a sense of how well the two are related, further enhancing the utility of the evaluation. It could also inform an understanding of the relationship between perceived and actual values of the other outcomes mentioned in this section, as observing their actual outcomes (for example, ability to advocate on one's own behalf) would be difficult.

documents as a key outcome of interest for a potential evaluation. In addition, BA intervention for one plan participant might affect numerous plan participants in a multiplier effect as improper actions by plan sponsors are corrected. Although there could be challenges to measuring this accurately, it should also be considered as an outcome for a potential evaluation.

- **More secure retirement and health of workers.** BAs ultimately hope to help inquirers achieve a more secure retirement and better health outcomes. Capturing these outcomes, however, could require significant long-term follow-up, as many inquirers seek assistance from the program while still of working age or before health issues emerge. An evaluation could ask inquirers about their perceptions of their health and retirement security. However, we do not recommend conducting extensive follow-up data collection to capture a more accurate measure of these outcomes because the cost and time delay of doing so would be large and potential data quality might be low.
- **Employer compliance with ERISA.** Increasing participants' knowledge of their benefits rights and the responsibilities of their employers or plan sponsors could result in greater employer compliance as informed employees hold them to account. However, capturing these outcomes would require challenging and costly collection of data from the employers of inquirers; therefore, we do not suggest that a potential evaluation include employer-level outcomes.

The program already has some measures of customer satisfaction, perceived knowledge of benefits rights, perceived self-sufficiency, and access to entitled benefits through its two main data collection sources, a survey conducted by Gallup Inc. and OEA's Technical Assistance Inquiry System (TAIS). The Gallup survey is conducted on a rolling basis using closed inquiries as the sampling frame. It collects information on customer satisfaction engagement and other outcomes, including individuals' perceived knowledge of benefits rights. TAIS is the information system that BAs use when handling inquiries. BAs record information about the inquiry, the caller's telephone number, the type of inquiry, and its resolution, among other data items. If the inquiry resulted in a document or monetary recovery, this is entered into TAIS as well.

Although the program already collects some data on outcomes of interest, both data collection systems would have to be modified—or a new data collection system developed by the evaluation team—to capture sufficient data to support an impact evaluation.[3] Chapters II and III contain further discussion about how key outcomes could be measured as part of the evaluation.[4]

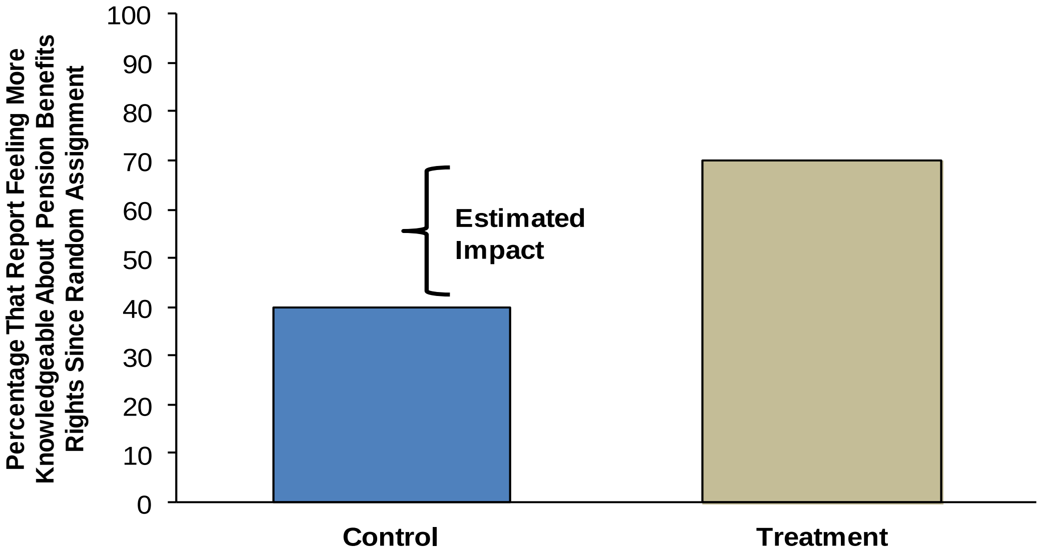
[3] OEA has expressed concern about the burden this data collection will have on BAs and customers given that it is not essential for the performance of the program's daily work. DOL and the evaluator will need to take this into consideration as decisions are made regarding a potential evaluation and the extent of data collection implemented.

[4] In addition, Chapter IV provides details on suggested changes to the measurement of outcomes for OEA's performance measurement system. Although there might be some overlap in the outcome measures needed for the purposes of an impact evaluation and performance measurement, there are also distinct differences in the measurement and data collection required for outcomes used to assess program impacts and those used to assess ongoing program performance.

D. Randomized Controlled Trials Versus Quasi-Experimental Impact Evaluation Designs

With the focus of the possible evaluations identified and the potential outcomes of interest in mind, the next step was to determine the feasibility and desirability of potential rigorous impact evaluation designs. The most rigorous way to conduct an evaluation of the impact of BAs' activities on participants would be to conduct a randomized controlled trial (RCT).[5] In a classic RCT, participants are assigned at random to different groups whose outcomes will be compared. Typically, those assigned to a treatment group continue to receive services as usual, while those assigned to a control group receive either no services or some variation on services. RCTs are considered the gold standard for determining program impacts because the random nature of the assignment ensures that there are no systematic differences, on average, between the treatment and control groups at the time of random assignment. Therefore, any differences in the observed outcomes of the two groups can be attributed to the impact of the treatment—which, in this case, would be access to BA services. Exhibit I.3 shows an example, using hypothetical data, of how impacts are estimated in an RCT.

Exhibit I.3. Example of Impacts from an RCT



Note: This exhibit uses hypothetical data to demonstrate an impact estimated from an RCT with an outcome such as feeling more knowledgeable about pension benefits rights.

[5] The study statement of work requires Mathematica to develop "2-3 evaluation designs to measure and test the impact of OEA, specifically BAs, and linking activities and outcomes to those impacts. At least one model shall not include a reduction of OEA service levels." The team asked DOL and the TWG whether a process and outcomes evaluation should be considered. CEO and OEA expressed a strong interest in an impact evaluation, and relayed a similar interest expressed by current members of DOL administration. TWG members raised concerns about an impact evaluation, given the relatively small size of the OEA budget and the possible need to deny services to inquirers. Nevertheless, CEO determined that Mathematica should focus its discussion on potential impact evaluation designs.

Second to RCTs in analytical rigor are various types of quasi-experimental designs (QEDs). These designs typically compare the outcomes of individuals who have received program services (the treatment group) to similar individuals who did not receive program services (a comparison group). The critical distinction between this type of design and an RCT is that, in a QED, individuals in the treatment group have chosen to receive services from the program, rather than being randomly assigned to receive them; in contrast, individuals in the comparison group have chosen not to receive program services. This introduces the potential that the two groups differ in more ways than just participation in the program and, as a result, this design is viewed as being less rigorous than an RCT. QEDs can come in a range of different formats, including but not limited to regression discontinuity designs, instrumental variable approaches, and propensity score matching. One of the most rigorous types of QEDs (which is still considered less rigorous than a well-implemented RCT) involves using a statistical matching process to select a comparison group from an existing data set that is as similar to the treatment group as possible on observed characteristics. The idea is that if the two groups are as similar as possible on observed characteristics, then any differences in their outcomes can be attributed at least in part to the program being examined. Unfortunately, however, there still could be unobserved differences between the treatment and comparison groups that cause differences in outcomes. Thus, one cannot state definitively that access to the program services is the sole, or even the main, cause of the differences in outcomes between the two groups.

Although we considered a QED and looked into several potential sources of data for such a design, we ultimately concluded and the TWG members agreed that a QED was not feasible. First, there are very few data sets that contain the types of outcomes that would be of interest to an evaluation of BA participant assistance activities. Table I.1 provides information on existing data sets that cover access to employer-provided benefits and Appendix B provides the wording of relevant survey questions in the data collection instruments used for these data sources. As illustrated in the table, the sources tend to ask questions about the availability and take-up of benefits, but not about potential issues with accessing entitled benefits. In addition, none of the available sources currently asks questions about perceived knowledge of benefit rights or ability to advocate on one’s own behalf, both of which we believe are important outcomes to examine in an evaluation of BA services.

In addition to the general lack of available data from which to draw a comparison group, a QED would require collecting extensive information about the characteristics of individuals who contact BAs for assistance in order to identify the individuals in the other data set who were similar; this would likely amount to more data than would have to be collected for an RCT. These characteristics would probably include items such as age, race/ethnicity, gender, annual income, union status, region, and other items. OEA and the two TWG members with the most extensive program knowledge found this potentially problematic because of the burden it would place on BAs to collect such information from each caller for an extended period. They also thought some callers would not be willing to provide such information. Although the RCT designs presented in later chapters do require the collection of some additional baseline data, the extent of that data collection would likely be more limited than would be required for a rigorous QED.

Table I.1. Existing Data Sets on Access to Workplace Benefits

Name of Data Set	Organization Collecting/ Housing the Data	Publicly Available (Y/N)	Sample Size	Description of Sample	Benefits-Related Topics Covered
Retirement Confidence Survey	Employee Benefit Research Institute	N	1,000 individuals	Random, nationally representative sample of individuals ages 25 and older	Retirement savings, retirement confidence, take-up of employer-sponsored retirement savings plans
Survey of Income and Program Participation, Retirement and Pension Plan Coverage Module	Census	Y	14,000 to 36,700 households	Sample of U.S. civilian noninstitutionalized population; all household members ages 15 or older are included	Retirement savings, type of retirement/ pension plan
Survey of Income and Program Participation, Medical Expenses/Utilization of Health Care Module	Census	Y	14,000 to 36,700 households	Sample of U.S. civilian noninstitutionalized population; all household members ages 15 or older are included	Health and insurance expenditures
Current Population Survey Annual Social and Economic Supplement	Census	Y	53,300 households	Sample of U.S. civilian noninstitutionalized population; all household members ages 15 or older are included	Pension/retirement income, participation in employer-sponsored retirement savings plans, employment-based health coverage, health expenditures
General Social Survey	NORC at the University of Chicago	Y	Approx. 3,000 individuals	Sample of English- and Spanish-speaking adults in U.S.	Receipt of fringe benefits, health insurance coverage

NORC = National Opinion Research Center.

E. Organization of the Rest of the Report

The rest of this report provides a detailed description of possible impact evaluation designs and presents recommendations for possible revisions to OEA performance measures. Chapter II discusses an RCT assessing the relative impacts of an alternative service delivery model that involves greater use of referrals to EBSA’s website. EBSA is committed to using new technology to enhance its service delivery, and making greater use of website referrals has the potential to serve more customers at lower cost; it could also result in increased BA job satisfaction. The chapter describes the considerations for implementing web referral service delivery models and a related impact evaluation of them. Chapter III presents a classic RCT assessing the impact of having access to BA assistance compared with not having access to it. CEO has decided not to pursue this design option at this time as a result of ethical and implementation concerns raised by EBSA and the TWG members. However, the design is presented in this report to document what was considered and why it was determined infeasible. Finally, Chapter IV presents recommendations for possible revisions to current program performance measurement strategies.

II. DESIGNS TO ASSESS THE RELATIVE IMPACT OF WEB REFERRAL SERVICE DELIVERY MODELS

OEA has indicated a desire to increase participants' use of its website both to submit inquiries to BAs and to acquire information so that participants can resolve inquiries on their own. Increased use of the website has the potential to increase the total number of participants assisted by the program, reduce the cost per participant served, and boost employee satisfaction and longevity. These considerations are especially important as the demand for BA services continues to rise with the passage of new legislation related to benefits rights; in 2014, OEA anticipates a surge of inquiries as a result of implementation of the Affordable Care Act (ACA). Moreover, increases in the use of the EBSA website might be achieved with minimal impact on participants' outcomes such as customer satisfaction and knowledge of benefits rights, and in fact participants' outcomes might increase with greater website use.

This chapter discusses a research design that explores the potential for making increased use of service delivery through EBSA's website. In particular, we discuss two impact evaluation designs testing—in slightly different ways—the impact of receiving assistance via EBSA's website compared with receiving services through the participant assistance telephone hotline.[6] Although the impact analysis portion of the evaluation would determine the effects of these different delivery strategies on participants' outcomes and satisfaction, cost and implementation studies would also be needed to learn whether these approaches offer the promise of being more cost-efficient and increasing BA job satisfaction.

As programs naturally evolve over time, they often try different service delivery strategies to determine the best approach for serving their customers. Rarely, however, do they have strong evidence on whether the resulting change had an impact on participants. The designs proposed in this chapter would use random assignment to determine what type of service delivery each inquirer would receive and, in this way, it would enable the program to capture rigorous evidence about the impact of web service delivery models on participants' outcomes, use of the website, and a host of other information. Importantly, because this design compares the relative impact of two service delivery approaches, it does not result in denial of service to anyone; everyone would have the opportunity to get services, either via the web or telephone.

In this chapter, we describe two potential random assignment designs for an impact evaluation of service delivery models that incorporate referrals to the EBSA website. We also discuss how this type of evaluation would be implemented, data collection needs, sample sizes and minimum detectable impacts (MDIs), and the strengths and drawbacks of the two potential web referral options. We conclude the chapter with a section on the importance of implementation and cost studies when comparing two service delivery models.

[6] Appendix D presents evaluation designs for two different alternate service delivery models: a model that would use junior BAs or receptionist staff to first categorize and prioritize calls as they were received, answering simple informational requests, and providing referrals as needed, but referring more complicated requests and emergencies directly to a BA; and a model with the same system of prioritization plus specialization of BAs into subject matters.

Summary of Designs to Assess the Relative Impact of Web Referral Service Delivery Models

- The designs would estimate the impact of receiving services delivered via the website (in one of two potential ways) compared with services delivered through the participant assistance telephone hotline.
    - In the web referral only model, telephone inquirers who did not have an emergency and have Internet access would be asked to participate in a research study. Then, those who volunteered to participate would be randomly assigned to one of two groups: (1) inquirers would be referred to the website for self-service and to submit a web inquiry if needed or (2) inquirers would receive telephone services from BAs in the usual manner.
    - In the web referral/telephone follow-up choice model, all inquirers who did not have an emergency would be randomly assigned to one of two groups: (1) inquirers would be given a choice to either access the website for self-service and to submit a web inquiry if needed or to await a phone follow-up call from a BA or (2) inquirers would receive telephone services from BAs in the usual manner.
- The designs estimate the relative impact of two service delivery strategies and answer the research question, “What is the impact of referring inquirers to the EBSA website (either with or without a phone follow-up option) on their customer satisfaction and knowledge of and access to their entitled pension/health benefits, compared to receiving services through the telephone hotline?”
- For both design options, baseline data would have to be collected from inquirers and a follow-up survey would have to be administered to collect data on inquirers’ service receipt and outcomes.
- These design options would enable the program to determine how changing the current service delivery strategy to incorporate web referrals would affect the following:
    1. Inquirers’ outcomes, such as customer satisfaction, perceived knowledge of benefits rights, access to benefits-related documents, and benefit recoveries
    2. Program-level outputs, such as the total number of inquiries received and closed, the average time inquiries were open, and average number of contacts made to resolve an inquiry
    3. The cost per participant of service delivery and BA satisfaction with the new model
- In the case of the web referral/telephone follow-up choice model, the evaluation would also provide information about inquirers’ preferred methods for accessing services; in other words, it would show the characteristics of individuals who actively chose to use the website when promised faster service compared with those who chose to wait for a follow-up telephone call from a BA. This information could help OEA target website referrals and maximize BA resources during periods of high inquiry volume.
- The information from the components of the evaluation could be used to inform management decisions as to how the program could operate in the future.

A. Models to Be Tested, Research Questions, and Hypotheses

The core of this evaluation design option would test differences in outcomes between individuals who received BA services delivered by the traditional telephone system and those who received a variation on service delivery that emphasized using the EBSA website for self-service and submitting web inquiries. There are two variants on this same basic service delivery model.

Web referral only. Under this service delivery model, inquirers to the BA hotline would first be screened to determine whether they had an emergency or did not have Internet access.[7] If either condition were true, the inquirer would be directed to a BA through the usual procedures. If both conditions were not true, the inquirer would be referred to the website for services; this includes both self-service—using the EBSA website to find information that resolves the issue—and submitting a web inquiry to which a BA would respond.

Web referral/telephone follow-up choice. Under this service delivery model, inquirers who did not have an emergency situation (irrespective of their Internet access) would be given a choice between (1) accessing the website for self-service and/or to submit a web inquiry or (2) to leaving their contact information for a BA to follow up with them by telephone. Inquirers would be informed that web inquiries are typically handled in one business day, whereas the telephone follow-up would require longer (OEA would determine the target it would set for conducting the telephone follow-ups).

There are two key advantages of these alternatives to the usual service delivery strategy. First, web referral—either with or without the telephone follow-up option—could result in some inquirers getting the information they need from the website and therefore not needing to submit a web inquiry in the first place, freeing BAs to spend additional time with the more complex inquiries that require their assistance. Although data on call lengths are not systematically collected, BAs said in interviews that from 75 to 90 percent of telephone inquiries can be handled in 5 to 10 minutes. To address these types of calls, the BA might provide a referral to a health insurer or other government agency, direction to written documentation on the EBSA website, or a quick explanation of Consolidated Omnibus Business Reconciliation Act (COBRA) notices. In short, these inquiries do not often require extensive assistance from a highly experienced BA; instead, the inquirers could be directed to the website to find that information on their own. Although BAs can provide this information in a matter of a few minutes, every minute counts when call volumes reach 250,000 or more inquiries per year. Therefore, a web referral model has the potential to enable BAs to spend more time on the types of complex issues that require their detailed knowledge about benefits rights, and free time for them to do more outreach or serve more inquirers. This could also potentially increase BAs' job satisfaction, as they can work on and resolve more complex cases and help develop a longer or more meaningful career progression that reduces staff turnover. While OEA expressed some concern that it would take more time to refer hotline callers to the website than to answer their question, a study of this design could help inform future policy decisions about whether website referrals are a reasonable service strategy that does not negatively affect inquirer outcomes. The program could use this information to decide if and under what circumstances automatic referrals that do not go through an evaluation screening process might be appropriate in the future.

The second advantage of the web referral approach is a reduction in the amount of data entry time needed on the part of BAs for inquiries submitted through the website. To submit a web inquiry, inquirers must fill out a form on the EBSA website that requests some basic information

[7] Because of a concern that some inquirers—particularly older individuals—might not have Internet access, OEA proposed that inquirers who reach the office by telephone be screened for whether they had Internet access before being referred to the website.

about the inquirer, the nature of the inquiry, and contact information so that a BA can call the inquirer if needed to resolve the issue. The information entered into the form automatically populates a TAIS entry for that inquiry. For a telephone inquiry, the BA would create a new TAIS record and enter this information while on the telephone with the inquirer. Thus, having the customer enter the information directly on the web form reduces the BA time spent per inquiry.

To determine the relative effectiveness of either of these web referral approaches relative to offering telephone assistance, we propose a random assignment evaluation in which half the inquirers would be assigned to one of the web referral models (the model selection would be made by EBSA and CEO) and the other half would be assigned to receive services through the telephone hotline. Because this design compares the relative effectiveness of two service delivery approaches, customer satisfaction—which is OEA's primary performance measure—could be examined as an outcome of interest. This adds an important dimension to this type of evaluation.

This evaluation would be designed to answer the specific research question:

- What is the impact of referring inquirers to the website (either with or without a telephone follow-up option) on their customer satisfaction and knowledge of and access to their entitled pension/health benefits, compared with receiving services through the telephone hotline?

It is unclear whether customer satisfaction would differ under these models; on one hand, some people prefer to access information using the Internet, especially if they have a simple question. On the other hand, people might be frustrated if they are referred to the website and are unable to resolve their issue or cannot easily figure out how to submit a web inquiry. Given the extent of information available on the website and the ability to submit web inquiries for in-depth issues, we hypothesize that web referral in either of these alternative service delivery models will not harm inquirers' outcomes related to perceived knowledge of benefits rights, self-sufficiency, and security of retirement and health.[8] In fact, if BAs have more time to focus on complex issues related to access to benefits, access to documents, and recovery of benefits, inquirers' outcomes might actually improve under the web referral models.

An evaluation of this kind would also generate useful information for the program about how receiving services through the website affects the individuals submitting inquiries. For instance, it would determine whether people with similar demographic characteristics and similar inquiries are able to resolve their problems by using the website as quickly and effectively as when speaking with a BA on the telephone. When combined with information about the implementation and cost of those services, the results would enable the program to make informed choices about its service delivery strategies.

[8] OEA did express concern that inquiries are often time sensitive and there is the possibility that an inquirer who does not chose the Internet option could miss an eligibility deadline if telephone follow-up is delayed. As discussed later in the chapter, customers who are approaching deadlines would have the option to identify their call as an emergency and be referred directly to a BA, avoiding any delay in sevice.

In addition to learning about the relative impacts of web referral compared with services delivered in the usual manner on participants' outcomes, the evaluation could also examine several other descriptive items of interest. These could include the characteristics of inquirers, such as age, which the program does not currently collect and which could be helpful in targeting future outreach efforts; the proportion of inquirers referred to the website who resolve their issues through self-service, submit a web inquiry, or do neither; and program-level outcomes, such as the total number of inquiries received and closed by mode of contact, whether web inquiries require more or less time to respond to than telephone inquiries, and BAs' satisfaction. In the case of the web referral/telephone follow-up choice model, the study would also provide information about inquirers' preferred methods for accessing services; in other words, it would show the characteristics of individuals who actively chose to use the website when promised faster service compared with those who chose to wait for a follow-up telephone call from a BA. This information could help OEA target website referrals and maximize BA resources, especially during periods of high inquiry volume.

B. Implementing the Evaluation Designs

Although they are closely related, implementation of the RCT designs to test the two web referral models and what they would demonstrate would differ slightly. Neither model would apply to inquirers who said they had an emergency—all such calls would be routed directly to a BA.[9] The web referral only model would also not be applied to people who reported no access to the Internet.

1. To What Would the Web Service Delivery Models Be Compared?

In both potential designs, individual inquirers to the program would be randomly assigned to receive either the web referral model or telephone services as usual. Telephone services as usual might not necessarily encompass an identical way of offering services in every regional office because, although the national office provides guidance to the regional offices, they are still allowed flexibility in their operations. As a result, some offices have evolved to deliver services in slightly different ways—for example, by having a receptionist answer all calls and open TAIS records before passing calls to BAs or by referring calls about certain issues to BAs with expertise in those issues. Because this is how the program ordinarily operates, it is important to preserve these differences in service provision across the regional offices during the course of the impact evaluation to ensure that the evaluation is able to estimate impacts that apply to the program as it normally operates. This also implies that procedures put in place for the evaluation should alter normal service delivery as little as possible.

2. How Would Random Assignment Work?

The process of randomly assigning participants would vary based on the option selected—web referral only or web referral/telephone follow-up choice. The difference between the two

[9] Discussions with OEA and the TWG members who were retired from EBSA indicated that a small fraction of inquiries are emergency situations. Nevertheless, it is important that those inquirers receive immediate assistance. The evaluation team would have to work with OEA to determine how to assess whether inquirers have an emergency; it might be sufficient to simply ask the inquirer if he or she has an emergency.

lies in who would participate in the study. In general, if a program is operating normally—which can include delivering services in a slightly different way from other programs—its customers would not have to consent to participate in a research study involving random assignment; that is, customers would not have to be informed that a study was taking place, what the possible outcomes of random assignment would be, and any implications of not agreeing to participate in the study.[10] However, OEA expressed a preference that, for the web referral only option, inquirers volunteer to participate, and only those who volunteer would undergo random assignment. Therefore, we incorporated this design feature into an evaluation of the web referral only model and discuss the implications.

a. Random Assignment Under the Web Referral Only Model

Under the design to test this model, all inquirers to the program would be diverted to contractor staff[11] who would first collect some basic information about the inquirer (for example, age, gender, race/ethnicity, and topic of inquiry) and then ask the inquirer two things: whether he or she had an emergency and whether he or she did not have Internet access at home. If the inquirer responded yes to either of these questions, he or she would be transferred immediately to a BA for assistance. If he or she responded no to both questions, the contractor staff would inform the inquirer about the study and encourage the inquirer to participate in it. If the inquirer volunteered to participate in the study, he or she would be randomly assigned to be transferred to a BA for assistance by telephone or be referred to the website for self-service with the option to submit a web inquiry. If the inquirer did not volunteer to participate in the study, he or she would be transferred to a BA to receive services by telephone. Exhibit II.1 provides a flow chart for this process.

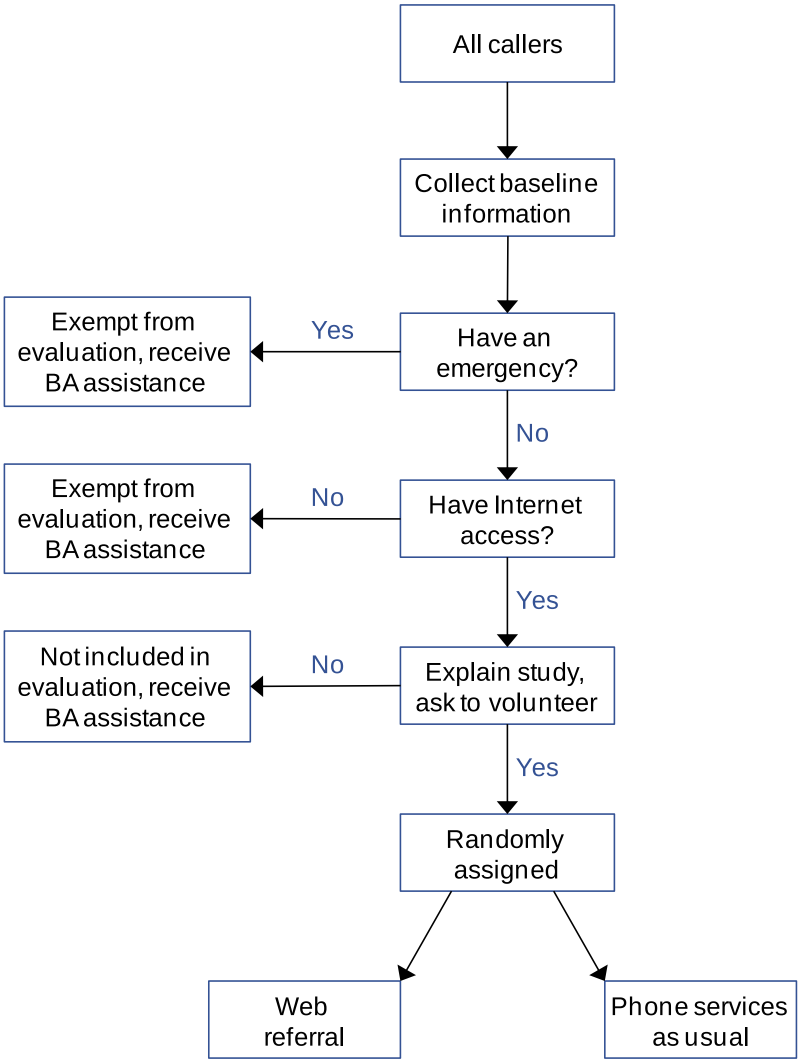
Staff conducting this process would be trained to deliver a standard approach to all inquirers in asking for personal information, asking about emergencies, informing inquirers about the study, and encouraging their participation. This would ensure that inquirers are fully informed of what the evaluation entails and that their information might be used for research purposes. Staff would also be trained to inform study volunteers who were randomly assigned to the web referral only group about their study group assignment, the website address, and perhaps some specific information about how to submit a web inquiry (to be determined by OEA).

Two outstanding issues would have to be resolved before implementing this evaluation design. The first is how to handle inquirers who were referred to the website but were not able to find the information they sought and/or were not able to submit a web inquiry. These customers

[10] However, if personally identifiable information were collected by or provided to a third-party evaluator, it would be advisable to inform customers that their information might be used for research purposes. An institutional review board governing the treatment of human subjects could make the ultimate determination of this. While inquirers are currently not informed that their information might be used for research (that is, the Gallup survey), this data collection is permissible under the Privacy Act. Program staff indicated that complaints from inquirers about the Gallup survey are very rare.

[11] We refer to contractor staff conducting intake throughout this chapter because it was determined in consultation with EBSA, CEO, and the TWG members that having BAs perform this role would add too much burden to their already heavy workloads.

Exhibit II.1. Study Intake for Evaluation of Web Referral-Only Model


All callers
Collect baseline information
Have an emergency?
Yes
Exempt from evaluation, receive BA assistance
No
Have Internet access?
No
Exempt from evaluation, receive BA assistance
Yes
Explain study, ask to volunteer
No
Not included in evaluation, receive BA assistance
Yes
Randomly assigned
Web referral
Phone services as usual

may call the program back for further assistance. OEA also indicated that some inquirers may "shop around" for answers by contacting different BAs or offices if they are unhappy with the assistance provided. If there were substantial numbers of these callbacks and they were assisted by BAs by telephone, this could compromise the integrity of the study groups, potentially calling into doubt the validity of the impact estimates. However, this could still provide useful information about the efficacy of making website referrals and suggest possible improvements to the EBSA website.

The second issue is whether and how to screen inquirers with compliance-related issues and inquirers who have reached the program in error, two groups that EBSA indicated should *not* be included in an impact evaluation. If EBSA were interested in the impact of requiring website use among employers and plan sponsors, inquirers with compliance-related issues could be put through the random assignment process. However, many inquirers with compliance-related issues are repeat callers who might have developed a relationship with a BA and call the BA directly; these inquirers would have to be directed to the hotline if they were to be included in the study, which might compromise the relationship between the BA and the employer or plan sponsor seeking assistance. If EBSA wanted to exclude compliance inquiries from the study, an additional question would have to be added to the intake process so those calls could be immediately routed to the BA hunt group or the BA with whom the inquirer has an existing relationship. For inquirers who have reached the program in error, it seems possible that a web referral option could be a good way to inform them that they have reached the program in error. In that sense, it might be worthwhile to include them in the evaluation of web referral models. EBSA, CEO, and the evaluator would have to determine how to handle these two issues before implementing the random assignment design.

b. Random Assignment Under the Web Referral/Telephone Follow-Up Choice Model

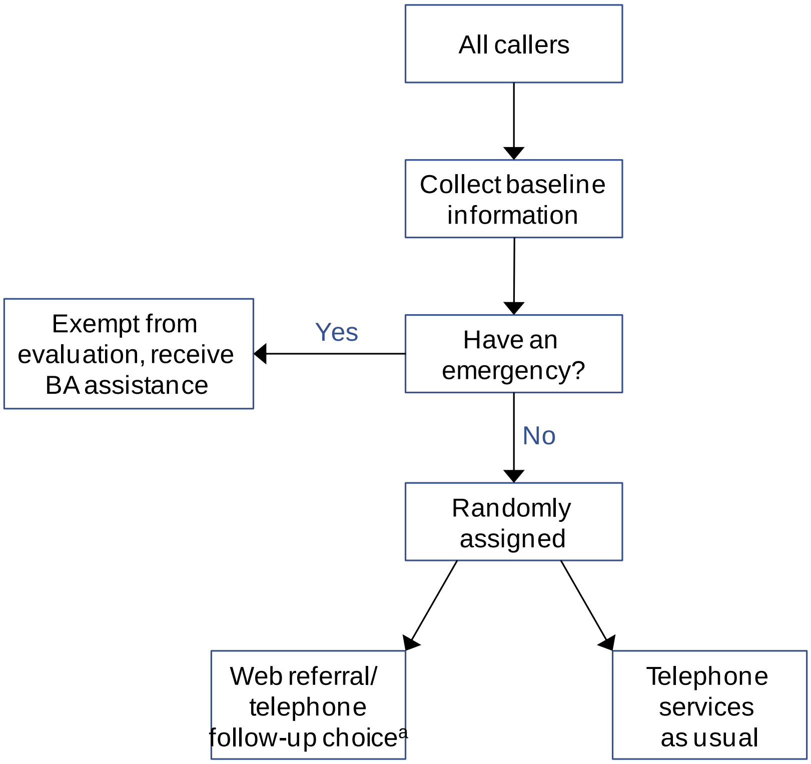
This design is very similar to the one for the web referral only option, with two key differences: First, only inquirers with emergencies would be exempt from the study, whereas inquirers without Internet access would be included; because all inquirers would have the option to wait for a follow-up telephone call from a BA, the concern about potentially including inquirers without Internet access in the study is mitigated. Second, inquirers would not be asked to volunteer for the study and instead would simply be randomly assigned to receive one or the other sets of services. EBSA staff indicated that they were comfortable not seeking volunteers for an evaluation of this model because inquirers in both study groups would have the possibility of receiving telephone services.

Thus, under the design to test this model, all inquirers to the program would be diverted to contractor staff who would first collect some basic information about the inquirer (for example, age, gender, race/ethnicity, and topic of inquiry)[12] and then ask the inquirer whether he or she had an emergency. If the inquirer responded yes, he or she would be transferred immediately to a BA for assistance. If the response was no, the contractor staff would randomly assign the inquirer to either (1) be referred directly to a BA for assistance or (2) choose between going to the website for services or waiting for a follow-up call from a BA. Given that consent is not needed for this type of study, inquirers would be unaware that the random assignment process was taking place. Inquirers in the web referral/telephone follow-up choice condition would be told that they could use the website for self-service with the option of submitting a web inquiry that would result in a BA response within one business day. They would also be informed that if they chose to await a follow-up telephone call from a BA, they could expect to receive it within a period longer than one business day (exact duration to be determined by OEA). The inquirer would make a decision while on the telephone with contractor staff about the preferred option. Exhibit II.2 provides a flow chart for this process.

The random assignment system would be developed to be seamless—perhaps even integrated with the existing TAIS system—so that inquirers did not experience an interruption in service while they were on the telephone with intake staff and so that staff could perform random assignment in seconds. The exact details of this system would have to be worked out with OEA.

[12] As discussed below in the Baseline Data section, the data collected prior to random assignment would be limited to those items required to ensure the validity of the study. The evaluator could also pilot test the random assignment process to ensure that callers were not deterred by the required data collection.

Exhibit II.2. Study Intake for Evaluation of Web Referral/Telephone Follow-Up Choice Model


All callers
Collect baseline information
Have an emergency?
Yes
Exempt from evaluation, receive BA assistance
No
Randomly assigned
Web referral/ telephone follow-up choice[a]
Telephone services as usual

[a] Inquirers would choose whether they preferred to use the website, in which case they could expect a response within one day, or wait for a telephone follow-up call, which would take longer.

Also, the evaluator would have to work with OEA to determine the best phrasing for the contractor staff to use for those individuals assigned to the web referral/telephone follow-up choice condition to inform them of their options. Staff collecting baseline information and performing random assignment would be trained in study procedures and delivering consistent information to those randomly assigned to the web referral/telephone follow-up choice group about their options, the website address, and information about how to submit a web inquiry (exact information to be determined).

Like the web referral only design, the evaluator would have to work with the program office to develop a method for handling call-backs from those individuals randomly assigned to the web referral/telephone follow-up choice condition, selected the web access option, and then were unable to find the information they sought or could not submit a web inquiry. In addition, for those customers who chose the phone follow-up option, they would need to consider putting a mechanism in place for checking whether inquirers had submitted a web inquiry before conducting the follow-up phone call, as sophisticated inquirers might hedge their bets by choosing the phone follow-up option and accessing the website anyway. The issues of including compliance-related calls and inquirers who reached the program in error would also have to be resolved before proceeding with this evaluation design.

C. Data Collection Needs

Any random assignment design would require the collection of a range of data from inquirers before and after assistance is provided. Specifically, the evaluation would have to collect data on baseline characteristics and contact information of inquirers, and follow-up data on services received and outcomes of interest.

1. Baseline Data

Baseline data would be collected *before* random assignment. BAs currently use TAIS to collect a basic set of baseline data about inquirers, including name, zip code, and telephone number. They also capture information about the nature of the inquiry during the course of providing assistance. TAIS can store data on inquirers' demographics, employers, and plan information, but BAs typically record this information only if it is needed to resolve the inquiry.

For the purposes of an impact evaluation, an expanded set of baseline data would have to be collected from all inquirers. This would include some additional demographic data on the characteristics of inquirers and additional contact information. These data would be used to assess whether random assignment successfully created similar groups, allow the evaluator to contact participants for follow-up data collection, conduct impact analyses on subgroups of interest, and examine nonresponse patterns for outcome data collection.

In terms of demographics, we suggest that age is an important new variable to collect, given potential differences among age groups in the types and extent of benefits-related issues, likely access to and comfort levels with online or community resources, and the ability to advocate for oneself. The evaluation team could also work with OEA to determine whether other demographics are important to capture, either for the impact analysis or to simply learn more about the types of people contacting the program for assistance. Possible data items could include the inquirers' current or former occupation, income, household size, and number of dependents. Collecting additional demographic data must be weighed against the burden on inquirers and the resources needed to collect and record new data items. In developing a final list of data elements, the evaluator would need to work with CEO and OEA to carefully weigh the analytic value of each item.

In addition to demographic data, we strongly suggest that an evaluation should include collection of additional telephone numbers, email addresses, and potentially even family contact information from inquirers. This additional contact information would be used to ensure that the evaluation team could reach inquirers for a follow-up survey and achieve a reasonable response rate. If the available contact information is limited and resulting response rates are low, the impact results might not be representative of the majority of participants.

Baseline data would have to be collected from all inquirers, including those who did not volunteer to participate in the evaluation under the web referral only design, so that the evaluator could compare the characteristics of volunteers with those who did not volunteer; this would enable the evaluator to determine whether the results of the impact evaluation of the web referral only model could be reasonably inferred to apply to all inquirers. Collecting additional baseline data would also enable the program, under the web referral/telephone follow-up choice model, to see the differences between those who chose to use the website for faster services versus waiting for delayed telephone follow-up. Table II.1 presents some suggested baseline data items for an evaluation along with example wording that could be used to collect this information (the evaluation team would refine the actual wording in conjunction with DOL).

Table II.1. Example Baseline Data Items for Web Referral Impact Evaluation

Data Type	Data Element	Example Wording
Demographics	Age	What is your birth date [or year]?
Contact Information	Telephone number	We might contact you in the future for a follow-up survey and I have to know how to get in touch with you. What is your telephone number?
	Alternate telephone number	Can you give me a different telephone number at which you can be reached, perhaps a cell phone number or your home telephone number?
	Name and telephone number of friend or family member	In case we cannot reach you for a follow-up survey, we would like to have the name and telephone number of one person who does not live with you who will know how to reach you. We would contact this person only if we have trouble getting in touch with you directly.
[For web referral/ phone follow-up choice model] Choice of Web or Telephone Follow-Up	Intent to use the website	Would you prefer to access the website for service within one business day or await a telephone call from a BA? Why? (For example, is that because you prefer faster service, you prefer to use the website in general, or you do not have Internet access?)

2. Follow-Up Data

Follow-up data—to be collected using a follow-up survey of evaluation participants—include data on services received and the outcomes of interest. A follow-up survey would have to be administered to both study groups to capture this information in a consistent way for an evaluation, even though some of the data might be available for a subset of participants in TAIS.

Service receipt data. Data on services received would have to be collected from a follow-up survey of both groups and not from TAIS data because the people who access the website (under both the web referral only and web referral/telephone follow-up choice options) might be able to resolve their issues using information on the website and without submitting a formal inquiry to the program. In that case, the program would have no information about providing direct BA assistance to the inquirer in TAIS, yet the program could have imparted the relevant knowledge to the inquirer through the website. In other words, the website could have provided the services to the inquirer. Questions on the follow-up survey would ask about whether and how the person's inquiry was resolved by using the EBSA website and/or submitting a web inquiry. Although it might seem appealing to use TAIS data for those inquirers for whom it is available and only survey those not represented in TAIS, doing so would compromise the validity of any impacts because the data from the two sources would not be comparable. For instance, survey data are subject to recall bias, or the difficulty people have recalling information after a period of time has passed. The top panel of Table II.2 contains data types and example wording for collecting service receipt data on the follow-up survey.

Outcomes data. The key outcomes for an impact evaluation of the web referral service delivery models would be customer satisfaction, perceived knowledge of benefits rights, perceived ability to advocate on one's own behalf, access to entitled benefits and related documents, and perceptions of a secure retirement and health. OEA already collects measures of some of these outcomes through the Gallup survey and TAIS; however, as with data on service receipt, because the web referral only and web referral/telephone follow-up choice designs imply that some people might use the website for self-service—and therefore not submit a formal inquiry—TAIS data would not represent the outcomes of inquirers in these groups. It would be critical to capture these outcomes for the study. Also, data on outcomes must be collected in the same way across both study groups to be comparable. For instance, monetary recoveries as currently calculated in TAIS include the recoveries for plan participants other than the initial inquirer. It is unlikely that survey respondents would be able to calculate recoveries in a comparable way. The evaluation might be able to use survey responses measuring recoveries per inquirer along with TAIS data on total recoveries per inquiry to estimate upper and lower bounds on the impact of the likely value of monetary recoveries reported through the survey when other affected participants are taken into account.

Sample survey questions to collect outcome data required for an evaluation are provided in the bottom panel of Table II.2. The actual question wording, particularly for estimated values of benefits recovered, would have to be developed in conjunction with OEA to ensure that the question is appropriate and meaningful for all respondents and that they will be able to answer it accurately.

Administering a follow-up survey to both treatment and control group members has several advantages in addition to satisfying the data collection requirements of the evaluation. First, the follow-up survey could include additional outcomes not currently collected anywhere, such as the amount of time it took to receive benefits or benefits-related documents; this could be of interest because one of the impacts of BA assistance might be that people receive benefits sooner than they would have without BA assistance. Also, the survey could be designed to capture specific aspects of participants' knowledge about and access to benefits. For instance, if the participant had a pension-related issue, the survey could ask questions to directly test the participant's knowledge of, for instance, his employer's responsibilities, timing of contributions, and so on. Finally, because the treatment group would also be surveyed, the evaluation team could compare treatment group members' survey responses with their corresponding TAIS records on receipt of benefits and benefit/document recoveries. This information could help OEA understand how to better interpret the data collected in TAIS.

3. Additional Data of Interest

In addition to examining the impacts of the program on inquirers' outcomes, the evaluation could be used to examine impacts of the selected web referral model on program-level outputs. The program already collects a great amount of program-level data, making it straightforward to analyze. Examples of program-level outputs that could be of interest are the number of total inquiries received and closed, website traffic, the number of inquiries submitted through the website as a fraction of the total, and the average time a case is open. The evaluator could work with OEA and CEO to determine which additional program-level outputs were of interest and, if the program were not already collecting them, a data collection strategy. Analyzing how these program-level outputs vary with the implementation of the evaluation would give the program a sense of how it would be likely to change if one of the web referral service delivery models were implemented program-wide.

Table II.2. Follow-Up Survey Data and Example Wording for Web Referral Impact Evaluation

Data Type	Data Element	Example Wording
Service Receipt Data	Leading questions to focus the respondent on benefits-related issues	To refresh your memory, you called the Employee Benefits Security Administration, or EBSA, on [date] to discuss an issue about your benefits, [*for web referral only* and you agreed to be part of a national study]. What type of benefit-related issue were you calling about? [Develop list with OEA]
	Received assistance	Did someone help you resolve your issue, either on the telephone at that time or later?
	If received assistance, from whom?	Who helped you with your issue (for example, family member, community organization, BA)? Indicate all that apply.
	Use of EBSA website	Have you used the EBSA website at any point since your initial call to EBSA on [date]?
	If used website, how?	Did you use the EBSA website to get information about your benefits-related issue? Was this related to the same issue that you called EBSA about on [date]? If not, what was the issue? Were you able to use this information to resolve your issue without submitting a web inquiry? If not, how did you resolve the issue? Did you submit an inquiry for a Benefits Advisor through the EBSA website? Was the inquiry about the same issue that you called EBSA about on [date]? If not, what was the issue? Was your issue resolved? If not, how did you resolve the issue?
Outcome Data	Customer satisfaction	How satisfied are you with EBSA overall?[a] How likely would you be to contact EBSA again using the telephone hotline? Through the EBSA website? How likely would you be to recommend EBSA to a friend?[a]
	Ability to advocate on own behalf	How confident are you that, when issues related to your benefits arise in the future, you can solve them without seeking assistance?
	Access to entitled benefits	Since contacting EBSA, did you receive a document from your employer or benefit plan sponsor related to your benefits? What type of document was it? When did you receive that document? Since contacting EBSA, did you receive benefits that you had previously been improperly denied? When did you receive those benefits? Thinking about all of the benefits that had previously been improperly denied but you have received since contacting EBSA, what would you estimate is their monetary value? For instance, a certain amount per month or per pay period?
	Perceptions of a more secure retirement and health benefits	How confident are you that you (and your spouse) will have enough money to live comfortably throughout your retirement years? How confident are you that you will have enough money to take care of your medical expenses?
	Knowledge of benefits	After your interaction with EBSA, did you feel much more knowledgeable about your benefit rights, somewhat more knowledgeable about your benefit rights, or not any more knowledgeable about your benefit rights?[a]

[a] Currently asked on Gallup survey.

Another potential outcome of interest to the evaluation is BA satisfaction; this could be measured before the evaluation was implemented and again during or after the evaluation to determine how BAs would be likely to respond if the models were implemented more broadly. This could be collected using a separate survey effort as part of an implementation study, described later in this chapter.

4. Using the Gallup Survey or a New Survey for Data Collection

As discussed earlier, data on service receipt and inquirers' outcomes would have to be collected using a follow-up survey to ensure consistent measures were available for everyone randomly assigned, including those who do not have TAIS records. This survey could either be a modification of the existing Gallup survey or a new survey effort.[13] In general, considerable changes would have to be made to the Gallup contract to add resources and ensure the availability of adequate survey staff during the relatively short follow-up period proposed for the evaluation designs. In addition, the survey methodology would differ enough that the results of a modified Gallup survey fielded instead of the usual Gallup survey would not be comparable to data that Gallup has historically provided to the program on outcomes such as customer satisfaction. Given these considerations, EBSA and CEO might consider whether it is feasible and worth the investment of resources to continue running one Gallup survey in its existing form, just as the program always does, and having a separate survey effort for evaluation purposes.

The main considerations in whether the Gallup survey could be used to capture data for an impact evaluation are (1) the ability to add or modify questions, (2) the sampling strategy, (3) the timing of the survey, (4) the ability to achieve reasonable response rates, and (5) the desirability of comparing evaluation responses with historical data from Gallup.

Question items. Many items would have to be added to the Gallup survey or modified for the purposes of an impact evaluation (Table II.2 indicates those items currently collected on the Gallup survey). These include baseline data items, questions on services received, ability to advocate on one's own behalf,[14] whether the person received a benefits-related document or recovered benefits, and questions designed to elicit estimates from the participants of the value of those benefits. Questions would also have to be added about website use and whether the inquirer was able to resolve the issue using information on the website or by submitting a web inquiry.

The Gallup survey in its current form takes an average of eight minutes to complete, and adding questions relevant for an impact evaluation would substantially increase the response time. However, the evaluation team could work with OEA to decide whether other questions currently asked on the survey might be lower priority and could be removed to help compensate for the new questions.

[13] If DOL decides to maintain separate survey efforts for Gallup and the evaluation, the two efforts would not use the same sample frame. Therefore, few, if any, customers would be asked to respond to both surveys.

[14] The existing Gallup survey asks respondents if they feel "better informed to protect my benefit rights." This suggests the ability to recognize a problem, but not necessarily to seek a resolution independently. We suggest that revised questions would have to be added to fully capture the concept of self-sufficiency or ability to advocate on one's own behalf.

Sampling strategy. Only participants whose inquiries are closed in TAIS are currently included in the sampling frame for the Gallup survey; therefore, the sampling frame would have to be adjusted for the purposes of an evaluation to include inquirers who had open inquiries at the time the follow-up survey was administered so that outcomes for all evaluation participants are collected after a similar amount of time has elapsed, even if their inquiries have not been fully resolved. The sampling frame would also have to be adjusted to include members of the web referral group. This is because inquirers in this group might use the EBSA website for self-service, resolve their inquiries, and never contact the program again; therefore, TAIS would not capture them even though they received services through the website. The program or the evaluation team could use the random assignment database to generate lists of the contact information for people in both evaluation groups.

Timing of the survey. Currently, OEA releases contact information for participants within two weeks after their cases are closed and Gallup attempts to survey them within one month. The time frame for collecting service receipt and outcomes data for the impact evaluation should be based on expectations of when most issues would likely be resolved and an impact of services could be detected. An evaluator would have to work with OEA to determine the most appropriate length of the follow-up period, but it is likely to fall somewhere within two to six months from the day of random assignment. That time frame would provide time for people to receive COBRA notices—which generate many calls—and give people an opportunity to try to advocate on their own behalf, if needed. Therefore, the timing of the Gallup survey would have to be adjusted to contact people within a specific period after random assignment. In addition, evaluation intake could occur very quickly—perhaps over the course of only a couple of months—with the follow-up survey having to be conducted over a similar amount of time. Thus, the process of submitting information to Gallup on a rolling basis over the entire year for survey sampling would have to be adjusted.

Survey response rates. In FY 2011, the response rate for the Gallup survey was approximately 54 percent. In a typical impact evaluation, an 80 percent response rate is considered appropriate. This is because a response rate of about 80 percent overall is often considered to imply that the program impacts are being estimated on a representative sample of the study population. Although the survey sample size for the evaluation would have to be finalized during development of the evaluation design, the completed survey sample sizes proposed in this report are quite a bit larger than the annual number of completes Gallup typically achieves under its existing contract with EBSA.[15] This means that the total number of complete surveys would have to increase dramatically. Because of the relatively low current response rate on the Gallup survey, and because we suggest a longer time frame for follow-up, we recommend collecting additional contact information for study participants, as discussed earlier. The evaluator could also consider offering a small monetary incentive to encourage respondents to complete the survey. CEO and EBSA would also need to consider the investment of resources required to achieve this response rate and ensure that the evaluator has designed a thorough non-response analysis in the event that an ultimate response rate of 80 percent cannot be achieved.

[15] Gallup typically samples about 11,000 participants per year and achieves an average response rate of 54 percent, for a total of about 6,000 complete surveys.

Desirability of comparing evaluation responses with historical data from Gallup. Because the sampling frame, timing, and response rates for an evaluation follow-up survey would differ from Gallup's usual survey administration processes, the information gleaned from the evaluation follow-up survey would systematically differ from what Gallup typically collects, even for identical questions. Therefore, if the Gallup survey were adapted to meet the needs of an evaluation, the customer service performance data generated for the program would not be comparable to historical data. This might make it difficult for the program to assess its ongoing versus past performance.

D. Evaluation Sample Sizes and Minimum Detectable Impacts

An MDI is the smallest true impact that an impact evaluation has a high probability of detecting. It is largely a function of the impact evaluation sample size—the total number of people enrolled in the evaluation. Typically, a target MDI is selected during the design phase of an impact evaluation, and then the evaluation sample sizes needed to achieve that MDI are set; this means that the evaluators know in advance how many people would have to enter the evaluation, which has implications for the length of the intake period and fielding the follow-up survey. Thus, before embarking on an evaluation of the relative impact of a web referral service delivery strategy, EBSA, CEO, and the evaluator would have to select a target MDI.

As discussed earlier, the web referral service delivery strategies considered in this chapter have the potential to enable BAs to serve more customers at a lower cost per customer, along with ancillary benefits such as increased BA job satisfaction. If the web referral model also resulted in similar or better outcomes for inquirers, then the program might consider adopting the model more broadly. Therefore, the target MDIs should be set with the goal of determining whether the two service delivery strategies result in similar outcomes for customers. This can be done by setting a threshold for the impact to fall within.

For example, if EBSA and CEO determined that it would be acceptable for there to be a 3 percentage point or lower difference in customer satisfaction or knowledge of benefits rights between the two service delivery approaches, then the study could be powered to detect this. If the study did not find a statistically significant impact, then it could be concluded that the web referral strategy was roughly equivalent—in terms of inquirers' outcomes—as delivering telephone services as usual because the impact was no larger than the specified threshold. Thus, EBSA, CEO, and the evaluator would have to work together to determine the largest difference they would be comfortable with for each of the outcomes of interest, and power the evaluation appropriately. Looking at historical data on customer satisfaction and perceived knowledge of benefits rights might provide some insight into acceptable thresholds for these measures.

To quantify the tradeoffs between sample sizes and MDIs, we computed them using two key outcomes of interest—binary variables such as customer satisfaction, perceived knowledge of benefits rights, and ability to advocate on one's own behalf; and monetary recoveries—based on three different sample sizes: 25,000, 15,000, and 10,000 inquirers. Note that these sample sizes correspond to the number of participants in an evaluation, and therefore the number the evaluator would attempt to contact through a follow-up survey. The number of completed surveys determines the statistical power, not the number of attempted surveys. When deciding which sample sizes to consider, we kept in mind that the current Gallup survey targets slightly more than 11,000 survey respondents per year (although only about 6,000 actually complete the survey), and that CEO could potentially provide resources for additional data collection efforts for an impact evaluation.

A full set of assumptions underlying the MDI calculations is contained in Appendix C. The key assumptions follow:

- **Power.** We assumed a power of 80 percent. This means the evaluation would detect true impacts with a probability of 80 percent.
- **Statistical significance.** We computed hypothesis tests assuming a 0.05 significance level for a two-tailed test.
- **Random assignment ratio.** We assumed that half the study sample would be randomly assigned to the web referral model and the other half to the services as usual group. This is the most efficient ratio possible and would minimize the total study sample size needed.
- **Key outcomes of interest.** We focused on two types of outcomes. The first type is binary outcomes, which include customer satisfaction, perceived knowledge of benefits rights, ability to advocate on one's own behalf, and secure retirement and health (the discussion generalizes to other binary variables as well). Binary variables take the value of 1 if, for instance, the respondent feels much more or somewhat more knowledgeable about his or her benefits rights, and the value of 0 otherwise. For exposition purposes, we use the perceived knowledge of benefits rights as an example throughout this section. Binary outcomes with similar mean values as the knowledge of benefits measure would have similar MDIs because of the statistical properties of binary variables.[16]

  The second outcome is monetary recoveries, which would reflect the dollar amount of benefits recovered for both groups. It is likely that resulting MDIs would differ somewhat if other types of outcomes were used.
- **Mean and standard deviation of perceived knowledge of benefits rights.** OEA provided us with information on the mean value of knowledge of benefits rights from the Gallup survey.[17] More than three-fourths (76 percent) of respondents to the 2012 survey indicated that their knowledge level was much more or somewhat more after interacting with the program. However, the standard deviation of the group receiving the web referrals is needed for computing MDIs. Although we cannot be sure what amount of knowledge about their benefits the inquirers in the web referral groups are likely to acquire on their own, we assumed that 65 percent would somewhat or strongly agree that they felt better about these outcomes, on average, reflecting a

[16] OEA currently uses a five point scale for customer satisfaction. For each of exposition, we suggest converting this measure into a binary measure, dividing the scale in ways that are most meaningful to OEA. For example, if they want to increase the proportion of customers responding with a 1 or 2 on the current scale, the outcome measure could be 0 for those responding 3,4,5 on the current scale and 1 for those responding 1 or 2. Alternatively, DOL might consider examining impacts on the mean customer satisfaction score.

[17] We discussed with CEO the possibility of capturing actual gains in inquirers knowledge of rights and determined that the cost, complexity, and difficulty of measuring such an outcome made them infeasible to consider.

treatment impact of 11 percentage points or roughly 15 percent. The standard deviation is derived from that mean as the square root of 0.65*(1 - 0.65) = 0.48. Because of the statistical properties of binary variables, the MDIs are not sensitive to small differences in the assumed standard deviations. For instance, shifting to an assumption that the web referral group mean is 50 percent results in a 0.01 percentage point difference in the MDI for a sample size of 25,000. Even shifting to an assumption of a web referral group mean of 20 percent or 80 percent results in only a 0.03 percentage point difference in the MDI.

- **Mean and standard deviation of monetary recoveries.** OEA also provided us with data on monetary recoveries from TAIS for the first three quarters of FY 2013.[18] The program currently computes monetary recoveries as a total that includes the monetary recoveries of other participants at the same employer/plan sponsor who were also affected as a result of BAs' informal interventions. However, we do not expect to be able to gather this information in a follow-up survey; respondents will probably not be able to accurately estimate the amount of benefits recovered for others in their plan or the number of other participants assisted, and collecting this information directly from employers or plan sponsors would require considerable resources for the impact evaluation. In addition, the program uses a complex calculation to determine these amounts, and inquirers are unlikely to be able to report estimates in the same way.[19] Therefore, this measure will differ from what the program currently collects, and the analysis would have to explain that this is likely an underestimate of the true impact on this outcome. However, the evaluation could use survey responses measuring recoveries per inquirer along with TAIS data on total recoveries per inquiry for the telephone services as usual group to get an estimate of the likely value of monetary recoveries reported through the survey when these issues are taken into account.

  To obtain the monetary recoveries per participant, we divided the total amount of benefits recovered by the number of participants assisted to arrive at a recovery amount per participant. The vast majority of inquiries do not result in a document or benefit recovery, so we accounted for this when computing means and standard deviations. Once again, it is difficult to know what to expect about the dollar amount of benefit recoveries or its likely standard deviation for the web referral group. Presumably, both would be about the same as for the telephone services group, so we used the standard deviation estimated from TAIS.

[18] Monetary recoveries are not a cumulative measure; rather, they are calculated at a single point in time. Depending on the type of recovery, the monetary recoveries computed by the BAs might factor in not only current benefits recovered but also future benefits recovered. However, this is all included in the calculation. Therefore, it is not necessary to adjust for the fact that these estimates were based on only three quarters of program data.

[19] For example, suppose an inquirer called because he was unable to locate the trustee of his former employer's pension plan. The BA located the trustee and explained its fiduciary responsibilities, as a result of which the plan sponsor decided to terminate the plan. If 10 plan participants received distributions totaling $150,000, that full amount is entered as a monetary recovery covering 10 participants. In this scenario, it is unlikely that survey respondents would know about the others in their plans who were affected and be able to report on the amount of distributions they received. SOP 3-12 contains further information on how monetary recoveries are calculated.

- **Survey response rate.** We assumed a survey response rate of 80 percent, which is a common and achievable target for impact evaluation designs. The ability to achieve this rate would depend on the amount and quality of contact information collected during the intake period and the length of the follow-up period.
- **Subgroups.** We computed MDIs for three potential subgroup sizes: a 50 percent subgroup of inquirers in the study sample—for example, those with a health benefit issue versus a pension benefit issue; a 25 percent subgroup of inquirers, such as those of retirement age; and a subgroup of 10 percent of inquirers, such as those ages 70 or older. Although we do not know the exact proportions these subgroups of interest might represent in the study, these calculations give some indication of the range of impacts we would be able to detect when looking only at inquirers in those types of subgroups.

Table II.3 presents the MDIs for the web referral only and web referral/telephone follow-up choice options. As shown in column 1 of the table, with an evaluation sample of 25,000 participants (with the assumed 80 percent response rate, this would result in 20,000 completed follow-up surveys), the study could detect an impact as small as 1.9 percentage points in perceived knowledge of the web referral group compared with that of the telephone services as usual group. This means that if the true difference between web referral group members' perceived knowledge of benefits rights and that of telephone services group members were 1.9 or more percentage points, the study would be able to detect that difference; if the true difference were less than that, the study would not find a statistically significant impact and would conclude that the web referral and telephone services as usual affect inquirers' outcomes in roughly similar ways. Similarly, for monetary recoveries, an evaluation sample of 25,000 inquirers would mean the evaluation could expect to detect an impact of $436 in recoveries per participant. For reference, the average monetary recovery per person is $814, based on data from FY 2013.

Table II.3. MDIs for Web Referral Only and Web Referral/Telephone Follow-Up Choice Options

	(1)	(2)	(3)
Sample Size	25,000 Randomly Assigned, 20,000 Complete	15,000 Randomly Assigned, 12,000 Complete	10,000 Randomly Assigned, 8,000 Complete
MDIs—Binary Variables			
Overall	**1.89%**	**2.44%**	**2.99%**
50% subgroup	2.67%	3.45%	4.22%
25% subgroup	3.78%	4.88%	5.97%
10% subgroup	5.97%	7.71%	9.44%
MDIs—Recoveries			
Overall	**$436**	**$562**	**$689**
50% subgroup	$616	$795	$974
25% subgroup	$871	$1,125	$1,377
10% subgroup	$1,377	$1,778	$2,178

Note: Binary variables include customer satisfaction and self-reported knowledge of benefits rights, ability to advocate on one's behalf, access to benefits-related documents, and perceptions of a secure retirement and health. MDIs for binary variables are expressed in percentage points. MDIs for recoveries are expressed in dollars. See Appendix C for a description of the full set of assumptions used to calculate the MDIs.

As a more concrete—and completely hypothetical—example, suppose that EBSA and CEO determined that, in pursuing the web referral service delivery strategy, the potential benefits to the program in terms of reduced BA time spent on erroneous and/or simple informational calls and data entry would be worth the tradeoff of a 2.5 percentage point impact (either positive or negative) on inquirers' perceived knowledge of their benefits rights, self-sufficiency, and other binary outcomes. Then selecting an evaluation sample size of about 15,000 inquirers would be appropriate because the study would be powered to detect a difference of 2.44 percentage points in either direction in binary variables. Thus, if no statistically significant differences were found, it would indicate that the difference between the two strategies was within the acceptable range of 2.5 percentage points set by EBSA and CEO; it could be concluded that the web referral service delivery strategy being tested was roughly as successful in terms of inquirers' outcomes. If instead EBSA and CEO felt that a 2.0 percentage point impact was the maximum acceptable impact on inquirers' perceived knowledge of benefits rights, a sample size of about 25,000 inquirers would be needed, whereas a 3.0 percentage point impact would require a sample size of about 10,000 inquirers.

The evaluator would have to work closely with EBSA, CEO, and other potential stakeholders to determine the target MDIs for an evaluation of a web referral service delivery model. Based on that decision, the sample sizes necessary to achieve the target MDIs would be determined based on calculations similar to those presented here.

Given a call volume of about 250,000 inquirers per year, it might not take very long to achieve the sample targets of 10,000, 15,000, or 25,000 inquirers, particularly for the web referral/telephone follow-up choice option in which everyone with a nonemergency call would be randomly assigned. However, EBSA might be interested in conducting the evaluation for a longer period than this, such as six months or a year, to become familiar with the different way of operating and tweak the service delivery model if needed. If this were desired, all inquirers could enter the evaluation in the intake process described here, but only a random subsample of them would be contacted for the follow-up survey.

Another approach to extending the length of the evaluation while still targeting sample sizes in the range of those discussed here could be to capture only every fifth or tenth call for inclusion in the evaluation. However, one important consideration is that evaluation participants could more easily cross over into the telephone services as usual condition if they called the hotline again after random assignment. For instance, it would be easy for inquirers in an evaluation of the web referral/telephone follow-up choice option to call the program back and go straight through to receive immediate BA telephone assistance if fewer than 100 percent of calls were screened for inclusion in the evaluation. The program could consider having BAs do a database check to see if new callers had been previously randomly assigned. However, this would likely be too burdensome to be practical. The evaluation could also be limited to a subset of regional offices. However, this poses the same problem in that it would be easy for evaluation participants to cross over and receive immediate BA assistance by calling another office that was not participating in the evaluation. In addition, conducting the evaluation in only a few offices would limit the generalizability of its results.

E. Strengths and Drawbacks of the Web Referral Evaluation Design Options

The web referral impact evaluation options have several strengths overall and relative to each other. In particular, both designs would provide EBSA and CEO with actionable information on which to base future operational plans given that both options reflect service delivery strategies the program has considered implementing, especially in times of high call volume. If the evaluation found that the two service delivery models resulted in similar outcomes for inquirers, it would provide OEA with strong and specific evidence about how the program could change its current service delivery model. For example, if the web referral model can be delivered at a lower cost per participant than the current model or results in improved job satisfaction of BAs, OEA could choose to implement it more broadly or permanently. If the evaluation found that the web referral model improved outcomes for inquirers relative to the current model, this would further bolster the case for expansion of the model. If the program was more effective for a certain subgroup of inquirers, the program might also learn how to target web referrals most effectively. However, these designs also have some potential drawbacks; we discuss both strengths and drawbacks here.

Strengths and drawbacks of web referral only evaluation design. One of the main strengths of this design is that no one would be denied services; everyone would receive services, either by telephone or via the website. This reduces the ethical concerns associated with an evaluation design in which some inquirers would be denied BA services. On a related note, because only the inquirers who indicated they had access to the Internet would be part of the evaluation, there would be minimal chance of referring people to the website who could not then access BA services through it. And, only people who consented to participate in the evaluation—and therefore already knew that they would have a chance of being assigned to the web referral only group—would be referred to the web in the first place.

In addition, collecting baseline data on demographic characteristics for all customers, as proposed here, would provide OEA with information that could help focus its outreach activities and inform future service delivery options. For example, it would provide information on the types of customers who call for services in the first place; characteristics of those who have Internet access and those who do not; characteristics of those who have emergency situations and the nature of those situations, which could be compared with the program's definition of emergency situations; what customers think their topic of inquiry is and how well that relates to what the topic actually is, as coded by a BA in TAIS; and much more. It would also facilitate an analysis of those who do not volunteer to be part of the evaluation, so that the evaluator can determine to what extent the evaluation's findings can be generalized to the broader population of people seeking BA assistance.

The main appeal to this design—randomly assigning only those who have Internet access and volunteer to participate in the evaluation—is also its main drawback because it is unclear what fraction of those asked to participate in the evaluation actually would volunteer to do so. If the inquirer could simply refuse to participate in the evaluation and receive immediate services, it seems likely that many—if not most—inquirers would do that. Even inquirers who might be receptive to participating in an evaluation in general might find it easier to not volunteer to participate in the study rather than disconnect the call, access the website, and potentially have to call back again if they are unable to resolve their problems or submit web inquiries. To better inform the likely outcome of the encouragement design needed for this impact evaluation option, OEA and/or the evaluator could undertake a pilot study to see the extent to which obtaining volunteers might be an issue, and perhaps consider offering some kind of incentive to participate. Alternatively, given that consent would not be required for this type of study, EBSA and CEO might consider implementing random assignment without asking for volunteers, as outlined in the web referral/phone follow-up choice option.

Another drawback to this design is that if the analysis of inquirers who did not volunteer for the evaluation found that they were very different from those who did volunteer—for instance, they were much older on average, or had different types of inquiries—then the evaluation's findings would not be generalizable to the population of inquirers, greatly limiting its usefulness. In short, if those who volunteered differed from those who did not, then implementing the web referral only model more broadly would be less likely to result in the same impacts as were found through the evaluation.

Finally, the evaluation could have an impact on customer satisfaction; this is potentially problematic because OEA currently uses customer satisfaction as its primary performance measure. If customer satisfaction declined under the web referral only model—perhaps because people did not find the information they needed on the website and had to call back—then the program's performance would reflect this. In addition, as mentioned in Section II.C, the customer satisfaction measures on the follow-up survey would not be directly comparable to historical data unless a separate data collection effort was undertaken for the evaluation, further complicating performance reporting. Thus, EBSA, CEO, and the evaluation team would have to work together to adjust expectations about the customer satisfaction standard, perhaps encouraging DOL leadership to temporarily adjust the goals upon which performance is gauged, implement the usual Gallup customer satisfaction survey as a separate survey effort, or remove the measure temporarily as a performance indicator for the program.

The strengths and drawbacks of both options are summarized in Table II.4.

Table II.4. Strengths and Drawbacks of the Web Referral Evaluation Design Options

Web Referral Only	Web Referral/Telephone Follow-Up Choice
Strengths • Provides EBSA and CEO with actionable information on which to base future operational plans • No service denial: everyone gets services via the web or telephone • Minimal risk of referring people without Internet access to the web for services • Collection of baseline data would provide information on various items of interest to the program and on those who do not volunteer to participate in the evaluation	**Strengths** All the strengths of the web referral only option plus: • Little concern about achieving target sample sizes because volunteering is not required • Evaluation participants will be representative of the population of inquirers because volunteering is not required • Learn about why people choose to access web versus wait for telephone follow-up • Provides an incentive of faster service to use the website, unlike web referral only option, so people might be more likely to do this
Drawbacks • Might be difficult to get enough volunteers to participate in the evaluation • If analysis of those who did not volunteer to participate found they differed from those who volunteered, study results could not be generalized to the population of inquirers, limiting the study's usefulness • Customer satisfaction—a performance measure—might be affected; it would not be comparable to historical customer service data	**Drawbacks** • Might not reduce BAs' workload as much as web referral only model if many inquirers choose to wait for a telephone follow-up • Same issue with customer satisfaction performance measures as web referral only model

Strengths and drawbacks of web referral/telephone follow-up choice evaluation design. The web referral/telephone follow-up choice design has the same key strengths as the web referral only model: the program would learn actionable information on which to base future operational plans, no one would be denied services, no one would be forced to use the Internet for services because they could choose to await a follow-up telephone call (whereas they would choose to not consent in the web referral only model), and the rich baseline data collected would enable the program to learn about various characteristics of its customer base.

In addition to these strengths, this design also has some key advantages over the web referral only option. First, because all calls (minus emergencies) would be included in this design option and inquirers would not have to volunteer to participate, there is much less concern about achieving target sample sizes. Also because of this design feature, the results will be generalizable to the population of people who contact the program for assistance, and not only to the subgroup that volunteers to be included in a research study in which web referral is a possibility. This means EBSA would have high confidence that adopting this service delivery model would result in the same types of impacts as those estimated in the evaluation.

Also, because inquirers under this option would choose whether they prefer to access services via the website or wait for telephone follow-up, this design option would provide information about characteristics of the inquirers who make these choices. This could be further enhanced by gathering information about why the inquirer chose the particular option he or she did, either on the follow-up survey or at the time the choice was made. EBSA could use this information in the future when considering strategies for encouraging web usage, especially during periods of high call volume.

Finally, from the inquirer’s perspective, having access to faster service via the website compared with waiting for telephone follow-up provides an incentive to use the website. This could result in more people choosing to access the website than would choose to consent to the web referral only evaluation, potentially shortening the study intake period relative to that option.

The main drawback to this design relative to the web referral only option is that it might not reduce the BAs’ workload as much as the web referral only option. If any of the people who do have Internet access choose the telephone follow-up option, this could result in the BAs handling more inquiries directly than if all inquirers had to use the website, for the same sample size. In addition, it might require more time to contact people for follow-up than it would to respond to their inquiries immediately, particularly if inquirers are not available or do not answer their telephones when a BA calls. However, this would still result in valuable information for the program—namely, why people with Internet access would rather wait a few days for a telephone call than attempt self-service themselves. It also has the drawback of potentially altering customer service performance measures, which would have to be addressed under this evaluation design as well.

F. Implementation and Cost Studies

For an evaluation comparing two service delivery strategies, it would be critical for the evaluator to conduct an in-depth implementation study to further understand how web service delivery unfolded over the course of the evaluation. This information would help the evaluator interpret the impact results and help identify best practices that could be applied more broadly if the new model were adopted program-wide. For instance, BAs might develop strategies for smoothing out their workloads by responding to web inquiries in a certain order. This information could be documented and shared with other BAs. Or, the contractor staff and evaluation team might develop effective ways of tactfully directing inquirers to the website during times of high call volume as they gain familiarity with directing people as part of the evaluation; this too could be used if one of the web referral options were implemented program-wide.

A cost-benefit study would also be critical for comparing the resources needed to implement each of the service delivery strategies. The relative costs of the approaches being compared would provide EBSA with information necessary to determine whether to adopt the new strategy more broadly or retain the existing one. For instance, if the evaluation was powered to detect a 3 percentage point impact on knowledge of benefits rights but the analysis showed no statistically significant differences on this outcome between the two service delivery models, it would suggest that OEA might want to adopt the less-expensive approach, barring other important implementation factors. This would result in a savings to the program and a better return on taxpayers’ investment in the program.

Cost studies can sometimes be done fairly easily by taking the total budget and dividing by the number of participants assisted to get a cost per participant. However, when comparing two service delivery approaches, there are likely to be subtle variations in costs across the two models that would have to be captured and would be difficult to parse out using this approach. Instead, costs would have to be generated by collecting detailed information on how staff time and other resources are allocated for each approach; this would be done as part of an implementation study and could be structured to minimize burden on BAs.

III. CONSIDERATIONS FOR A CLASSIC RCT TO ASSESS PROGRAM IMPACTS

Chapter II discussed options for an evaluation designed to determine the relative impact of an alternative BA service delivery model on participants' outcomes. It was hypothesized that the web referral models might be less costly, enable the program to serve more inquirers, and increase BAs' job satisfaction; if participants' outcomes were roughly the same under this service delivery model as under traditional telephone service delivery, it could justify expanding the use of web referrals in the program's normal activities.

In this chapter, we shift from the relative impact evaluation design to an evaluation design that would determine the overall impact of BA assistance. Known as a classic RCT, this impact evaluation would randomly assign inquirers either to a group that could receive BA services as usual or a group that would not be able to access BA services. By comparing the outcomes of these two groups, the study would be able to determine what would happen to participants' outcomes in the absence of BA services.

After reviewing the design considerations for a classic RCT in the draft version of this report, CEO determined that it will not pursue this design option at this time because of ethical and implementation concerns raised by EBSA and TWG members. Although EBSA and the TWG members acknowledged that, in theory, a classic RCT would be the most rigorous way to conduct an impact study of the overall impact of the program, they felt that denial of services to a subset of inquirers would compromise EBSA's core mission and raise ethical concerns. Even in periods of extremely high call volume, the program has managed to serve all inquirers with some level of service, and they felt that refusing to provide timely assistance could result in participants missing important eligibility deadlines and appeal time frames, not obtaining necessary documents needed to obtain benefits guaranteed by law, and losing retirement and/or health benefits, or necessary medical treatments. In addition, the two TWG members with substantial evaluation design expertise felt that not enough information is known about the program, who it serves, and the quality of services delivered to warrant a classic RCT at this time; these TWG members recommended pursuing one or more studies to first shed light on these issues and then consider implementing an RCT at a later point to determine the overall impact of the program.

This chapter provides important information on the key issues that would have to be considered if a classic RCT were to be implemented in the future. The chapter begins by describing the research question and hypothesized links between program activities and outcomes that could be tested with this design. We then describe how the evaluation could be implemented (if it were to be pursued at a later time), discuss data collection needs, explore sample sizes and MDIs, and review the potential use of implementation and cost studies to support an impact evaluation. We end with a discussion of challenges for this option and potential solutions.

A. Research Question and Hypothesis for a Classic RCT

When considering the impact of a program, one of the most basic questions that could be asked is how program participants are affected by the program's services, compared with how they would have fared if they had not had access to the services. The following is the relevant research question:

- What is the impact of having access to direct assistance provided by BAs on participants' perceived knowledge of and access to their entitled pension and health benefits compared to not having access to the full range of BA participant assistance activities?[20]

Before considering a study to answer this research question, it is important to understand the expected relationships between program activities and outcomes. That is, how do we expect the direct participant assistance provided by BAs to affect participants' knowledge of and access to benefits? Based on the logic model created for this design study, we hypothesize that BAs providing information and/or informal intervention to inquirers with benefits-related questions and issues should lead to increased knowledge about their benefits issues and recovery of benefits-related documents and entitled benefits. Ultimately, these short-term outcomes are expected to lead to increased knowledge about benefits rights, increased ability to advocate on one's own behalf, increased access to benefits-related documents and entitled benefits, and greater perceptions of a secure retirement and health in the long run (see Exhibit I.2). By implementing a study that randomly assigns inquirers to receive or not receive BA assistance, DOL would be able to test the hypothesized link between BAs' activities and participants' outcomes. Positive impacts could potentially justify further expansion of the program.

In addition to estimating overall impacts of the program, this type of evaluation could also estimate impacts for different subgroups—by type of inquiry, age of the participant, or other characteristics of interest—to determine whether impacts were larger or smaller for certain groups. This information could help OEA consider enhancements or modifications to its current service delivery strategy, including outreach and education efforts. For example, if an evaluation determined that the program had greater impacts for those with pension-related questions, OEA might consider enhancing online resources related to pension issues, placing greater focus on education efforts related to pension issues, or targeting populations during outreach efforts who are likely to be in need of pension advice and assistance. Similarly, if an evaluation indicated that BAs' direct participant assistance had a greater impact on older inquirers, OEA might consider targeting outreach activities to increase awareness of the program among this population. Thus, the same evaluation design could estimate both overall impacts and impacts for subgroups of interest.

B. Implementation of a Classic RCT

For the duration of a classic RCT evaluation, inquirers would contact the program in the usual manner. When they were connected to a regional office, they would be informed about the evaluation, asked to consent to participate in it, and—if they provided that consent—randomly assigned to the treatment or control group. Those in the treatment group would receive BA services as usual while those assigned to the control group would not be able to receive the full range of available BA services. At some point after the initial telephone call to the program—to be determined by EBSA, CEO and the evaluation team—data would be collected from both

[20] Because inquirers assigned to the control group in this design would not receive any BA services, it would be impossible to measure their satisfaction with services they did not receive. Thus, this design could not estimate an impact on customer satisfaction.

groups to measure their benefits-related outcomes of interest; the outcomes of the two groups would be compared to determine the impact of the assistance BAs provided.

1. To What Would the BAs' Participant Assistance Services Be Compared?

Comparing the outcomes of those who were randomly assigned to receive BA services with those who were randomly assigned to not receive services would provide a strong test of what would happen in the absence of services being offered, known as the counterfactual. In this case, those assigned to the control group would not be offered assistance from BAs upon learning of their assignment; however, they would be able to access whatever other resources were available in their communities, from family members, their employers, or other sources. Control group members could also access the EBSA website for self-service if they found it on their own.

This type of evaluation could also be designed with two slight variations on the type of assistance offered to the control group, which would enable all callers to receive some guidance from EBSA. In one variation, immediately after being informed of their assignment, control group members could be directed to the EBSA website and told that they could access self-service materials on benefits-related issues. This so-called light touch treatment approach would not allow control group members to receive one-on-one assistance from BAs, but they might be more likely to access the website for assistance. This could be a reasonable way to offer some basic services to control group members with minimal potential for influencing the impact estimates. However, this variation would change the counterfactual in the study, because the estimated program impacts would be only the impacts of the telephone BA assistance above and beyond referral to the website.

In a second variation, control group members could be directed to the EBSA website and told that they could access self-service materials on benefits-related issues *and* submit a web inquiry if they needed additional assistance from a BA. In this case, the counterfactual is quite different, as individuals would be able to receive direct one-on-one assistance if they chose to submit a web inquiry. Such a design would answer a different research question: What is the impact of receiving BA assistance through the telephone on participants' knowledge of and access to their entitled pension and health benefits compared with receiving BA assistance only through the website? This study design is similar to the two evaluation designs, discussed in detail in Chapter II, that assess the relative impacts of web-referral service delivery strategies.

2. How Would Random Assignment Work?

The process for enrolling inquirers in the evaluation and randomizing them to the treatment or control group would involve several steps. First, inquirers would have to be screened and determined eligible for random assignment. As discussed in Chapter I, we suggest that those inquirers who reach the program in error or require resource assistance resulting in a simple referral to another entity would not be enrolled in the study. This would be determined by the contractor staff answering the hotline or regional telephone number or responding to the web inquiry.[21] In addition, as with the evaluation designs discussed in Chapter II, those with

[21] It is possible that BAs could conduct study intake and random assignment. However, this could be problematic given the burden of this additional effort on top of staff's existing workloads, and EBSA, CEO and the

emergency situations would be excluded from the evaluation and directed to a BA for immediate assistance. Second, the contractor would inform eligible inquirers about the study and seek their consent to participate. During this process, which could take as few as three minutes, the contractor would tell inquirers the purpose of the study; that their data would be shared with an external evaluator; that confidentially was assured; and the details of the random assignment process, including the potential to be randomly assigned to either a treatment or control group. The contractor would also advise inquirers that refusing to consent to participate in the study would mean they would be unable to receive any services from the BAs.[22] Then, the contractor would seek verbal consent from the inquirer to be in the study. If the inquirer consented, the contractor would collect a limited number of baseline data items, discussed in detail below. Finally, the contractor would submit the individual's data to a random assignment system (to be either developed by the evaluator or integrated into TAIS by OEA), receive the results, and inform the individual of his or her assignment, passing those in the treatment group to the BA hunt group as usual.

There are several important considerations when implementing a classic RCT. First, it would be important to direct all inquiries—from the toll-free hotline, the regional office numbers, and web inquiries—through the study procedures to ensure that everyone was included in the evaluation; if everyone was not, it could call into question the validity of the impact estimates. In the case of web inquiries, this would occur during an initial telephone contact after submission of the inquiry.

Second, ensuring adherence to the assigned research group status would be essential for maintaining the integrity of the study. If inquirers in the control group are able to access BA assistance, the distinction between the treatment and control groups becomes less meaningful. Comparisons between the groups would not be valid, and the analysis would underestimate the true impact of BA assistance. This type of control group crossover into treatment service receipt could be mitigated by implementing procedures to automatically check the random assignment database for duplicate records when a case is submitted for random assignment. This could be done using the inquirer's name and telephone number (or a different combination of variables that the evaluator and DOL determine are appropriate). Those found to have already been assigned to the control group would be informed that they are not eligible for BA assistance. To

(continued)

TWG members expressed a strong preference for using contractor staff for conducting study consent and gathering study-specific data that are not currently collected by BAs. There is the potential for CEO to provide resources for an evaluator to hire contractor staff to perform this function, thus reducing the burden on existing program staff. We therefore refer to contractor staff performing these functions throughout the remainder of this chapter.

[22] We considered the possibility of a design in which individuals are encouraged to participate in the study but are not denied access to services if they chose not to consent. We determined that the lack of incentive to participate would make it extremely difficult to generate enough sample for the study; we expect that, given the choice to not consent and receive services as usual or to consent and have the chance of not being able to access services, very few people would consent to participate in the evaluation. In addition, there could be significant unobservable differences between those who consented to participate and those who did not; we would not have any data to examine this. Therefore, the study results could not be generalized to the full population served by EBSA and would be of limited use. If CEO chose to pursue a classic RCT evaluation design in the future, an institutional review board would have to assess whether the denial of services to those who chose not to consent to participate in the study would be acceptable.

prevent treatment group members from having to go through this step more than once, BAs working with treatment group members after random assignment could provide a direct telephone number to use for follow-up discussions. Although these procedures would minimize the amount of crossover, some might occur. The evaluator would have to determine through the follow-up survey whether control group members were able to receive BA assistance and adjust the impact analysis to account for low levels of crossover.

Third, it would be very important for the evaluator to provide training to the contractor staff who would administer the consent process, conduct random assignment, inform individuals of their assignments, and check study group status. The evaluator would provide detailed procedures manuals with formal scripts, talking points, and other resources to use during the consent and random assignment processes. Training would also have to be provided to BAs to impress upon them the importance of the study and adhering to study procedures. If BAs chose to purposely circumvent the system, such as by providing services to individuals assigned to the control group, the results might suggest that BA assistance had no impact on participants' outcomes even if it did, in fact, have an impact. The evaluation team would also have to provide ongoing technical assistance to address evaluation issues as they arose and to ensure that everyone followed study procedures.

Finally, the evaluation team and OEA would have to determine whether random assignment could be done within TAIS or an external data system would be needed. Most evaluations use an external system because it does not require modifications or disruption to existing program management systems and can be more easily monitored by an external evaluator. If an external system were used, however, it would be important to investigate whether it would be feasible to automatically transfer baseline data and random assignment results to the TAIS database. This would eliminate the need for duplicate data entry in the two systems. If this were not feasible, CEO might consider providing resources to support the salaries of dedicated staff members to perform the required data entry in TAIS.

C. Data Collection Needs

A classic RCT would require the collection of a range of data from inquirers before and after assistance is provided. Specifically, contractor staff would have to collect data on baseline characteristics and contact information of callers, and follow-up data on services received and outcomes of interest. The types of data needed for a classic RCT are generally the same as those discussed in Chapter II, Section C with respect to an evaluation of web referral service delivery models. Here we point out a few additional considerations relevant for a classic RCT.

Baseline data. As mentioned previously, an expanded set of baseline data would have to be collected for the purposes of the impact evaluation, including some additional demographic data on the characteristics of callers. These data would be used to assess whether random assignment successfully created similar groups, conduct impact analyses on subgroups of interest, and examine nonresponse patterns for outcome data collection. Table II.1 in Chapter II presents the baseline data items for an evaluation of web referral service delivery options, along with example wording that could be used to collect this information. The same baseline data elements would be collected in a classic RCT, with the exception of asking whether the study group member would prefer to use the website or wait for telephone follow-up, because that would not be not an option in the classic RCT.

Service receipt data. Data would be collected on the services received by both groups. This would provide information not only on the assistance the treatment group members received from BAs, but also whether the control group members were able to access similar services available in the community. It would also serve to verify if control group members were able to access BA assistance by somehow circumventing study procedures. This information would help the evaluator interpret the study's impacts and shed light on the assistance other than that provided by BAs that is currently available to participants in need, which is not currently well understood. This data would have to be collected using a follow-up survey administered to both study groups to ensure the information is captured in a consistent way for the impact analysis. Although TAIS would contain information on services received by the treatment group, it obviously would not contain information on services received by the control group. Sample survey questions on service receipt are provided in Table II.2.

Outcome data. The key outcomes for an evaluation of the overall impact of the program would be perceived knowledge of benefits rights, perceived ability to advocate on one's own behalf, access to entitled benefits and related documents, and perceptions of a secure retirement and health. The analysis would use these outcomes to determine the impact of BA assistance. The considerations for collecting this data are similar to those discussed in Chapter II, Section C. As discussed there, existing data sources would not be sufficient to capture data on outcomes of both groups because both the Gallup survey and TAIS capture data only for inquirers who receive BA assistance; those assigned to the control group as part of an evaluation would not be surveyed by Gallup and would not have TAIS records tracking their access to benefits or other outcomes of interest.

Sample survey questions to measure outcomes for a classic RCT are largely the same as those provided in Table II.2. However, as mentioned previously, although customer satisfaction is an important performance measure for the program, it would be impossible to conduct an impact analysis on this outcome in a classic RCT because the control group would, by definition, not receive any BA services. The follow-up survey could also be used to collect outcomes of interest that are not currently collected by the program, such as the amount of time it took to receive benefits or related documents. Additional details are in Chapter II, Section C.

Finally, the same considerations hold for whether to use the Gallup survey or a new survey for data collection for an evaluation. Chapter II Section C has a detailed discussion of these considerations.

D. Evaluation Sample Sizes and Minimum Detectable Impacts for a Classic RCT

When discussing the evaluation of the web referral service delivery models compared with telephone services as usual, we discussed how to set target MDIs for the evaluation. In essence, EBSA, CEO, and the evaluator would work together to establish a threshold below which impacts would be considered acceptable because the goal of the evaluation would be to show that outcomes are approximately the same in the two service delivery models. For a classic RCT, the goal is different because we want to determine the size of the true impact of the program. Therefore, the impact evaluation sample size must be large enough so that the analysis can detect true impacts of BA assistance; the larger the sample size, the smaller the impact the study will be powered to detect. However, given that some inquirers would be denied access to BA services under this design option, an evaluator's goal should be to keep the sample size of the control group as small as possible while still being able to detect meaningful impacts.[23]

We used a series of assumptions and calculations similar to those presented in Chapter II to estimate MDIs for different sample sizes. If the impact evaluation were powered to detect an MDI of $2,000 in average benefits, the analysis would be able to detect an impact with high probability if the average benefits recovered for the treatment group were $2,000 more or less than the average benefits recovered for the control group. However, if the true impact were smaller than $2,000, the analysis would not be able to conclude as confidently that BA assistance had an impact on benefit recoveries. Therefore, the target MDIs (and the sample sizes and other design features that lead to the MDIs) have to be selected so that a study using this type of design would have a high probability of detecting meaningful impacts.

Typically, target MDIs for a classic RCT can be selected in several ways. Some evaluations use cost as a basis for determining acceptable MDIs. For instance, if it costs $100 to deliver a service, a study could be designed to detect an MDI of $100 to ensure that the program is cost-effective. Another approach is to use estimated impacts from evaluations of similar programs to determine target MDIs. For instance, if previous research on a program offering similar services had estimated an impact of 3 percentage points on knowledge of benefits rights, it might be advisable to set 3 percentage points as the target MDI for the evaluation of BA assistance. However, neither of these approaches is likely to work for an evaluation of BA participant assistance activities for two reasons. First, perceived knowledge about benefits rights, ability to advocate on one's behalf, and perceptions of a secure retirement do not correspond to a dollar value, so it is difficult to identify a point at which the program is cost-effective for these outcomes. Second, we could find no evaluations of similar programs that examined similar outcomes from which to draw estimated impacts.[24]

Because the typical approaches to selecting MDIs will not apply to the classic RCT, the evaluator could instead work with OEA and CEO to select target MDIs based on existing hypotheses about the likely size of impacts of BA activities. For instance, if OEA and CEO were reasonably confident that the program would affect a 2, 5, or 10 percent increase in perceived knowledge, the study could be powered to detect that particular percentage increase. Looking at historical data on perceived knowledge might provide some insight into the magnitude of such an impact. To get an estimate of a reasonable MDI for monetary recoveries, the evaluator could work with OEA to generate some basic estimates of the cost of providing BA assistance (that is, net of education and outreach activities and compliance assistance) and set the MDI accordingly.

[23] Ethical concerns about denying services to even a small control group were a primary factor in CEO's decision not to pursue a classic RCT at this time.

[24] It is critical when using estimates from previous studies that the nature of the services delivered and the outcomes examined are similar, not only the method in which those services were delivered. For instance, it would not be appropriate to use the impact estimates for an evaluation of a 3-1-1 telephone hotline system to set MDIs for this evaluation. Although 3-1-1 services are delivered in a similar way as BA services, the subject matter of the inquiries and outcomes of interest would not be similar enough to those offered by BAs to inform the selection of MDIs for this evaluation.

Another consideration to keep in mind when selecting the target MDI is that allowing the control group to receive some services (such as referral to the EBSA website or the ability to receive BA responses to web inquiries) would narrow the difference in service receipt between the study groups and narrow the expected impact of BA assistance. Therefore, a study design allowing control group members to receive some services would likely have to be powered to detect a smaller MDI than a study design in which control group members receive no services, if the goal of the evaluation is to determine the true impact of the program. This implies the need for a greater sample size when control group members can access some EBSA services.

MDI calculations. Although the goal of selecting a target MDI for a classic RCT differs from that when selecting a target for the web referral-only evaluation, the assumptions and calculations themselves are the same.[25] We also used the same example sample sizes of 25,000, 15,000, and 10,000 inquirers. These sample sizes should be considered in the context of the approximately 250,000 inquiries the BAs receive per year. Even if the approximately 11 percent of callers who reach the program in error or require resource assistance and the 15 percent of compliance-related calls were screened out, that still leaves approximately 185,000 inquiries per year that could be included in a classic RCT. Therefore, even with an evaluation sample size of 25,000, only about 14 percent of the annual total number of eligible inquirers would be needed to achieve the target sample sizes. We reproduce the MDIs in Table III.1 for ease of reference.

Table III.1. Minimum Detectable Impacts for a Classic RCT at Various Sample Sizes

	(1)	(2)	(3)
Sample Size	25,000 Randomly Assigned, 20,000 Complete	15,000 Randomly Assigned, 12,000 Complete	10,000 Randomly Assigned, 8,000 Complete
MDIs—Binary Variables			
Overall	**1.89%**	**2.44%**	**2.99%**
50% subgroup	2.67%	3.45%	4.22%
25% subgroup	3.78%	4.88%	5.97%
10% subgroup	5.97%	7.71%	9.44%
MDIs—Recoveries			
Overall	**$436**	**$562**	**$689**
50% subgroup	$616	$795	$974
25% subgroup	$871	$1,125	$1,377
10% subgroup	$1,377	$1,778	$2,178

Note: Binary variables include self-reported knowledge of benefits rights, ability to advocate on one's behalf, access to benefits-related documents, and perceptions of a secure retirement and health. MDIs for binary variables are expressed in percentage points. MDIs for recoveries are expressed in dollars. See Appendix C for a description of the full set of assumptions used to calculate the MDIs.

[25] In addition, if random assignment were conducted at the regional office level, we would have to take into account the extent to which the outcomes vary across sample members within the same regional office territory. Because the classic RCT discussed here relies on random assignment of individuals and not regional offices, we do not discuss this issue here. However, see Appendix D for a discussion of cluster RCTs.

As shown in column 1 of the table, with an evaluation sample of 25,000 participants (with the assumed 80 percent response rate, this would result in 20,000 completed follow-up surveys), the study could detect an impact as small as 1.9 percentage points in perceived knowledge of treatment group members compared with that of control group members. This means that if the true difference between treatment group members' perceived knowledge of benefits rights and that of control group members were 1.9 or more percentage points, the study would be able to detect that difference. Similarly, for monetary recoveries, an evaluation sample of 25,000 inquirers would mean the evaluation could expect to detect an impact of $436 in recoveries per participant. For reference, the average monetary recovery per person is $814, based on data from FY 2013.

As the study sample size decreases, it becomes more difficult to detect true impacts. For instance, with a study sample of 15,000 (column 2), which implies 12,000 completed follow-up surveys, the study could expect to detect an impact of 2.4 percentage points or more on binary outcomes; with a sample of 10,000 (column 3), which implies 8,000 completed follow-up surveys, the MDI is even higher, at 3.0 percentage points. This means that if the true difference in binary outcomes between the two groups were less than 3.0 percentage points, we could not be confident that a sample size of 10,000 would provide enough statistical power for the study to detect that difference, and the study could not conclude that BA assistance had an impact on those outcomes. Although, as mentioned in Chapter II, this would be fine in an evaluation comparing two service delivery strategies; however, in a classic RCT, this would make it seem as if the availability of BA assistance had no impact on inquirers' outcomes compared with no BA assistance.

Varying the random assignment ratio. In the MDI calculations, we assumed that half the study sample would be randomly assigned to the treatment group and the other half to the control group. This is the most efficient ratio possible and would minimize the total number of people who have to go through the consent process to achieve the target MDI. However, it would also be worth considering other random assignment ratios that would assign a larger proportion of evaluation participants to the treatment group; for instance, a 75:25 percent treatment-to-control ratio. This would have the advantage of assigning inquirers to the control group at a lower rate. However, to maintain similar MDIs with this treatment-to-control ratio, it would be necessary to increase the total sample size,[26] which would require a longer study intake period, including baseline data collection and consent with many more inquirers. OEA and CEO could decide whether a reduction in the random assignment ratio would be worth these tradeoffs.

The evaluation team would have to work in conjunction with OEA and CEO to select the target MDIs for these outcomes that best balance the need to detect true program impacts with the desire for a small evaluation sample size and a relatively small control group.

[26] Because the number of control group members is a key driver of the MDIs, shifting the random assignment ratio to this proportion without altering the total sample size results in larger MDIs. For example, randomly assigning 25,000 individuals with a 75:25 treatment to control group ratio and an 80 percent survey response rate, the MDIs for binary variables shift from 1.89% to 2.18% and for monetary recoveries from $436 to $503. This amounts to about a 15 percent increase in MDIs.

E. Implementation and Cost Studies

As useful supplements to an impact evaluation, we recommend that any evaluation have supporting cost and implementations studies. Doing so could help put the estimated impacts of BA assistance into context, particularly because the key outcomes for the evaluation include some that do not have an obvious monetary value, such as an increase in perceived knowledge of benefits rights.

An implementation study could provide useful information to help interpret the factors influencing the direction and size of impact results. Data from interviews with program staff and observations of service delivery would shed light on how the program operates on a daily basis and attempts to affect participants' outcomes. If the evaluation found greater or lesser impacts for a particular subgroup or for inquirers with certain types of benefits issues, an implementation study could provide meaningful information on why that occurred and possible modifications or enhancements to make to program services given the impact results.

A cost study could be combined with the impact results to assess the relative cost-effectiveness of the program. For example, suppose the impact evaluation found that knowledge of benefits rights was 5 percentage points higher among the treatment group and that the cost per participant was $200. This would enable stakeholders to assess whether this seems like a reasonable tradeoff. Or, if the impact on monetary recoveries is $500 and the cost per participant is $200, then the program could conclude the program is cost-effective with respect to that outcome. Cost studies can be done in many ways, ranging from simple methods, such as taking the total program budget and dividing by the total number of participants assisted, to far more complex methods, such as building the costs component by component and developing cost estimates for different services.

F. Challenges for a Classic RCT and Possible Strategies to Address Them

CEO, EBSA and the TWG members agreed that an RCT is the most rigorous way of evaluating the overall impact of BAs' participant assistance activities, but there are challenges associated with implementing such a design. The extent of concerns raised regarding these challenges is what led CEO to determine that it would not pursue this design. To provide context for this decision, we highlight several potential challenges that emerged during conversations with these stakeholders and propose strategies to help mitigate the concerns associated with each of them, in the event that CEO decides to move forward with this type of design in the future. Table III.2 provides a summary.

Challenge 1: Some inquirers would not receive comprehensive, immediate assistance from BAs. If a classic RCT were to be implemented, it would involve randomly assigning individuals to either a treatment group that can receive BA services as usual or a control group that cannot access BA services. It is critical that the study groups are as similar as possible in every way except for their random assignment status, including their level of need and interest in receiving services. This implies that some people who are legitimately in need of services would not be able to receive assistance. EBSA and TWG members expressed concern about this approach from an ethical perspective. Should it be determined at a later time that rigorous evidence about the impacts of BA services, developed through this type of study design, is

Table III.2. Summary of Challenges and Recommendations for a Classic RCT

Challenge	Recommend Strategies to Address Challenge
1. Some inquirers would not receive comprehensive, immediate assistance from BAs.	1. Exempt inquirers with emergency situations from the evaluation. 2. Offer access to the website for the control group. 3. Offer control group members access to services after an embargo period. 4. Create excess demand for BA services by marketing to underserved populations.
2. The evaluation would require modifications to normal intake procedures.	1. Hire contractors to screen incoming calls, collect baseline data, and conduct randomization. 2. Implement the evaluation nationwide for a short time. 3. Implement the evaluation in a subset of regional offices. 4. Limit randomization to a minority of calls.
3. Some callers might disengage because of the additional time required for consent and data collection.	1. Minimize the length of the consent process. 2. Pilot study procedures. 3. Limit data collection to as few variables as possible. 4. Provide training on administering study intake efficiently.
4. The study could affect performance measures.	1. Develop a process for the evaluation team to handle complaints. 2. Work with OEA/EBSA to adjust performance measures that the evaluation might influence.

desirable, we suggest several possible strategies to minimize the implications of denying services to the control group:

1. **Exempt inquirers with emergency circumstances.** We have heard anecdotal reports of people who seek BA assistance in the midst of a crisis that needs immediate attention. As mentioned in the text of this option, the evaluation would offer exemptions (sometimes known as wildcards) for use in emergency situations, such as being denied health care benefits in a life-threatening situation. These individuals would not be randomized to the treatment or control groups and would not be tracked as part of the evaluation.

2. **Access to the website for the control group.** As discussed earlier, we believe it would be worth considering allowing the control group to access the website and perhaps submit web inquiries. This way the control group members would be able to receive services, though in a format different from the traditional one-on-one telephone assistance from BAs. Although Chapter II presents, in detail, a design very similar to this approach, a summary of the potential issues includes the following:

   a. If using the website and submitting web inquiries is almost as effective a strategy for conveying information to participants as speaking directly with a BA, then estimated impacts will be smaller and the study would have to be powered to detect those smaller differences. That is, the evaluation's necessary sample size would be larger. Alternatively, if the evaluation detects no difference, then it could use implementation and cost studies to determine whether other programmatic or cost benefits encourage use of the website.

   b. Relatedly, this type of evaluation design would answer a different research question: it would estimate the impact of the program as it currently exists against the impact of the program if it were administered online only.

c. Some control group members might not have Internet access, in which case they would not be able to access any services. It is hard to assess the extent of this without conducting additional data collection before implementation of an impact study. Alternatively, those without web access could receive wildcards and be exempt from the study. The analysis would have to take this into account when characterizing the study findings.

3. **Allow control group members to receive assistance after an embargo period.** Another strategy for ensuring that control group members are not entirely denied services is to allow them to receive assistance after the evaluation team has collected data on key outcomes; this waiting period for service receipt is referred to as a service embargo period. We suggested a long enough follow-up period after random assignment to ensure that the evaluation can observe important outcomes for the majority of the treatment and control groups. Under this scenario, a control group member would be able to recontact the program after that period if unable to resolve a benefits issue within that time frame. Alternatively, at the end of the embargo period, BAs could directly reach out to control group members to determine whether they still had an issue.

4. **Create excess demand for BA assistance through marketing to underserved populations.** The participant assistance program has explicit outreach goals to engage more vulnerable populations, such as workers who face job loss, women and minorities, and individuals for whom English is not their primary language. In an effort to reach these vulnerable populations, CEO could consider infusing resources into an evaluation to fund new marketing and outreach activities targeted to groups that are not currently being fully served by the program. The evaluation team could even carry out some of these strategies, such as helping to develop a public awareness campaign or implementing a social media strategy. As the program reached more people in need of services, one consequence would be an increase in the overall call volume and more demand for assistance than could actually be met without bringing in additional help or reducing the current level or quality of service. In this context, a random assignment design would serve as a fair way of distributing the available resources. Although some inquirers would be denied services, the determination of who would be helped would be made randomly rather than based on some other factor, such as the persistence of the inquirer or time of day at which he or she called. Moreover, by pursuing this strategy, the program would be able to reach a broader spectrum of participants, some of whom are members of vulnerable populations who need assistance and who might not have been aware of the program in the absence of the evaluation-funded outreach activities.

Challenge 2: The evaluation would require modifications to normal intake procedures. Implementing a classic RCT would potentially change the process by which BAs and other regional office staff answer and deal with incoming calls. Instead of using the normal office procedures for answering calls and routing them to BAs, the process would have to be altered so that consent could be obtained, baseline data collected, and randomization conducted before BAs handle inquirers' questions. Regional office staff would have to adjust to a new call answering process, receive training on data collection and consent, and spend additional time on evaluation activities before each call that is not quickly screened or referred to another agency. We suggest several possible strategies to mitigate the evaluation's impact on existing procedures:

1. **Hire contractors to answer incoming calls, collect baseline data, and conduct randomization.** A small adjustment to the current service delivery system could help the program implement the study processes with less burden on BA staff. Lower-level staff—hired as contractors by the evaluation—could answer incoming calls and web inquiries, siphon off those that reached the program in error or required resource assistance resulting in a simple referral, obtain consent, perform random assignment, inform customers of their group assignments, and collect baseline data. Customers assigned to the treatment group would be forwarded to BAs as usual, and BAs would provide services as usual. Under this approach, BAs' activities would be less affected by the presence of the evaluation. EBSA and TWG members expressed a preference for this approach.

   A potential disadvantage of this approach is the handoff it would require from intake staff to BAs. OEA staff members have expressed concern that callers might get frustrated. Keeping metrics on the number and proportion of disconnects could help inform the extent of this problem should inefficiencies in the process occur, facilitating mid-course corrections in procedures, if needed.

2. **Implement the evaluation nationwide for a short time.** The key to the success of the evaluation is reaching the targeted sample sizes, more specifically the targeted control group sample sizes. Even the largest total sample size proposed—25,000—would likely be reached within several weeks if intake were conducted nationwide, given the volume of calls the program receives every year. The advantage to this approach is that the evaluation's intake process could be completed quickly and program operations would be disrupted for only a short period. Hiring contractor staff would reduce a potential disadvantage about the significant start-up resources that would be required to train all the BAs, if they were to perform intake for the evaluation. Another potential disadvantage is that, by confining the study to a short period, it might be subject to unusual spikes in call volume or type, which could skew the study's results.

3. **Implement the evaluation in a subset of regional offices.** Under this approach, OEA could work with the evaluation team to select 2 or 3 of the 10 regional offices in which to conduct the evaluation. Depending on the call volume at the selected offices, it might take only a few months to achieve the target sample size. The advantages of this approach are that it would require training only a portion of BA staff on study procedures, normal operations in most offices would continue, and the evaluation would still be completed relatively quickly. However, limiting the evaluation to a few offices could limit the extent to which the study's results can generalize to the other offices because there is variation in the nature of inquiries and regional office organizational structures and processes. In addition, the evaluation team would have to devise a system so that inquirers assigned to the control group could not circumvent the evaluation by calling a regional office that is not participating in the study. One potential solution would be for the evaluation team to conduct a short training webinar with all BAs (even those not affected by the evaluation) to demonstrate how to look up all inquirers in the random assignment system designed for the evaluation and reinforce the inquirers' study group assignments.

4. **Limit the evaluation to a minority of calls.** Not every call has to be included in the study. Selecting only a small proportion of all calls to be part of the evaluation could

limit disruption to the program; for instance, one of every 10 callers could be singled out for inclusion in the evaluation. Although calls selected for the evaluation would still have to go through the evaluation activities—baseline data collection, consent, and randomization—all other calls would be answered and handled in the normal way. This could be implemented either nationwide or in a subset of regional offices. In addition to limiting program disruption, this variation on the design has the advantage that it is drawn out over a longer period, so that the study would be less influenced by unusual spikes in call volume or type. If the current telephone system were not capable of diverting one of every 10 calls, OEA and CEO would have to invest in a system to accomplish this. The evaluation team would also have to put procedures into place to limit crossovers, as mentioned previously; this would likely be more difficult if inquirers discovered that by simply calling back the program, they would be likely to be passed straight to a BA for immediate service.

Challenge 3: Some callers might disengage because of the additional time required for consent and data collection. Every caller included in this type of study would have to go through the consent and baseline data collection process. OEA and the TWG members were concerned that some participants would become frustrated with the additional time required for the evaluation activities before they could receive an answer to their benefits-related question and might therefore hang up. This could be the case especially for the 75 to 90 percent of calls that can be handled quickly, in fewer than 10 minutes. We discuss several possible solutions to this challenge:

1. **Make the consent process as short as possible.** We believe that the consent process for this study could be fairly short, with callers giving their consent verbally over the telephone. Mathematica's institutional review board (IRB) expert confirmed that this would be possible, but any final decision would have to be made by an IRB that governs the evaluation and the Office of Management and Budget. We estimate that the process of obtaining consent and collecting the data items necessary for random assignment could take about three minutes per call.
2. **Pilot the study procedures.** The evaluation team could establish a pilot phase for the evaluation during which it would record metrics on how many callers refuse to consent or disengage from the call, to determine whether the process of enrolling in the study itself turns people away from receiving services. If there is an unreasonably high rate of disengagement, steps could be taken at that point to refine the consent process or pursue another avenue to mitigate disengagement.
3. **Limit baseline data collection to as few variables as possible.** There are a few baseline data items that would be necessary to collect before random assignment could take place. These would likely include name, one or more telephone numbers, and a handful of basic demographic characteristics. Although additional information about callers would be interesting and potentially informative, it would not be strictly necessary. Data items that are not likely to be influenced by the study or change over time could be collected during the follow-up survey rather than at baseline, although the ability to do a nonresponse bias analysis would be compromised.
4. **Provide training to implement study procedures efficiently.** All staff administering the consent process and conducting random assignment would receive training from the evaluation team on study procedures. Staff would be given ample

opportunity to practice using scripts and other information to perform the study activities efficiently and with confidence. Evaluation staff would provide this training.

Challenge 4: The study could affect performance measures. OEA and TWG members were concerned that individuals assigned to the control group might be dissatisfied with their inability to receive one-on-one BA assistance by telephone, resulting in a decrease in customer satisfaction scores and an increase in congressional inquiries in response to participants' complaints. In addition, customer satisfaction for the treatment group could also decline because of the increased time required by the consent process and/or handoffs from screening staff to BAs. We suggest several avenues to pursue with respect to these issues:

1. **Develop a process for the evaluation team to handle complaints.** We recommend that the evaluator develop a process for handling complaints about the study that is external to the participant assistance program. Complaints about the study would be routed to the evaluation team leadership. BAs could even provide a study contact number as part of the random assignment process. Based on our previous experience, we do not believe that complaints will occur often as long as everyone adheres to study procedures. We have successfully implemented random assignment evaluations in many different contexts, including those in which control group members were denied access to thousands of dollars worth of services. For example, in the Workforce Investment Act Gold Standard Evaluation, in which more than 30,000 participants were randomly assigned, about 1.5 percent called a customer hotline set up for the evaluation and only 0.2 percent of the 30,000 participants complained about not having access to training funds.
2. **Work with OEA/EBSA to adjust performance measures that might be influenced by the evaluation.** If the Gallup survey were not adjusted for the evaluation, then only customers who received BA services (that is, the treatment group) would be in the Gallup sampling frame and control group inquirers would not be sampled. If the Gallup survey were adjusted in the ways necessary for the evaluation (as outlined in Section II, Section C), it would include control group members; however, these individuals would not, by definition, have received BA assistance and so would not be asked questions about their satisfaction with BA services. However, it is possible that treatment group members might experience a decline in customer satisfaction caused by the evaluation's procedures. The evaluator could assist OEA in advocating for the exclusion of evaluation members from its calculation of customer satisfaction.

   In addition, although even the maximum evaluation sample size proposed would involve assigning only 12,500 people to the control group (from about 250,000 served on a yearly basis), it is conceivable that this could result in, for instance, a lower number of inquiries being referred to OE or decreases in other performance measures based on the volume of calls. The evaluator could support OEA in adjusting the target performance measures to account for the effect of assigning some people to receive less than full BA assistance.

IV. RECOMMENDATIONS FOR PERFORMANCE MEASUREMENT

Beyond designing possible impact evaluations, Mathematica was also contracted to review OEA's existing performance measurement system. OEA currently gathers an extensive set of input, activity, and output counts, as well as some outcome measures, to assess program performance. Although there is some overlap in the outcome measures of interest to an impact evaluation and those needed to manage the day-to-day operations of the program, the two tasks are largely disparate. This chapter presents our recommendations for supplementing and revising existing performance measures and creating new measures to better suit the needs of OEA as it oversees its program activities. The discussion is informed by best practices in performance measurement and Mathematica's experience developing performance measurement systems for DOL.[27]

The measures addressed in this chapter include the five identified as FY 2013 priorities by OEA and additional priorities identified through the information-gathering activities used to develop the logic model presented in Chapter I. The recommendations in this chapter were developed while EBSA's 2014 operating plan was under development. In both years, customer satisfaction was the primary performance measure of the Participant Assistance program, but several of the other priorities shifted in 2014. In a memo from EBSA in response to our draft report, the program indicated that it had implemented many of the recommendations contained in this chapter. We document EBSA's planned revisions when appropriate throughout the chapter.

Summary of Performance Measurement Recommendations for the Participant Assistance Program:

- Maintain existing measures and counts of program operations.
- Supplement these with percentage measures that better align with program office goals and more accurately capture activities under the program's control.
- Revise the Gallup survey to enhance the program's ability to analyze customer satisfaction and include questions regarding access to and satisfaction with the program's website.

As shown in Table IV.1, we have organized the measures discussed in this chapter by the sections of the logic model to which they pertain (see Exhibit I.1 for a visual representation of the logic model). Each section provides recommendations for developing modified or new measures listed in the table. We do not recommend that the program stop collecting any existing measures or counts.

[27] According to Borden (2011) in *The Challenges of Measuring Performance*, "… identifying relatively good or bad performance and measuring improved or decreased performance (requires) a rate of success and not simply a count of activities.… Standards that identify minimally acceptable performance must be associated with measures." Borden also emphasized the need for clear definitions, stating that "Seemingly simple concepts such as … whom and when to count must be defined very precisely for performance results to have meaning."

Table IV.1. Existing Priority Measures and Suggested Supplemental, Revised, or New Measures

Section of Logic Model	Existing Priority Measures	Supplemental, Revised, or New Measures
Outputs	Number of inquiries, by type of inquiry	Percentage of inquiries by type of inquiry (supplements existing measure)
	Number of BA referrals to enforcement accepted for investigation within 60 days[a]	Percentage of BA cases referred to enforcement with a decision to accept, to reject, or not yet decided, within 30 days (supplements existing measure)
	Number of national and regional compliance activities;[a] number of Rapid Response sessions;[a,b] percentage of congressional offices briefed[a]	Percentage of outreach or education activities by type of activity, based on national and regional activity standards (supplements existing measure)
Outcomes	Participants' knowledge of health and pension benefits rights	Participants' knowledge of health and pension benefits rights (revision to existing measure)
	Participants' customer satisfaction[a]	Participants' customer satisfaction (revision to existing measure)
	n.a.	Knowledge of fiduciary responsibilities related to compliance assistance (new measure)
	n.a.	Compliance assistance customer satisfaction (new measure)

[a] Items identified by OEA as a priority for FY 2013.

[b] In EBSA's 2014 operating plan, the number of Rapid Response sessions was not included as a performance measure. OEA has indicated that it is still a priority, but it is now a measure of demand for services, driven by the magnitude and number of layoffs in each region.

n.a. = not applicable.

A. Output Measures Used to Monitor Program Performance

In FY 2013, OEA had two priority measures related to participant and compliance assistance: number of inquiries by source and number of BA referrals to enforcement accepted for investigation within 60 days. The agency also used counts of the number of national and regional compliance activities and Rapid Response sessions as priority measures for outreach and education activities. We discuss our recommendations for supplementing each of these measures next. Importantly, OEA indicated in response to the draft of this report that, in 2014, the agency decided not to prioritize the number of Rapid Response sessions as a performance measure. It will continue to prioritize Rapid Response but it is considered a measure of demand for services. Regional offices are directed to plan their Rapid Response activities based on the magnitude and number of layoffs in their regions, and to assess the cost effectiveness of traveling to an on-site location that requires overnight travel in light of travel restrictions in the budget. Outreach to dislocated workers after job loss is a priority, but OEA considers handling inquiries to be the top priority for BAs.

1. Participant and Compliance Assistance Priority Measures

We recommend supplementing existing output measures related to both participant and compliance assistance. We suggest supplementing the program's counts of the number of inquiries received by type of inquiry with percentages received by source. We also suggest shifting the focus of BA referrals to enforcement to percentages of inquiries accepted, rejected, or not yet decided within 30 days, while retaining the program's existing measure. Both of these changes enable the program to track activities based on measures within the control of the BAs.

a. Type of Inquiry

In FY 2012, BAs responded to 240,110 inquiries. As shown in the logic model, those seeking assistance can contact the program through a number of methods, including telephone, the Internet, mail, email, and visiting a regional office. Currently, the predominant mode of contact is by telephone, with more than 93 percent of inquiries received this way. About 3 percent of inquiries are submitted via the Internet. To support further growth in web inquiries and help reduce telephone inquiries to the BAs, OEA created a consumer assistance page to address simple inquiries. It also developed a web portal so that participants can submit web inquiries rather than calling a BA. Web inquiries automatically populate TAIS, eliminating the need for BAs to collect and enter some data.

Recent pressures on government services and costs, existing staffing constraints, and the move to have BAs conduct employer compliance reviews suggest that the program will have to direct participant and compliance assistance inquiries to web services to a greater extent than in the past. The program has noted this and developed a performance measure to gauge progress on this front; that measure is the total number of inquiries received via the web. To further enhance the agency's ability to track progress on web inquiries, we recommend new percentage-based measures along with an approach to setting manageable standards today and increasing them over time. To help OEA better focus its efforts on increasing web inquiries, we also recommend supplementing the existing inquiry measures with additional Gallup survey questions to assess the extent to which customers have access to Internet technology.

Recommended performance measure for type of inquiries. We recommend that, going forward, the program supplement its counts of inquiries by source with a measure of the percentage of total inquiries received by each source. The percentage of web, telephone, mail, and other sources would equal 100 percent of inquiries. In addition, we recommend establishing standards across inquiry types to incentivize the program to increase the percentage of web inquiries received over time. In their response to our draft report, OEA indicated that it is moving forward with implementing these changes to the inquiry type measures for FY 2014.

Adding this comprehensive measure to the existing counts of inquiries would have several benefits for the program. First, this measure would better align with the program office's goal of moving inquiries to the web, which would presumably reduce the time BAs spend on calls that simply require referrals or general inquiries and increase their time spent on data collection and more complex inquiries and activities (see Appendix A for ideas on how to drive traffic to the website). Second, this revised measure more accurately captures activities under the program's control. Because the total number of inquiries fluctuates from year to year and is driven by changes in federal policy and regulations; participants' and employers' needs; and the behaviors of other key stakeholders such as health insurance companies, pension administrators, and brokers, that number is outside the control of the Participant Assistance program. Using the percentage measure means that the performance standard can be modified each year to track and incentivize progress toward achieving the goal of increasing the proportion of inquiries received via the web, regardless of the total number of inquiries.

As mentioned previously, the program already measures counts for each type of inquiry and we recommend that it continues to do so because these data provide useful information about trends that can be used for policy development, outreach design, budget requests, and program modifications. The counts can be used very easily to develop the percentages we recommend adopting.

The program should also set standards against which to measure progress toward its goal of increasing web inquiries, understanding that directing callers to the web is challenging. We recommend starting with the percentages derived from the most recent FY or quarter as the baseline, and then setting modest and achievable standards for increases for each subsequent quarter or year. For instance, the baseline measure for web inquiries received in FY 2012 is 3 percent. The program could set a goal of increasing that by a few percentage points for FY 2013, pending redesign of the contact form and resolution of technical issues on the website. Table IV.2 demonstrates how the program might develop standards for shifting inquiries from telephone to the web over time, assuming other inquiry sources remain constant.

Table IV.2. Example of Standards, by Type of Inquiry and Year (percentages)

Inquiry Type	FY 2012	FY 2013[a]	FY 2014[a]
Telephone	93.0	90.0	87.0
Web	3.0	6.0	9.0
Mail	3.6	3.6	3.6
Email, Walk-In, and Other	0.4	0.4	0.4
Total	**100**	**100**	**100**

[a]These percentages are hypothetical targets.

Factors to consider in setting standards for future years might include how high and low volumes of inquiries are expected to affect the distribution across types of inquiries. The program might also choose to differentiate between standards by type of inquiry for participant and compliance assistance—for example, building the employer and plan sponsor relationship through telephone assistance could have different implications for ERISA compliance. Therefore, the program might choose to set different standards for telephone inquiries on compliance assistance accordingly.

In addition to setting goals to increase web inquiries over time, OEA might also be better able to focus its efforts on increasing use of the website by tracking how many and which inquirers have access to the Internet or feel comfortable using Internet technology. This could be tracked by adding a small number of questions to the Gallup survey based on standard Internet use and satisfaction questions, such as those used in the General Social Survey and the Current Population Survey supplement on computer and Internet use. The Gallup survey instrument has already been approved by the Office of Management and Budget for FY 2014, but these changes could be made in FY 2015 if funding is available.

Suggestions for reviewing BA data collection efforts. To help OEA identify trends in inquiries and better measure performance, we recommend a review of data collection efforts related to the types of inquiries received and their disposition. Table IV.3 categorizes closed inquiries by the codes captured in TAIS.

Table IV.3. Fiscal Year 2012 Closed Inquiries

Closed Inquiries	Volume	Percentage of Closed Inquiries
EBSA 15 Total Participant Assistance Inquiries Closed	239,520	100.00
EBSA 15a Benefit Claims Assistance	158,611	66.22
EBSA 15b Benefit Recovery or Referral for Investigation	6,995	2.92
EBSA 15c Complaint Analyzed (no referral or recovery)	5,361	2.24
EBSA 15d Compliance Assistance	34,722	14.50
EBSA 15e Resource Assistance or Other Calls	33,831	14.12

Source: FY 2012 production data worksheet.

About two-thirds of participants' inquiries are for benefit claims assistance (15a). Anecdotal reports from BAs suggest that many of the inquiries coded in this category are purely informational and that some proportion of them are referred to the website. However, we did not have access to a more detailed breakdown of the disposition of this large category, such as by the status codes used for closure analysis. In addition, inquiries coded under resource assistance or other calls (15e) likely include misdirected calls or people who contact the program to obtain contact information for other organizations, such as their insurance companies.

EBSA convened a working group in FY 2013 to conduct a quality review related to the technical sufficiency of assistance to inquirers for whom the interaction did not result in the identification of a valid claim or recovery of a benefit. This review is ongoing in FY 2014 and is intended to provide more detailed data about the range of complexity of the inquiries coded as Benefits Claims Assistance (15a). When this analysis is complete, OEA might consider whether routine reports or additional standardized data items could help the regional offices better understand the nature of such inquiries. This could include whether a call was related to an explanation of benefits or potential private pension notice and whether the caller was referred to the website. Unique action codes could be added to TAIS for these kinds of situations and documented in the TAIS manual to facilitate standardized coding and analysis. Without increasing data entry and reporting efforts significantly, the resulting data could provide a greater understanding of whether redirecting these types of inquiries could enable BAs to increase the value of services they can provide (that is, those requiring in-depth assistance) and could help the program better plan BA training and workflow.

In addition, our information gathering revealed that BAs do not have a good anecdotal sense of the characteristics of callers. As discussed in Chapter II, most data items on participants' demographics are currently optional for callers to report and BA staff to record in TAIS, as are plan and employer information (unless the participant has a complaint that requires intervention with the employer). Although collecting employer information for every participant might have some analytic value, this would have to be weighed against the staff time required to collect the data.[28]

[28] To facilitate collection of employers' names, OEA could consider working with the EFAST2 program to determine whether basic information on employers can be integrated into the TAIS system. For instance, it could prepopulate the employer field with the list of employers filing under the EFAST2 system.

b. Referrals to Enforcement

OEA established a participant assistance measure in FY 2013 for the number of BA referrals to enforcement opened for investigation within 60 days. A case is opened for investigation after a decision is made to accept a case for investigation, ideally within 30 days of referral to enforcement. We recommend that OEA continue to measure referrals opened for enforcement within 60 days, but place a higher emphasis on measuring performance of decision making at the existing 30-day benchmark to better reflect performance within the control of the Participant Assistance program and timely disposition of all cases referred to enforcement. In its response to our draft report, OEA indicated that it is moving forward with implementing these changes to the enforcement referral measures for FY 2014.

Recommended performance measure for referrals to enforcement. We suggest adjusting the focus of OEA's measure from a count of referrals to enforcement to percentages of BA cases referred to enforcement with a decision to accept, to reject, or not yet decided, within 30 days.

We recommend adopting this measure for several reasons. First, this approach shows the allocation of all cases rather than counts, which in isolation are difficult to interpret. The not yet decided rate would ideally approach zero; this measure would provide a quick and easy point of reference to determine the program's progress toward its goal of having decisions made on all cases within 30 days. Second, this measure would provide quicker feedback to BAs than the current system on the kinds of cases that are acceptable for enforcement and those that might require more in-depth consideration before acceptance for enforcement. Third, respondents in the national office noted that, despite their best efforts, they cannot predict how many inquiries they will receive in a given year because of economic shifts, new legislation, and other issues outside their control. They also cannot predict the types of inquiries received; very few require the involvement of enforcement. This measure can help to identify potential backlogs earlier in the process in a year with more referrals to enforcement than usual.

Further, the BAs have little control over what happens to an enforcement referral after it is submitted and offices vary in the criteria used for accepting these referrals; in who makes the final decision to refer; and, on the enforcement side, the rationale and acceptance rates for the referred cases. For example, one office noted a 50 percent referral acceptance rate and another noted a 98 percent rate. When a referral has been accepted, enforcement staff, not BA program staff, are responsible for how quickly the case is opened and resolved. Therefore, using percentages rather than counts in conjunction with a 30-day window shifts the focus to the outcomes that are more directly under the BAs' control and away from things they cannot control, such as the number and nature of inquiries received and whether enforcement decides to open a case. Finally, this measurement could also be a tool for the national and regional offices to use to assess differences in enforcement referral and acceptance approaches across offices.

Implementing this new measure would be straightforward because data on the number of referrals to enforcement and the number of referrals to enforcement opened as enforcement cases are already collected; these would only have to be converted into percentages. Again, the program would have to establish standards, most likely for each regional office because of the considerable variation in policies and procedures related to enforcement referrals across offices. The national office could use the percentage of cases decided within 30 days from FY 2012, or more recent quarterly numbers, as a baseline and set discrete targets for improvement in subsequent periods.

Table IV.4 provides an example of how this measure might look over time; analysis of the measure might lead the program to make operational changes to decide cases more quickly and to improve understanding of which cases to refer for enforcement.

Table IV.4. Example of Measure of Decisions on Referrals to Enforcement Within 30 Days (percentages)

Decisions Within 30 Days	FY 2013	FY 2014	FY 2015
Accept	20	40	60
Reject	40	30	20
Not Yet Decided	40	30	20
Total	**100**	**100**	**100**

Note: All percentages in the table are hypothetical and intended for expository purposes only.

2. Outreach and Education Activities

For FY 2013, OEA had three performance measures based upon outreach and education outputs: (1) the number of national and regional compliance assistance activities conducted, (2) the number of Rapid Response sessions conducted, and (3) the percentage of congressional staff briefings conducted. Like the other priority measures described in this chapter, each has advantages. All are relatively straightforward to measure, can be consistently measured across the regions, and are easily understandable by staff in the national and field offices. However, we recommend adjusting the first two measures to strengthen their linkage to the activities under staff control. The third measure was converted to a percentage in 2013—to brief 95 percent of the offices of all newly elected members, member offices that have never been reached, and member offices that have not been briefed within the past two years. OEA tracks offices briefed and can monitor this new congressional briefing measure. It is therefore not included in the following discussion.

Recommended performance measure for outreach and education. OEA did not make Rapid Response a priority measure for 2014 even though OEA required a minimum number of compliance assistance outreach activities more generally. We have included Rapid Response in this discussion because it represents an important component of the established targets. In that light, we recommend introducing flexible percentage standards across the compliance assistance and Rapid Response output measures. Unlike the current outreach and education measures, this would ensure that the total number of outreach activities remains constant but would allow the regions flexibility to adjust the types of activities that they conduct to accommodate region-specific needs while working to achieve comprehensive outreach and education performance standards set for the region and overall. The comprehensive performance standards will ensure that all regional offices are working aggressively toward the overall program goals. These standards will be established based on the target number of activities that can be conducted within national and regional budgets.

We recommend adopting this approach for performance measurement related to outreach and education activities because of the challenges respondents identified with having separate measures for different types of these activities. First, although conducting a certain number of compliance assistance activities seems straightforward to implement, national office respondents and several regional office respondents indicated that limited budget resources have hindered their ability to meet the targeted number of activities. In contrast, one regional director reported there were many alternative resources for employers in that region, so businesses did not see the need or value in OEA's compliance activities. That made it difficult for BAs in that regional office to achieve the required level of outreach activities.

Second, several respondents in the national office and one in a regional office viewed having to conduct a certain number of Rapid Response sessions to meet a performance target as problematic. One reason was the difficulty in setting and achieving targets for these activities because the need for them depends on the number of companies experiencing mass layoffs. It is difficult to predict how many mass layoff events will occur nationally or in a given region, even when using previous years' data as a baseline. In addition, according to one national office respondent, the deputy secretary's quarterly review process penalizes OEA for either falling short of or exceeding targets by more than 5 percent. If more mass layoffs occur than were anticipated and the regional offices respond by attending additional Rapid Response sessions to meet that demand, they are penalized. Alternatively, if fewer layoffs than expected occur that require Rapid Response, the program is also penalized.

OEA has acknowledged these issues and decided not to include the number of Rapid Response sessions as a priority performance measure in FY 2014. The performance measure we recommend here largely mitigates these concerns but also enables OEA to continue tracking Rapid Response sessions as a component of outreach and education. The recommended measure focuses on an overall total target number of activities to conduct within the region's budget, but provides the regions flexibility in how they attain the total across the compliance assistance and Rapid Response activities. Within the overall goal for outreach and education, OEA could maintain minimum standards for the number or percentage of each kind of activity to ensure that agency and departmental goals are met. For instance, regions experiencing a relatively high number of mass layoffs could increase the number of Rapid Response sessions held. Regions would continue to conduct some compliance assistance outreach activities to ensure that agency goals are met, but they would temporarily shift some outreach resources toward Rapid Response sessions to meet immediate demand. Each quarter, as performance is reviewed, the national office and each region could reassess performance against the previously set standards to determine whether outreach and education performance as a whole is on track and whether minimum standards are being met. The reasons that performance within each category or overall differ from the original standards can be documented to inform the development of future standards.

Table IV.5 provides a hypothetical example of how the standards, actual performance, and subsequent year's standards might look in a given region. OEA and regions can develop future standards based on the current year's actual performance; projections about whether trends affecting current performance will continue; and knowledge of program budgets and other factors, such as new legislation, that will influence outreach and education activities.

Table IV.5. Example of Outreach and Education Activities Measure in a Given Region (percentages)

Type of Activity	FY 2013 Standard	FY 2013 Actual	FY 2014 Standard
National/Regional Compliance Assistance	60.0	20.0	35.0
Rapid Response Sessions	40.0	80.0	65.0
Total	**100**	**100**	**100**

Note: All percentages in the table are hypothetical and intended for expository purposes only.

B. Outcome Measures Used to Measure Program Performance

OEA currently uses customer satisfaction and inquirers' perception of increases in their knowledge of benefit rights as key outcome measures for participant assistance activities. Little information is captured on the outcomes of compliance assistance activities. We discuss our recommendations for modifying and adding to these measures in this section.

1. Participant Assistance Outcome Measures

The Participant Assistance program currently measures customer satisfaction and perceived increase in knowledge of health and pension benefits rights through a telephone survey administered by Gallup. The customer satisfaction standard for FY 2013 is 69 percent. That is, the target is for at least 69 percent of customers to indicate that they are satisfied or extremely satisfied overall with the information, products, and services that they received from EBSA, using a 5-point Likert scale. This is the program's only measure related to the Government Performance and Results Act and is reported to the deputy secretary. By linking the survey responses with TAIS data, Gallup also reports customer satisfaction by region, individual BA, inquiry topic, and inquiry final disposition. The survey also includes a simple measure of the respondent's perceived increase in knowledge of health and pension benefits rights. To further expand potential analysis of these existing measures, we recommend that OEA consider adding some questions to the Gallup survey and modifying others. We also suggest limiting the sampling frame from which survey respondents are drawn to exclude those requiring resource assistance resulting in simple referrals. We describe additional recommendations for modifying the Gallup survey for purposes of the impact evaluation in the discussion of outcome data in Chapter II. Because the Office of Management and Budget has already approved the FY 2014 Gallup survey instrument, we recommend implementing these changes for FY 2015.

Recommended performance measures for participant assistance. To support OEA's measurement of customer satisfaction and knowledge of health and benefit rights, we suggest modifying the Gallup survey items in three ways. First, as discussed in Chapter II, we recommend that OEA consider adding questions to the customer satisfaction survey about the subject of the participant's inquiry. We also recommend moving the questions about the resolution of the inquiry to the beginning of the survey, before the customer satisfaction questions. Although the survey is currently administered only to individuals who recall contacting EBSA within the past few weeks, respondents are not asked to confirm the nature or resolution of their inquiries until the end of the survey. This confirmation should occur at the beginning of the survey to ensure that the respondent thinks about his or her specific interaction(s) with BAs when responding to the satisfaction questions and separates whether the issue was resolved from how that resolution affected the respondent.

Many of the staff in the national office and field offices with whom we spoke during the logic model development mentioned a concern about how to interpret customer satisfaction scores. For instance, a caller who has recently separated from his or her job might want to access pension benefits immediately but is informed by the BA that he or she must wait a period specified in the plan rules before doing so. Although this is a positive outcome—it helped achieve EBSA's and OEA's mission of providing information about pension beneficiaries' rights—the customer might respond negatively on the survey regardless of the quality of the BA service simply because he or she learned that it was necessary to wait to access benefits.

We therefore recommend inserting a type and resolution of inquiry confirmation before the customer satisfaction question. Currently, customers are asked about the resolution of their inquiry at the end of the survey, the response options are limited, and they do not directly correspond with TAIS closure analysis codes. The response options for this new question or questions could be revised to correspond to the closure analysis codes and other key codes that OEA uses for survey analysis. Among other benefits, this approach would enable OEA to analyze recall error when it matches survey responses to the respective closure analysis codes in TAIS for each survey respondent. A pre-test of these new questions could determine how much burden results from the change.

Eliminating the questions that ask the respondent to indicate how EBSA always behaves in a series of settings could offset the time required to respond to these new questions. (The survey documentation does not provide enough information to determine whether these questions feed into Gallup's proprietary customer satisfaction measure.) These questions as worded do not seem applicable when the majority of respondents to the survey do not have repeated contact with EBSA. If these questions cannot be eliminated because a multiquestion measure is preferred for calculating customer satisfaction, we recommend revising the question wording to reflect respondents' typical experiences with EBSA.

Second, as discussed earlier in this chapter, a small number of questions could be added to the Gallup survey to assess the extent of inquirers' access to the Internet, comfort using Internet technology, and satisfaction with OEA's website.

Third, we recommend limiting the sample frame for the Gallup survey based on TAIS closure analysis codes so that the survey is not administered to individuals requiring resource assistance resulting in simple referrals. In its response to our draft report, OEA indicated that it is implementing these changes to the Gallup sampling frame. Beyond this, additional data collection through TAIS would facilitate sampling by inquiry disposition for the majority of benefit claims assistance inquiries.[29]

Another approach would be to explore the validity of sampling all inquiries in closure analysis codes 15b and 15c: (1) benefit recovery or referral for investigation and (2) complaint analyzed—no referral or recovery, respectively. These dispositions could be prepopulated into new, detailed survey items that inquire about satisfaction with how the complaints were handled, as these cells cover the range of outcomes associated with the BAs' more complex cases. OEA indicated in its response to our draft report on August 23, 2013, that it has analyzed satisfaction by inquiry topic and disposition by linking the survey responses with TAIS data, and it has learned that satisfaction is not always positively correlated with recovery amounts. As noted earlier in this section, we recommend modifying the survey design and sampling so that OEA can expand this analysis.

[29] According to the survey's Office of Management and Budget package Part B, the current sample stratification is by the 10 regional offices. The expectation is that the random sample for each data collection period within each office is "likely to include proportional representation of cases (inquiries) by Closure types (those who need benefit claim-assistance—80 to 90% of cases, those who have a valid benefit claim, and those who have an invalid benefit claim)." If response rates vary significantly across closure types, nonresponse adjustment is considered. It is not clear which category calls about EOBs and others that are misdirected belong to, and how respondents are identified in the file.

2. Compliance Assistance Outcome Measures

At present, the program captures little information about the outcomes of employers and plan sponsors served by the BAs, even though these sources represented 14.5 percent of all FY 2012 closed cases. Notably, OEA administrators indicated that, in the past, the Gallup survey captured information on compliance assistance, but the agency eliminated that aspect of the survey because they found a high rate of satisfaction among compliance assistance customers (82 percent) and they wanted to focus efforts on participant assistance. Nevertheless, ensuring that employers and plan sponsors are satisfactorily served can yield important leverage for meeting EBSA’s goal of encouraging voluntary compliance with ERISA, because improving understanding and compliance for a single employer or plan sponsor has the potential to affect a substantial number of participants. We recommend including compliance assistance calls in the sample frame for the 2015 Gallup survey if funding is available.

Recommended performance measure for compliance assistance. Adaptations to questions can be made to capture knowledge of fiduciary responsibilities under ERISA and customer satisfaction for this sizeable portion of the program’s customer base. The outcomes that a compliance assistance survey could capture are parallel to those that a modified participant assistance survey could capture. The compliance assistance satisfaction survey could be developed and administered in tandem with the participant assistance customer satisfaction survey. However, the program might want to consider developing a web-based survey for this respondent population, because they are more likely to have repeated contact with the program. A web survey would have the added benefit of directing attention to the program’s website.

Focus groups of employers and/or plan sponsors are another option for obtaining feedback on compliance assistance. Although focus groups would reach a smaller set of respondents than would a survey, feedback through focus groups would be richer, more direct, and interactive; the format would also serve to further develop the relationship between the program and employers, plan sponsors, and others who participate in the focus groups.

This page has been left blank for double–sided copying.

REFERENCES

Borden, William. "The Challenge of Measuring Performance." In *The Workforce Investment Act: Implementation Experiences and Evaluation Findings*, edited by D.J. Besharov and P.H. Cottingham. Kalamazoo, MI: Upjohn Institute, 2011.

Bellotti, Jeanne, Annalisa Mastri, Julie Bruch, Grace Roemer, and Beenu Puri. "A Logic Model for the Outreach, Education, and Assistance Program." Memo submitted to the U.S. Department of Labor. Princeton, NJ: Mathematica Policy Research, August 5, 2013.

This page has been left blank for double–sided copying.

APPENDIX A

RECOMMENDATIONS TO ENCOURAGE USE OF THE EBSA WEBSITE

This page has been left blank for double–sided copying.

Over the course of developing the evaluation design and performance measures recommendations, we identified several potential approaches that the Office of Outreach, Education, and Assistance (OEA) might consider for encouraging individuals to use the Employee Benefits Security Administration (EBSA) website instead of calling the telephone hotline. In line with OEA's goals, our recommended approaches could help reduce the number of telephone inquiries, encourage individuals to resolve their issue through self-service on the website, and encourage individuals to use the web portal to submit inquiries about any issues they could not resolve without assistance. If these activities are successful, the Benefits Advisors (BAs) would be able to spend more time providing assistance with in-depth inquiries that potentially add the greatest value within the participant assistance program.

We suggest that OEA consider implementing four strategies to direct inquirers to the web. These include (1) coordinating with other entities and agencies to list the EBSA website rather than the hotline on prominent materials, (2) changing the recorded message when individuals call the hotline and are placed on hold, (3) exploring additional features of an advanced interactive voice response system (IVRS) to direct more callers to the website before being connected to a BA, and (4) increasing public awareness efforts. In its response to the draft version of this report, EBSA indicated that it is actively pursuing strategies related to several of these recommendations. We document those strategies in the following discussion, as well as ones not currently being implemented by EBSA but which we think are worthy of additional consideration. In addition, variations on some of these strategies have been implemented in the past, and we provide suggestions for potential enhancements.

List the EBSA website rather than the hotline on prominent materials. As noted in Chapter I, approximately 11 percent of telephone inquiries are from callers who reached the program in error or require resource assistance to be referred to other agencies or companies. During our information-gathering activities, BAs in all of the regional offices estimated that 75 to 90 percent of participants' calls could be handled in a single interaction and take fewer than 10 minutes to resolve. Although the time to respond to each of these inquiries individually is minimal, the cumulative resources required contribute notably to the BAs' workload. Using the estimates provided, BAs might have spent more than 4,000 total hours in fiscal year (FY) 2012 or the time of more than two full-time equivalent staff members providing information to these inquirers. (As context for this figure, the national office reported that the program supports a total of 108 BAs in its 12 field offices.)

To promote use of the website, OEA recently began to highlight the website address on every press release, new publication, fact sheet, frequently asked questions (FAQs), and notice issued by the agency. Although this has resulted in increased web traffic, the number of web inquiries has remained stable. In order to further increase web traffic and to increase web inquiries, OEA could consider promoting the website address on other types of key materials commonly seen by callers who could potentially benefit from the website. For instance, BAs anecdotally reported that many of the single-interaction calls—which could potentially be addressed through information on the website or through web inquiries—are generated because the toll-free hotline number is listed prominently on two key types of documents: (1) health insurers' explanation of benefits (EOB) forms and (2) Social Security Administration notices of potential private pension (PPP) benefits. BAs reported that most callers who obtain the toll-free number from their EOB forms intend to call their health care providers with questions about specific medical claims and do not intend to contact a BA with questions or concerns about denial of benefits. They also suggested that many callers responding to PPP notices expect to receive information on how to access the listed benefits and need very simple instructions on how to contact their prior employer or pension plan to find out whether they are indeed eligible for those benefits.

Interviews with regional staff suggest that some portion of these inquiries could be resolved by directing participants to review available information on the website. If health insurance providers and the Social Security Administration replaced the OEA hotline number with the website address on EOB forms and PPP notices, respectively, a substantial portion of single-interaction inquiries would be directed to the website. This change could significantly reduce the number of misdirected telephone calls after EOB and PPP mailings and thus increase the amount of time BAs have to spend on inquiries directly related to issues of health care coverage and pension benefits rights.

Changes to the telephone message. Currently, the telephone message that people hear when they call the program provides the agency's website address, but does not provide specific information on what one could do by going to the website. The agency's website home page includes links to FAQs and the web inquiry portal; the telephone message could be enhanced to be more explicit about where to find these resources and their use. We suggest that the message be enhanced to do the following:

- Direct participants to an FAQ document posted on the website with specific language describing where to find the document and what information it provides:
    - For instance, "Go to www.askebsa.dol.gov and click on the FAQ link on the left-hand side. You can use this to find information about COBRA notices, potential private pension notices from the Social Security Administration, and other topics."
    - Note that the content on the FAQs page would probably have to be simplified.
- Encourage the submission of web inquiries:
    - For instance, "Go to www.askebsa.dol.gov and click "Contact EBSA" at the top of the page. Our trained staff will review your inquiry and contact you by telephone or email within two business days."

This simple change could inform callers about the resources available on the website, the extent of which they might not be aware. With this additional information, more callers might choose to use the website to find information and submit inquiries instead of speaking to a BA over the telephone.

Notably, EBSA indicated in its response to the draft of this report that adding instructions for getting to the web portal from the homepage as recommended by Mathematica was unnecessary. EBSA staff indicated that the web address currently given on the hotline message takes users directly to the consumer assistance page. That page links to the most commonly asked questions and FAQs about health and retirement laws and has three different buttons with different headings that lead directly to the web portal for submitting an electronic inquiry. EBSA hoped that, if the lines are busy or there is a wait to speak with a BA, callers would contact the BAs by the web.

Although the current message might be sufficient to encourage some callers to hang up and use the website, a larger fraction of callers could be enticed to choose the Internet over the telephone if they received more specific instructions about what is available on the website and how to submit a web inquiry. Our recommendation would entail only minor changes to the hotline message, but these small changes could help callers better understand that they might be able to address their question quickly using information available on the website and they can submit a web inquiry if needed that will be answered by the same BAs who answer telephone calls.

Expanding use of IVRSs. IVRSs use recorded or dynamically generated audio to direct callers through a telephone system so they can receive service in the most efficient manner. Currently, the IVRS on the national toll-free number provides basic information about EBSA, gives the agency website, and enables callers to be routed to the field office closest to their location or to select a specific field office. OEA has attempted to use an expanded IVRS in the past without success; when it implemented the expanded IVRS, it reported that many callers became frustrated and hung up before connecting with a BA. The program also found that it was difficult to convey information about the nuances of ERISA through a recorded message and to keep it up to date as the law changed.

Despite this previous experience, we believe that the IVRS technology—and customers' familiarity with it—has advanced over the years to the point it would be worth exploring the use of an expanded IVRS. This could be implemented in a pilot period without rolling it out program-wide immediately, so that it can be tested and refined to determine what features are most effective at conveying useful information in a short period while minimizing hang ups.

A new IVRS could incorporate more sophisticated features that are currently available. For instance, before directing callers to a field office for BA assistance, the IVRS could provide the caller the option of accessing basic information through an FAQ menu. One FAQ option could address why the caller received an EOB letter with instructions to contact his or her health care plan; another could address why the caller received a PPP notice and provide simple instructions to verify eligibility for the listed benefits. This would be particularly useful for those callers without access to the Internet. To minimize caller frustration and disengagement, callers could be informed that they could choose to be connected directly to a BA at any point during the recorded message. The IVRS could incorporate speech recognition technology so that callers could state their issues and not have to go through the entire FAQ menu as well.

Beyond these relatively simple changes, customized IVRS software packages could also enable regional offices to obtain real-time statistics on telephone and web inquiries and forecast resource needs. For example, the offices could use the average time spent per inquiry to identify topics that require more BA intervention and research and perhaps identify training opportunities. This could also facilitate measurement of BA time spent actively responding to inquiries, which is not currently possible.[30]

[30] One example of integrated call center software that might be useful for the BA program to review is used by the EFAST2 contractor. This software combines both an IVRS and a management information system to log

Given that OEA has attempted to use an IVRS in the past without success, EBSA indicated in response to the draft of this report that it could better leverage resources and reach more people by keeping the website material current. If the agency decides to consider a more sophisticated IVRS in the future, it would be important to pilot the system before implementing it program-wide. To ensure its success, the program might want to pilot multiple IVRS versions to determine the most effective structure and message for conveying useful information in a short period.

Increase public awareness efforts. Finally, OEA might be able to increase overall traffic to the website through public relations efforts. Based on national office discussions, the website is consistently mentioned during outreach activities. However, OEA would like to expand use of the website while expanding its client base. OEA could potentially expand use of its website by engaging concerned stakeholders who could help the program increase traffic to the website. When the website redesign is complete, OEA could send a series of email blasts with links to its website to a master list of related federal agencies, such as the U.S. Administration on Aging (which funds the Pension Rights Center), advocacy organizations, consumer rights groups, members of Congress, and other potential stakeholders. These stakeholders could then link directly to the website from their websites. It would be quite helpful, for example, if the Pension Rights Center were to link directly from the Pension Rights Center to the EBSA Consumer Assistance page. In its response to the draft of this report, EBSA indicated plans to send this type of email blast after launching the redesigned website.

OEA should also ensure that in redesigning its website, search engine optimization is used to ensure that the website ranks highly when consumers search for pension and benefits information. At present, web searches might not rank the website highly enough for consumers to find easily.

These strategies of driving traffic to the website would be difficult to evaluate using a random assignment design. However, OEA might consider nonexperimental designs to examine changes in aggregate outcomes such as overall website use and the volume of web inquiries captured by its existing analytics. Programs often make changes to these processes and track their success by comparing, for instance, the use of the website with the old telephone message with the use of the website after the new message has been implemented. Monitoring website use over time might lead to a determination of whether website use is approaching the desired level or if more has to be done.

(continued)
inquiries, calculate the number of calls by subject area, and track time spent on responding to inquiries. There are many providers of IVRS software whose products offer a range of different features.

APPENDIX B

RELEVANT SURVEY ITEMS FROM EXISTING DATA SETS

EBRI Retirement Confidence Survey[31]

Overall, how confident are you that you (and your spouse) will have enough money to live comfortably throughout your retirement years?

- Very
- Somewhat
- Not too
- Not at all
- Don't know

I would like to know how confident you (and your spouse) are about certain aspects related to retirement.

- You will have enough money to take care of your basic expenses during your retirement
- You are doing/did a good job of preparing financially for retirement
- You will have enough money to take care of your medical expenses during your retirement
- You will have enough money to pay for long-term care should you need it during your retirement

Not including Social Security taxes or employer- provided money, have you (and/or your spouse) personally saved any money for retirement? These savings could include money you personally put into a retirement plan at work.

- Yes
- No

Are you (and/or your spouse) currently saving for retirement?

- Yes
- No

Survey of Income and Program Participation, Retirement and Pension Plan Coverage Module[32]

Now I'd like to ask about retirement plans offered on this job, not Social Security, but plans that are sponsored by your job. This includes regular pension plans as well as other kinds of retirement plans like th rift and savings plans, 401(k) and 403(b) plans, and deferred profit-sharing and stock plans…

[31] Additional information about this survey can be found at: http://www.ebri.org/surveys/rcs/2013/

[32] Additional information about this survey can be found at: http://www.census.gov/sipp

Does your job have any kind of pension or retirement plans for anyone in your company or organization?

- Yes
- No

Are you included in such a plan?

- Yes
- No

Why are you not included?

- No one in my type of job is allowed in the plan
- Don’t work enough hours, weeks, or months per year
- Haven’t worked long enough for this employer
- Started job too close to retirement date
- Too young
- Can’t afford to contribute
- Don’t want to tie up money
- Employer doesn’t contribute, or doesn’t contribute enough
- Don’t plan to be in job long enough
- Don’t need it
- Have an IRA or other pension plan coverage
- Spouse has pension plan
- Haven’t thought about it
- Some other reason

Is the plan something like a 401(k) plan, where workers contribute to the plan and their contributions are tax deferred?

- Yes
- No

Which type of plan are you in?

- Plan based on earnings and years on the job
- Individual account plan
- Cash balance plan

Do you contribute any money to this plan, for example, through payroll deductions?

- Yes
- No

If you were to leave your job now or within the next few months, could you eventually receive some benefits from this plan when you are of retirement age?

- Yes
- No

If you left your job now, could you get a lump-sum payment from this plan when you left?

- Yes
- No

How much has your job contributed to your plan within the last year?

- $ amount

I'd like to make sure about a particular type of retirement plan that allows workers to make tax deferred contributions. For example, you might choose to have your employer put part of your salary into a retirement savings account and you do not have to pay taxes on this money until you withdraw the money. These plans are called by different names, including 401(k) plans, pre-tax plans, salary reduction plans, and 403(b) plans. Does your job offer a plan like this to anyone in your company or organization?

- Yes
- No

Are you participating in this plan?

- Yes
- No

Why are you not included?

- No one in my type of job is allowed in the plan
- Don't work enough hours, weeks, or months per year
- Haven't worked long enough for this employer
- Started job too close to retirement date
- Too young
- Can't afford to contribute
- Don't want to tie up money
- Employer doesn't contribute, or doesn't contribute enough
- Don't plan to be in job long enough
- Don't need it
- Have an IRA or other pension plan coverage
- Spouse has pension plan
- Haven't thought about it
- Some other reason

Does your employer provide a matching contribution, or contribute to the plan in any other way?

- Yes
- No

Are you able to choose how any of the money in the plan is invested?

- Yes
- No

Are you able to choose how all of the money is invested, or just part of it?

- All of the money
- Part of the money

Could you withdraw the money in your retirement account now, or will you have to wait until retirement age to get the money?

- Could withdraw money now
- Must wait until retirement

Survey of Income and Program Participation, Medical Expenses/Utilization of Health Care Module[33]

During the last 12 months, about how much did [name] pay for health insurance premiums?

- $ amount

During the last 12 months, about how much was paid for his/her own medical care, including payments for hospital visits, medical providers, dentists, medicine, or medical supplies? (exclude any costs for health insurance premiums)

- $ amount

Current Population Survey Annual Social and Economic Supplement[34]

During 2011 did you/anyone in this household receive any pension or retirement income from a previous employer or union, or any other type of retirement income (other than Social Security/other than VA benefits/other than Social Security or VA benefits)?

- Yes
- No

What was the source of your income?

- Company or union pension

[33] Additional information about this survey can be found at: http://www.census.gov/sipp

[34] Additional information about this survey can be found at: http://www.census.gov/sipp

- Federal government
- U.S. military retirement
- State or local government pension
- U.S. railroad retirement
- Regular payments from annuities or paid up insurance policies
- Regular payments from IRA, KEOGH, 401(k), 403(b), and 457(b) and (f) accounts
- Other sources or don't know

Other than Social Security, did any employer or union that (name/you) worked for in 2011 have a pension or other type of retirement plan for any of its employees?

- Yes
- No

Were (name/you) included in that plan?

- Yes
- No

At any time in 2011, (was/were) (you/anyone in this household) covered by a health insurance plan provided through (their/your) current or former employer or union?

- Yes
- No

Did (name's/your) former or current employer or union pay for all, part, or none of the health insurance premium?

- All
- Part
- None

During 2011, about how much did (name/you) pay for health insurance premiums for (yourself/himself/herself) or others in the household, after any reimbursements? Please include premiums paid for HMOs, Fee for Service Plans, Commercial Medicare Supplements, or other special purpose plans, such as vision or dental plans. Include prescription drug insurance such as Medicare Part D premiums, and Medicare Advantage premiums. DO NOT include Medicare Part B premiums.

- $ amount

During 2011, about how much was paid for (name/you) for over-the-counter health-related products such as aspirin, cold remedies, bandages, first aid supplies, and other items?

Include any amount paid on (your/his/her) behalf by anyone in this household that was not reimbursed.

- $ amount

Aside from over-the-counter items, during 2011, about how much was paid for (name's/your) own medical care, including payments and co-payments for hospital visits, medical providers, dental services, prescription medicine, vision aids, and medical supplies? Include any amount paid on (your/his/her) behalf by anyone in this household that was not reimbursed.

- $ amount

General Social Survey[35]

I am going to read a list of fringe benefits that workers sometimes get in addition to their wages. Whether you receive it or not, please tell me whether you are eligible to receive each fringe benefit.

- Medical or hospital insurance?
- Dental care benefits?
- Life insurance?
- Sick leave with full pay?
- Maternity or paternity leave with full re-employment rights?
- Flexible hours or flextime scheduling?
- Cash or stock bonuses for performance or merit?
- A pension or retirement program?
- Information about child care services in the community?
- Assistance with the costs of day care for children?

Are you, yourself, covered by health insurance, a government plan like Medicare or Medicaid, or some other plan that pays for your medical care?

- Yes
- No

Were you ever denied (mental health) services under your plan's benefit package?

- Yes
- No

[35] Additional information about this survey can be found at: http://www3.norc.org/GSS+Website/Browse+GSS+Variables/

APPENDIX C

ASSUMPTIONS AND SENSITIVITY CHECKS FOR POWER CALCULATIONS

This page has been left blank for double–sided copying.

This appendix provides the specific set of assumptions used in the calculation of MDIs for the classic RCT presented in Chapter II, the evaluation of web referral service delivery models presented in Chapter III and the evaluation of the prioritization service delivery models discussed in Appendix D. Because a follow-up survey would have to be used to collect information on service receipt and outcomes for both study groups in the evaluation of web referral models presented in Chapter III, the MDIs for this option are the same as for the classic RCT presented in Chapter II.

All power calculations assume:

- Significance level of alpha = 0.05
- Two-tailed test
- Power of 0.80
- Survey response rate of 80 percent
- Outcomes:

3. The first outcome type is binary variables. We have information on four binary variables of interest, but this discussion can be generalized to any binary variable that has a similar mean value because of the properties of binary variables. Gallup fields a survey that has resulted in the following data:
    - 69 percent of respondents somewhat or strongly agreed in 2012 that they were satisfied with the services they had received through EBSA (this outcome is only relevant for the relative impact design discussed in Chapter III)
    - 70 percent of respondents somewhat or strongly agreed in 2012 that they felt better informed to protect their benefits after interacting with the program
    - 64 percent of respondents somewhat or strongly agreed in 2012 that they felt more secure about their benefits after interacting with the program
    - 76 percent of respondents indicated that their knowledge level was much higher or somewhat higher after interacting with the program
4. The second outcome is monetary recoveries. Although only a small fraction of calls result in monetary recoveries, they are considered to be an important outcome for the program. The BAs also spend more time on the types of inquiries that result in monetary recoveries.
    - OEA provided data on the total amount of the recovery and the number of participants covered by that dollar amount for the first three quarters of FY 2013. Because this is an amount calculated at a single point in time, and not an aggregate measure that increases over time, it is not necessary to inflate this to represent a year’s worth of data, as it would be if we were examining an outcome such as annual earnings based on three quarters of data. We used the data to compute a recovery per person. We also filled out the data set with zeroes for those inquiries that did not result in any

recovery, then computed the standard deviation of that number. The mean recovery per person was $814 and the standard deviation was $11,267.

- We assumed an individual-level R-squared of 0 because we have no information on how much of the variance individual-level variables would be able to explain. Most likely, including individual-level variables would be able to explain some portion of the variance, which would increase power somewhat and decrease the needed sample size to achieve a given MDI.

In addition to these assumptions, we made other assumptions specific to the designs being considered:

- *Classic RCT and evaluation of web referral models: randomization at the individual level.* We considered three target sample sizes: 25,000, 15,000, and 10,000.
  - For these designs, we cannot be sure what knowledge about their benefits members of the control or services as usual group are likely to acquire on their own. Given that the responses to binary variables ranged from 64 percent to 76 percent in the Gallup survey, we used an assumption that 65 percent of control group members will somewhat or strongly agree that they felt better about these outcomes (and derive the corresponding standard deviation as the square root of 0.65*(1-0.65), which is 0.48). Because of the statistical properties of binary variables, the MDIs are not highly sensitive to differences in a few percentage points in the outcome standard deviation assumed.
  - Similarly, we do not know what to expect about the dollar amount of benefit recoveries or the likely standard deviation of this outcome for the control or services as usual groups. This is especially challenging for these designs because the evaluation would need to collect this information in a very different way than the program currently collects it, and this would also affect the standard deviation, though it is not clear how. Because power decreases with the standard deviation, a conservative approach is to use the standard deviation calculated using the data provided by OEA, rather than shrinking it to account for the fact that control or services as usual group members would likely have lower means and standard deviations of monetary recoveries.
  - We tried two options for the treatment to control ratio—50:50 and 75:25—without increasing the total sample size. However, because half the number of sample members would be assigned to the control or services as usual group under this higher ratio, the MDIs are much larger, and exceed the magnitude that would likely be detectable for monetary recoveries. The power calculations for this scenario are available upon request.

- The MDI formula used for the benefit recovery calculations is as follows:

  $$factor \times \sigma \times \sqrt{\frac{1}{p(1-p)}} \times \sqrt{(1-R_{ind}^2)\left(\frac{1}{r \times N}\right)}$$

  where *factor* is 2.8; *σ* is the standard deviation of benefit recoveries ($11,267) or binary variables (0.48) based on data provided by the program as described above; *p* is the control group sampling rate (0.50); R^2 at the individual level is 0, as discussed above; *r* is the response rate (0.80) for the follow-up survey; and *N* is the total sample size (not the number of completed follow-up surveys). The MDI calculations assume 80 percent power, two-tailed tests, and a significance level of alpha = 0.05. No arcsine adjustment for the binary variables was made; making this adjustment has a trivial effect on the computed MDIs.

- *Evaluation of the prioritization model.* Under this design, there would be a couple of key features of the design that would have implications for the power calculations. First, administrative data on both study groups would be available. Second, a clustered design would be feasible.

  - For this design, we would have administrative data on both groups for monetary recoveries because both would receive services, albeit in different ways. This means the sample sizes for this outcome can be much larger than under either the classic RCT or the evaluation of web referral models. We used the same mean and standard deviations of the monetary recoveries as in the classic RCT for the power calculations, but the difference between the MDIs under that design and this one arises from the much larger sample size available for an evaluation of the prioritization model and the fact that the coverage rate for the TAIS data would be greater than for a follow-up survey; we assumed a TAIS coverage rate of 98 percent.

  - However, data for the binary outcomes would still need to be gathered using a follow-up survey. Therefore, the assumed sample sizes for the classic RCT apply to the evaluation of the prioritization model as well for these variables.

- *Evaluation of the prioritization model with random assignment at the individual level.* The power is greater under this design than under a design in which regional offices are randomly assigned. We computed a couple of sample size scenarios:

  - Sample sizes of 100,000, 50,000, 25,000, 15,000, and 10,000 for the monetary recoveries outcome. The sample sizes for the binary outcomes are the same as under the classic RCT and the evaluation of web referral models.

  - Treatment-control ratios of 50:50 and 75:25. As discussed above, there is not much to be gained by using a 75:25 ratio, especially when everyone is receiving services. The results of these power calculations are available on request.

- *Evaluation of the prioritization model with random assignment at the office level.* There are 14 total offices that handle inquiries from callers, plus the national office. We excluded the national office because it handles special inquiries that are typically referred to it by other offices or come about as a result of a Congressional inquiry. Three offices are satellites to larger regional offices and have the same supervision and processes, so we folded the data on those offices into their larger regional offices. This leaves 10 offices.
  - We used the same sample size scenarios as discussed immediately above, but in addition, for any clustered design, we need information on the intra-class correlation (ICC), or the extent to which outcomes are similar for people in the same territory. Unfortunately, we were not able to obtain respondent-level data on the binary outcomes of interest, which are held by Gallup. Therefore it was not possible to calculate ICCs for those outcomes. We used the data provided by the program on monetary recoveries to estimate the ICC for monetary recoveries. The estimated ICC was 0.00235, which is fairly low. We applied this to both the monetary recovery and binary variables. Because the MDI calculations are highly sensitive to the ICC, and because we are not sure about the extent to which the ICC from the monetary outcomes can be applied to the binary outcomes, we did sensitivity tests varying the ICC to 0.02 and 0.04 (presented below). The ICCs assumed have significant impacts on the power calculations. In general, the larger the assumed ICC, the larger the MDI for a given sample size (or, alternatively, the larger the sample size needs to be for a given MDI).
  - For any clustered design, we also need to account for the cluster-level R-squared, or the portion of the variance at the cluster level that can be explained by including additional cluster-level variables in the impact analysis. OEA provided data at the office level for the number and amount of benefit recoveries by office for the past three years; we used data in 2012 and 2011 to predict the recoveries per person in 2013. We tried a couple of specifications (i.e., total inquiry volume, total number of recoveries, total number of zero recoveries, total number of non-zero recoveries) at the cluster level. The R-squared was consistently in the .70 to .95 range. Therefore, we used an R-squared of 0.80.
  - For the binary outcomes, we had only two years of data for each of the questions of interest. We conducted similar analyses to those described for monetary recoveries, and the R-squared on those ranged from 0.4 to 0.6. Therefore, we used a cluster-level R-squared of 0.5.
  - The MDI formula used for the calculations in the clustered design is as follows:

$$factor \times \sigma \times \sqrt{\frac{1}{p(1-p)}} \times \sqrt{\frac{(1-\rho)(1-R_{ind}^2)}{r \times N} + \frac{\rho(1-R_{site}^2)}{\# sites}}$$

  where factor is 2.8; σ is the standard deviation of benefit recoveries ($11,267) or binary variables (0.48) based on data provided by the program as described above; p is the control group sampling rate (0.50);

r is the response rate (0.80 for the survey, 0.98 for TAIS records); *N* is the total sample size (not the number of completed surveys); the ICC ρ is 0.002, 0.02, or 0.04; R^2 is 0 at the individual level and 0.80 at the site level; and *#sites* is the total number of sites selected in the approach being considered. The MDI calculations assume 80 percent power, two-tailed tests, and a significance level of alpha = 0.05. No arcsine adjustment for the binary variables was made; making this adjustment has a trivial effect on the computed MDIs.

Table C.1. MDIs for Clustered Design Under an Assumption of ICC = 0.02

	Clustered Designs				
	(1)	(2)	(3)	(4)	(5)
MDIs—Binary Variables					
Sample Size			25K Sampled, 20K Complete, 10 Offices	15K Sampled, 12K Complete, 10 Offices	10K Sampled, 8K Complete, 10 Offices
Overall			**8.65%**	**8.78%**	**8.95%**
50% subgroup			8.90%	9.10%	9.40%
25% subgroup			9.20%	9.70%	10.30%
10% subgroup			10.30%	11.40%	12.60%
MDIs—Recoveries					
Sample Size	100K Participants 10 Offices	50K Participants 10 Offices	25K Participants 10 Offices	15K Participants 10 Offices	10K Participants 10 Offices
Overall	**$1,247**	**$1,262**	**$1,292**	**$1,331**	**$1,377**
50% subgroup	$1,262	$1,292	$1,350	$1,423	$1,509
25% subgroup	$1,292	$1,350	$1,458	$1,591	$1,742
10% subgroup	$1,377	$1,509	$1,742	$2,012	$2,305

Note: Binary variables include customer satisfaction and self-reported knowledge of benefits rights, ability to advocate on one's behalf, access to benefits-related documents, and perceptions of a secure retirement and health. MDIs for binary variables are expressed in percentage points. MDIs for recoveries are expressed in dollars. Sample sizes for recoveries indicate the number of participants' whose TAIS records would be extracted for the evaluation. We assume a 98 percent coverage rate for these data elements, so actual sample sizes would be slightly lower than indicated in the table.

Table C.2. MDIs for Clustered Design Under an Assumption of ICC = 0.04

	Option 2B (1)	Option 2B (2)	Option 2B (3)	Option 2B (4)	Option 2B (5)
MDIs—Binary Variables					
Sample Size			25K Sampled, 20K Complete, 10 Offices	15K Sampled, 12K Complete, 10 Offices	10K Sampled, 8K Complete, 10 Offices
Overall			**12.09%**	**12.18%**	**12.30%**
50% subgroup			12.20%	12.40%	12.60%
25% subgroup			12.50%	12.90%	13.30%
10% subgroup			13.30%	14.10%	15.10%
MDIs—Recoveries					
Sample Size	100K Participants 10 Offices	50K Participants 10 Offices	25K Participants 10 Offices	15K Participants 10 Offices	10K Participants 10 Offices
Overall	**$1,753**	**$1,764**	**$1,784**	**$1,812**	**$1,846**
50% subgroup	$1,764	$1,784	$1,826	$1,879	$1,944
25% subgroup	$1,784	$1,826	$1,905	$2,007	$2,127
10% subgroup	$1,846	$1,944	$2,127	$2,348	$2,599

Note: Binary variables include customer satisfaction and self-reported knowledge of benefits rights, ability to advocate on one's behalf, access to benefits-related documents, and perceptions of a secure retirement and health. MDIs for binary variables are expressed in percentage points. MDIs for recoveries are expressed in dollars. Sample sizes for recoveries indicate the number of participants' whose TAIS records would be extracted for the evaluation. We assume a 98 percent coverage rate for these data elements, so actual sample sizes would be slightly lower than indicated in the table.

APPENDIX D

OTHER SERVICE DELIVERY MODELS THAT COULD BE TESTED IN AN IMPACT EVALUATION

This page has been left blank for double–sided copying.

In this appendix, we discuss two alternative service delivery models that could be tested with a relative impact evaluation design similar to that described in Chapter II. The two models presented are: (1) a model that would use junior BAs or receptionist staff to first prioritize calls as they were received, answering simple informational requests and providing referrals as needed, but referring more complicated requests and emergencies directly to a BA; and (2) a model with this same type of prioritization plus specialization of BAs into subject matters. These service delivery models were developed as a result of discussions with OEA, but after they were more fully fleshed out in the draft design report, EBSA and CEO decided not to pursue them at this time because of potential implementation difficulties. They also decided that, in light of the expected increase in demand for BA services in 2014 upon implementation of the Affordable Care Act (ACA), they would prefer to focus a potential evaluation on alternative service delivery models that could assess the impact of using the EBSA website and web inquiry system.

Although EBSA and CEO have decided in the short run not to implement an evaluation testing either of the prioritization models described here, a discussion of an evaluation of these models is included in this appendix to serve as a reference if EBSA and CEO decide to revisit them in the future. We describe how this type of evaluation would be implemented, discuss data collection needs, review sample sizes and MDIs, and discuss the importance of implementation and cost studies for this option. We conclude the appendix with a review of potential challenges with this design and recommended solutions.

A. Prioritization Service Delivery Models

Through interviews with regional office staff and discussions with CEO and EBSA, we identified two potential alternative service delivery models that involve prioritizing calls before assigning them to a BA. Both of these service delivery models could be tested with a differential impact evaluation design similar to that discussed in Chapter II. These alternative models include:

1. Using junior BAs to prioritize calls
2. Using junior BAs to prioritize calls combined with senior BAs who specialize by topic area

Using junior BAs to prioritize calls. Responding to telephone calls is one of the primary activities that BAs engage in on a daily basis. EBSA’s customer service standards require BAs to answer calls live whenever possible, and offices are expected to have procedures in place to support this goal. The specific structure of the phone answering process varies somewhat across offices, with most using half-day telephone shifts for BAs to respond live to phone inquiries, and others using full-day telephone shifts. Slightly more than half of the offices route calls directly to BAs, while the other offices route calls first to a receptionist who then transfers calls to BAs. None of the offices use receptionists to answer benefits-related questions from inquirers. Bellotti et al. (2013) provides more details on variations in process across regional offices.

Under this alternative model, junior BAs would be trained to take incoming calls, respond to straightforward informational requests, and provide referrals to other agencies when needed.[36] If the junior BA determined that the caller had an in-depth question, needed informal intervention, or the junior BA could not quickly assess the issue about which the inquirer was calling, the junior BA would forward the call to the BA "hunt group" or queue system to handle in the normal manner.

Junior BAs would be provided with adequate training and materials to handle these tasks. They would be given a scripted set of questions and trained to use this at the beginning of each call to identify the caller's immediate issue. They would also be trained to ask a series of questions at the end of each call about any other benefits-related questions to ensure that they had not missed any issues that should be forwarded to a BA. The evaluation team could work with OEA to create easy-to-follow written conversation guides and lists of referral phone numbers that the junior BAs could have on hand to guide them through these calls. These procedures could be pilot tested before a formal evaluation was conducted to refine the training materials, identify any unanticipated issues, and determine whether indeed the model was worth testing. If the model was implemented, the evaluation staff, Senior Benefits Advisors (SBAs), or lead BAs could conduct periodic oversight to ensure that junior BAs were handling inquirers' questions appropriately and forwarding calls to BAs when necessary.

This alternative service delivery model has the potential to allow more senior BAs to spend more of their time on the types of complex issues that require their detailed knowledge about benefits rights, and free up time for them to do more outreach or serve more inquirers who need their in-depth assistance. Based on feedback from the BAs, this model could possibly also increase BA job satisfaction and help develop a longer or more meaningful career progression that reduces staff turnover. BAs currently spend a considerable amount of time responding to callers who reached the program in error, require resource assistance resulting in a simple referral, or have a simple question that does not require the assistance of a highly experienced BA. Approximately 11 percent of phone inquiries are from these types of callers. Many of these callers contact EBSA because the hotline number is listed on explanation of benefits (EOB) forms from health insurance companies. These callers do not typically need the assistance of a highly experienced BA to be redirected to the appropriate organization. Instead, junior staff could be trained to identify these types of calls and provide the referral information necessary. Although data on call lengths are not systematically collected, we were told by BAs in interviews that between 75 and 90 percent of phone inquiries can be handled in 5 to 10 minutes. These calls tend to require a simple explanation of the inquirers' rights, direction to written documentation on the EBSA website, or quick responses about COBRA notices. Some of these calls will still need to be handled by BAs, but with training and written guidance, junior staff might be able to

[36] BAs vary widely in their experience and educational backgrounds, and they range from federal pay grade 7 to pay grade 12. OEA could choose to use more junior BAs (e.g., pay grades 7-8) to prioritize calls and respond to less complex questions. This division of labor among BAs is currently done informally at some offices, but an evaluation would formalize these arrangements. This would not result in a demotion for any current BAs; all BAs would maintain their current pay grade, but their job duties might shift depending on their level of experience. Alternatively, OEA could choose to hire staff at a more junior level than current BAs (e.g., pay grades 5-6) to serve as junior BAs. In addition to assisting the study, this could serve as a valuable training opportunity for new or junior staff.

handle many of these types of quick resolution calls and identify ones that should be forwarded to BAs.

For this alternative service delivery model, a relative impact evaluation would answer the question:

- What is the impact of using junior BAs to prioritize phone inquiries on participants' knowledge of and access to their entitled pension/health benefits when compared to receiving services as they are currently delivered?

Given the substantial proportion of calls that require less in-depth assistance and could potentially be addressed by junior staff, we hypothesize that altering the delivery of BA assistance in this way will not harm participant outcomes. If BAs have more time to focus on complex issues related to access to benefits, access to documents, and recovery of benefits, participant outcomes may actually improve under the alternative service delivery model.

Beyond this general research question, this type of study could also look at the effect of this alternative service delivery model on other outcomes such as customer satisfaction, BA job satisfaction, and cost. For example, BAs reported that inserting another layer of staff could potentially reduce customer satisfaction if a portion of callers had to be transferred to a senior BA before their issue was resolved. However, this type of change might be acceptable to OEA if there were other positive results to implementing the model, such as increased job satisfaction among BAs or reduced cost per inquiry. These types of issues could be examined through an implementation study that would run side-by-side with an impact evaluation to examine how each of the models unfolds in practice.

One potential challenge associated with using junior BAs to prioritize calls is that there may be budget and staffing constraints in hiring new staff to serve as junior BAs. To overcome this challenge, the evaluation could possibly fund and hire temporary contractors to serve as junior BAs during the course of a study. This would reduce the recruitment and hiring burden for OEA and could help circumvent some of the staffing constraints in place for hiring full-time staff. If the evaluation generated evidence supporting the implementation of this alternative service delivery model, OEA could possibly get more support for hiring permanent staff to fill these roles in the future.

Another potential challenge is the need to ensure that changes in working conditions are respectful of the BA union contract. It would be important to implement this alternative service delivery model in close consultation with the BA union to ensure that it was done in a way that did not change the pay grades or job duties of current BAs. Hiring contractors to serve as junior BAs during the course of the study might be one approach to gaining union support for this alternative service delivery model because it would not change the work that BAs do and contracted staff would serve a very distinct role related to the evaluation.

Using junior BAs to prioritize calls combined with senior BAs who specialize by topic area. The prioritization model described above could be further enhanced by combining it with senior BA specialization by topic. Under the current system, calls are routed as they are received using a BA hunt group. In most offices, each inquiry, regardless of the topic, is routed to whichever BA is available and next in the queue. Even though some BAs reported in interviews that they have prior work experience or an interest in specific benefits areas—for example, some have worked for the IRS or health insurance companies—they have to be prepared to answer all types of benefits-related questions.

An alternative to the current process for distributing calls to available BAs would be to develop some BAs as experts in specific areas and route relevant calls to those specialists. For example, based on experience, knowledge, and interest, some BAs might specialize in questions related to COBRA, others on pensions, and others on the ACA. All BAs would still be trained to ensure basic knowledge of all major topics and would be prepared to answer any type of inquiry, but they could have the option to specialize in specific areas. When a junior BA determined that an inquiry required more specialized assistance, the junior BA would forward the caller to the specialist BA in that topic area. If no specialist were available, the caller could be given the option to leave a message or be forwarded to another available BA for assistance. In current practice, OEA reported that some BAs do already specialize, and their colleagues can request their assistance with specific inquiries or transfer an inquiry to the specialist directly if needed. However, there is no formal system in place for this process.

In theory, this alternative model might result in inquirers receiving more useful or detailed assistance as BAs build up extensive knowledge about a specific area rather than serving as generalists across all areas. It might also reduce the amount of research and call-backs that are necessary, as the specialist BAs would be more likely to answer the callers' questions immediately because they are more familiar with the topics. In addition, allowing BAs to specialize might lead to more satisfaction with their work because they will feel they have more control over the topics that they work on and their specific knowledge and experience are used to the fullest extent possible.

An evaluation of this alternative service delivery model would be designed to answer the question:

- What is the impact of combining caller prioritization by junior BAs with use of BAs as topical specialists on participants' knowledge of and access to their entitled pension/health benefits, compared to receiving services as usual?

We hypothesize that this alternative service delivery model would improve participant outcomes relative to the current model because it has the potential to lead to more comprehensive customer service. If BAs enjoy increased job satisfaction, this might also feed into improvements in customer satisfaction and participant outcomes as well as cost savings due to reduced BA staff turnover coupled with use of more junior staff for prioritization.

Beyond the challenges mentioned in the prioritization only model, another potential challenge in implementing this model is that it could result in an uneven workload for BAs who specialize in relatively common or rare issues. In addition, some regional offices might not have BAs who are interested or have the background to specialize in certain areas. This model might also increase waiting times for inquiries about common topics, potentially leading to callers becoming disengaged or dissatisfied. To avoid these issues, a flexible model could be instituted in which inquiries are routed to specialists only if they are available. If a specialist is not available, the call could be routed to another available BA who is trained to answer calls across all topic areas.

A second potential challenge is that BAs perceive that callers cannot always articulate the nature of their benefits-related question, and at times inquirers' most pressing issues are not clear until after a lengthy conversation. Under the specialist model, inquirers who cannot clearly articulate the nature of their issue might be forwarded to the wrong specialist. To address this concern, BAs would be trained, as they currently are, to question callers about all of their issues and to provide assistance on all of the major topics that arise. Inquirers could initially be transferred to a specialist BA based on the main issue that they raise at the beginning of their call, and the specialist BA could assist them with that issue and any others that arise in the course of their conversation. If another issue arises that another available BA specializes in, the initial BA could choose to consult with or forward the call to the other specialist.

Finally, the evaluation would have to consider that some offices may already implement informal practices that result in BA specialization. Depending on the mix of BA experience and background, some offices may already informally route calls to BAs with specific experience in the topic area. In addition, offices currently route very complex calls to the national office. To successfully test this type of model, the evaluator would have to work with OEA to more fully understand the current practices in all offices and how an alternative could be designed to be both feasible and sufficiently different from the current model, as well as structured to provide information of value to OEA and policymakers.

B. Implementation of the Design

If OEA chooses to test one of the prioritization service delivery models described above, an RCT could be implemented to evaluate the impact of the new model relative to the current service delivery model. Our recommendation would be to conduct individual-level random assignment within all regional offices, similar to the evaluation design described in Chapter II. Individuals seeking BA assistance would be randomly assigned to receive one of two treatments: BA assistance as usual or the alternative model of providing assistance. This would require each regional office to operate both treatment models simultaneously.

A clustered design, in which regional offices rather than individual callers are randomly assigned to different treatments, could also be considered; in fact, it would likely be easier to implement than one in which individuals would be randomly assigned. In this case, one set of offices would continue operating as usual while the other set of offices would implement the new service delivery strategy. This approach would make it easier to train staff, implement the service delivery models, conduct random assignment, and make sure that individuals maintain their study groups. However, this design has significant drawbacks in terms of the ability of the evaluation to detect impacts, which is discussed in the section on sample sizes below. A clustered design also raises concerns about the generalizability of results, given that what works in one office or set of offices may not work in others.

1. What Would the Alternative Service Delivery Model Be Compared to?

In Chapter II, we described a relative impact evaluation design in which both study groups are assigned to different types of treatment. If OEA chose to evaluate either of the prioritization service delivery models described in this appendix, a similar relative impact evaluation design could be used. The prioritization service delivery model tested would be compared to "business as usual"—that is, normal BA assistance, as it is currently being delivered. Both groups would have full access to all types of BA assistance, albeit delivered in different ways. This type of evaluation design would allow OEA to learn about the impact of a new model of service delivery compared to the current model.

2. How Would Random Assignment Work?

The actual process of randomly assigning participants would vary based on whether the evaluation used an unclustered or clustered design. For a clustered design, the random assignment process is straightforward because regional offices would be randomly assigned to either continue conducting services as usual or implement the alternative service delivery approach. Regional offices that are similar in terms of the number and type of inquiries and average level of BA staff experience could be paired and then one office in each pair randomly would be assigned to the treatment group and the other to the control group.

If an unclustered design is selected, each participant would be randomly assigned to either receive services delivered in the usual manner or to receive services delivered in the alternate model. The evaluator and OEA could explore the potential for using an automated phone system to randomly assign callers to one treatment or another before the call is answered by a BA or receptionist. Appropriate technology would be needed to automatically route incoming calls to the appropriate service delivery model and record the group to which the model was assigned. If it is not possible or determined undesirable to automatically randomize calls to study groups using the phone system, random assignment could be done by a staff member or contractor when the call is received. It may be preferable to have random assignment handled by a contractor so that BAs do not have to take time away from delivering assistance for study activities, and to ensure consistency and fidelity across offices.

Whether individual-level randomization is conducted by an automated system or by a person, it would likely not be necessary to explain the study design to callers prior to randomization. No callers would be denied services or offered a clearly less effective service, so the risk to callers of participating in the study is minimal. Programs routinely adjust their services over time to explore or try new service delivery options, and no consent is needed. We do, however, recommend that program staff members who are in contact with callers inform them that the program is participating in a research study and that any information they provide might be used for research purposes. They would also indicate that no personally identifiable information would be reported. Informing the customers in this way would reduce the chances that they would be surprised to be contacted later for follow-up data collection, and protect the program and the research team from liability. To ensure that this is an appropriate strategy, the study design could be submitted to an IRB to independently assess compliance with ethics related to research with human subjects.

If the IRB, OMB, or OEA decides that a consent process is needed, the evaluator will need to implement a short verbal consent process similar to that described in Chapter III. This consent process would have to occur prior to randomization, so an automated random assignment system could not be used in this case. Instead, the staff member or contractor who initially answered the phone call would describe the study and random assignment process to the caller, ask for verbal consent, and then randomly assign the caller to one of the two treatment groups. This process could be structured to take about three minutes. OEA and CEO would have to decide if and how to provide assistance to callers that did not consent to participate in the study. Denying services to these callers would incentivize participation, but would conflict with EBSA’s goal of providing assistance to all inquirers. The evaluation team would need to work with DOL to determine the best approach.

C. Data Collection Needs

An evaluation of a prioritization service delivery model would require similar types of data as the evaluation designs described in Chapter II:

- **Baseline data.** Some baseline data in addition to what is currently collected by the program would need to be collected to assess the similarity of the two treatment groups and to define subgroups. As mentioned above, the evaluation could try to limit, to the extent possible, the number of variables collected to reduce burden.
- **Service receipt data.** To track the amount of assistance received by each inquirer, the evaluation would need to examine service receipt data. In contrast to the evaluation designs described in Chapters II and III, service receipt data collection for the prioritization models could be done entirely through TAIS, rather than a follow-up survey. This is because all study participants would receive some form of BA phone assistance and therefore have service data captured in TAIS.
- **Outcome data.** To estimate the impact of the prioritization service delivery models, the evaluation would need data on the primary outcomes of interest—knowledge about benefits, access to benefits, and customer satisfaction—for all inquirers.[37] All study participants would have some outcomes, such as access to benefits and documents, recorded in TAIS. Other outcomes, such as knowledge about benefits rights and ability to advocate on one's own behalf, would need to be collected through a follow-up survey.

D. Evaluation Sample Sizes and Minimum Detectable Impacts

The availability of TAIS data on all study participants for some of the outcomes in an evaluation of these alternate service delivery models has considerable implications for the sample sizes of a potential evaluation. Because all study participants would have TAIS data on benefits-related documents and recoveries, a much larger sample size would be available for analysis on those outcomes than if they had to be measured using a follow-up survey. This, in turn, implies that it would be possible to detect very small impacts on those outcomes captured in TAIS.

As mentioned above, when choosing a target MDI, the study must take into account the size of the likely difference in impacts between the two treatments. As the services become more similar, the likely difference in impacts becomes smaller. If the goal were to show which service delivery strategy were better at improving outcomes, then the target MDI would have to be very small. However, that might not necessarily be the goal for an evaluation comparing two service delivery models; rather, the goal could be to simply determine whether the two service delivery strategies provide similar levels of outcomes. For instance, if it were acceptable that the impact

[37] For more details on these data elements, see the Data Collection Needs section in Chapter II.

of the alternative service strategy were less than a three percentage point difference in knowledge, then the study could be powered to detect a three percentage point or larger difference. If the study did not find a statistically significant impact, then it could be concluded that the alternate service delivery strategy did not cause an important difference in outcomes between the two groups. If one of the strategies were less costly than the other to implement or provided other benefits such as higher staff satisfaction and reduced staff turnover, this might argue for adopting one of the strategies more broadly. The evaluation team would need to work with OEA and CEO to determine the appropriate target MDI.

MDI calculations. For the prioritization service delivery models, we considered the study's statistical power under two designs: an unclustered design involving random assignment of individuals and a clustered design involving random assignment of regional offices. The assumptions for the unclustered design are the same as those described in Chapter II, with one exception:

- ***Availability of larger sample sizes using administrative data.*** For the outcome recovery amount per participant, we computed MDIs for much larger sample sizes because of the availability of TAIS data on both study groups.[38] The sample sizes are 100,000; 50,000; 25,000; 15,000; and 10,000 participants (the last three sample sizes can be compared to the MDIs in Chapters II and III). Because this outcome would be examined using administrative data, we assumed a coverage rate of 98 percent. This higher coverage rate decreases the MDIs relative to what they would be using an 80 percent survey response rate for the same sample sizes.

In addition to the factors related to the statistical power mentioned in Chapter II, when sample members are clustered within the same regional office territory, they are likely to face similar economic conditions and have similar demographic characteristics; therefore, their outcomes are likely to be more similar to one another than to sample members in other regional office territories. This correlation in outcomes—known as the intra-cluster correlation (ICC)—within regional offices would increase the variances of the impact estimates relative to those from a simple random sample of the same number of individuals. Thus, a larger sample size is needed when sample members are clustered to achieve the same MDI as with a smaller sample size using a random sample of individuals. Therefore, some additional assumptions were needed to compute MDIs for a clustered design:

- ***Assumed number of offices.*** There are 14 total offices that handle inquiries from callers, plus the national office. We excluded the national office from consideration in this design because it handles special inquiries that are typically referred to it by other offices or come about as a result of a Congressional inquiry. Three offices are satellites to larger regional offices and have the same supervision and processes, so

[38] Even though data on this outcome would be available for both groups, we still transformed them in the same way as for the MDIs in Chapters II and III; namely, by computing a recovery per participant. This enables comparison of the MDIs to those presented for the other study designs in the report. However, if OEA chose to evaluate a prioritization model, the evaluator could also compute the MDIs without making this transformation.

we folded the data on those offices into their larger regional offices. This leaves 10 regional offices.

- ***Assumptions for intra-cluster correlation.*** We used OEA data on monetary recoveries from TAIS for the first three quarters of FY 2013 to compute an ICC of 0.002. Unfortunately, we were not able to obtain respondent-level data on the binary outcomes of interest, which are held by Gallup. Therefore, it was not possible to calculate ICCs for those outcomes. It is not clear to what extent the ICC from the monetary recoveries outcome would apply to the binary variables, but it was the only available estimate so we used it for the binary outcomes as well. Appendix C shows sensitivity tests increasing the ICC to 0.02 and 0.04. These have large implications for the MDIs; the larger the ICC, the larger the MDI. We applied this to both the monetary recovery and binary variables.

- ***Assumptions for the cluster-level R-squared for binary outcomes.*** For the binary outcomes, we had only two years of data, so we used the data in 2011 to predict the outcomes in 2012. The R-squared on those ranged from 0.4 to 0.6. Therefore, we assumed a cluster-level R-squared of 0.5.

- ***Assumptions for the cluster-level R-squared for monetary recoveries.*** For any clustered design, we also need to account for the cluster-level R-squared, or the portion of the variance at the cluster level that can be explained by including additional cluster-level variables in the impact analysis. OEA provided data at the office level for the number and amount of benefit recoveries by office for the past three years; we used data in 2012 and 2011 to predict the outcomes in 2013. We tried several model specifications (that is, total inquiry volume, total number of recoveries, total number of zero recoveries, and total number of non-zero recoveries) at the cluster level. The R-squared was consistently in the .70-.95 range. Therefore, we assumed an R-squared of 0.80.

Table D.1 presents the MDIs for both unclustered and clustered designs. The MDIs for binary variables under the unclustered design—random assignment of individuals to the current service model versus an alternative model—assume the same number of survey participants as in Chapters II and III and therefore they are the same. However, for recoveries under the unclustered design, the sample size could be much larger than under any of the designs discussed in Chapters II and III because of the availability of TAIS data for a larger sample. In addition, the proportion of the sample covered would be greater using TAIS (about 98 percent) compared to a survey (for which we typically assume an 80 percent response rate). Therefore, the MDIs for monetary recoveries are slightly lower than in the classic RCT.

Using a much larger administrative sample of 100,000 participants and an unclustered design, the study would be powered to detect a difference as small as $197 per participant in monetary recoveries. This is about a quarter of the mean recovery amount per participant of $814. This presents a substantial advantage over the classic RCT and the evaluations of web referral models in terms of statistical power.

Table D.1. MDIs for the Unclustered and Clustered Designs of Prioritization Service Delivery Models

	Unclustered Design					Clustered Design				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
MDIs—Binary Outcomes										
Sample Size			25K Sampled, 20K Complete	15K Sampled, 12K Complete	10K Sampled, 8K Complete			25K Sampled, 20K Complete, 10 Offices	15K Sampled, 12K Complete 10 Offices	10K Sampled, 8K Complete 10 Offices
Overall			**1.89%**	**2.44%**	**2.99%**			**3.27%**	**3.61%**	**4.00%**
50% subgroup			2.67%	3.45%	4.22%			3.80%	4.40%	5.00%
25% subgroup			3.78%	4.88%	5.97%			4.60%	5.60%	6.50%
10% subgroup			5.97%	7.71%	9.44%			6.50%	8.20%	9.80%
MDIs—Recoveries										
Sample Size	100K Participants	50K Participants	25K Participants	15K Participants	10K Participants	100K Participants 10 Offices	50K Participants 10 Offices	25K Participants 10 Offices	15K Participants 10 Offices	10K Participants 10 Offices
Overall	**$197**	**$278**	**$394**	**$508**	**$622**	**$436**	**$479**	**$553**	**$640**	**$734**
50% subgroup	$278	$394	$557	$719	$880	$479	$553	$679	$817	$962
25% subgroup	$394	$557	$787	$1,016	$1,245	$553	$679	$878	$1,087	$1,303
10% subgroup	$622	$880	$1,245	$1,607	$1,968	$734	$962	$1,303	$1,652	$2,004

Notes: Binary variables include customer satisfaction and self-reported knowledge of benefits rights, ability to advocate on one's behalf, access to benefits-related documents, and perceptions of a secure retirement and health. MDIs for binary variables are expressed in percentage points. MDIs for recoveries are expressed in dollars. The unclustered design involves randomly assigning individuals, whereas the clustered design involves randomly assigning regional offices to treatment groups. Assumes ICC of 0.002 for the clustered design. See Appendix C for a description of the full set of assumptions used to calculate the MDIs. Sample sizes for recoveries indicate the number of participants' whose TAIS records would be extracted for the evaluation. We assume a 98 percent coverage rate for these data elements, so actual sample sizes will be slightly lower than indicated in the table.

However, note the substantial declines in power that accompany a clustered design. Under a clustered design, a survey sample of 25,000 participants achieves about the same MDI on a binary variable (3.27 percent, shown in column 8) as is achieved by a survey sample of only 10,000 participants in an unclustered design (2.99 percent, as shown in column 5). The MDIs under the clustered design are also much greater for the recovery amounts compared to what can be achieved in an unclustered design. In fact, the benefits of being able to use very large samples through the TAIS data are roughly offset by the drawback (from a statistical perspective) of clustering; that is, the MDIs from the use of a clustered sample and large administrative data are in line with the survey-based MDIs shown for the comparison of web delivery models and the classic RCT (see Table II.3).

The evaluation team would have to work closely with OEA, CEO, and other potential stakeholders to determine the target MDIs for an evaluation of prioritization service delivery models. Based on that decision, the design option and sample sizes necessary to achieve the target MDIs could be selected.

E. Implementation and Cost Studies

As is described in Chapter II, it would be critical for the evaluator to conduct an in-depth implementation study alongside an evaluation comparing two service delivery strategies, particularly if random assignment is conducted at the individual level. This should be done for two reasons. First, it would ensure that program staff were conducting random assignment properly and adhering to the model assignment for each individual; any deviations could compromise the integrity of the study and invalidate the impact results. There would likely need to be a pilot period during which the evaluation team would conduct observations of the intake and random assignment process as well as the delivery of services to determine whether any inadvertent deviations from the study design occurred. Subsequent staff training could be conducted, if needed, to correct any issues. Second, an implementation study would be important for understanding how the alternative models unfolded over time. This information would not only help identify best practices, but also provide essential information for interpreting the impact results.

A cost study would also be important for comparing the resources needed to implement each of the service delivery strategies. The relative costs of the two approaches would provide OEA with information critical to deciding whether to adopt the new strategy more broadly or retain the existing one. For instance, if the evaluation was powered to detect a three percentage point impact on knowledge of benefits rights but the analysis showed no statistically significant differences on this outcome between the two approaches, it would suggest that OEA might want to use the less expensive approach, barring other important implementation factors.

As mentioned above with respect to the cost study recommended in Chapter II, cost studies can be done fairly generally by taking total budget and dividing by the number of participants assisted or much more specifically by building costs from a more detailed collection of data about how staff time and other resources are allocated. When comparing two service delivery approaches, we recommend the more specific approach. This is because there are likely to be subtle variations in costs across the two models that would need to be captured and it would be difficult to disentangle costs for each of the two models with the aggregate approach.

F. Potential Challenges and Recommendations to Address Them

There are several challenges associated with implementing a study to evaluate one of the prioritization service delivery models. Below we highlight two key challenges that emerged during conversations with OEA, OPR, and CEO and propose recommendations to mitigate the concerns associated with them.

Challenge 1: BAs would have to spend additional time collecting baseline data. Similar to the evaluation design described in Chapter II, the primary challenge associated with this evaluation design is the additional time that BAs might have to spend on baseline data collection during each call. This could reduce the number of calls that BAs could answer live and result in lower levels of customer satisfaction due to wait times or required call-backs.

To mitigate this concern, the evaluation could hire contractors to collect and enter baseline data before forwarding calls to BAs, as was suggested in Chapter II. In addition, the evaluation could track the rate of disengagement and modify the baseline data collection process as needed, limit baseline data collection to as few variables as possible, and provide training on collecting data efficiently. These recommendations are described in detail in Chapter II.

It would also be possible to conduct random assignment of individuals within a purposefully selected subset of regional offices. This would reduce the number of staff involved in the evaluation, but would also mean that certain offices would not get the chance to test the alternate delivery model. In addition, the evaluator would need to work with OEA and train staff across the nation to avoid crossovers that could occur if callers who went through random assignment sought additional services from regional offices that were not part of the experiment. OEA could consider including offices based on key characteristics or including a subset of high-, medium-, and lower-performing offices. Although the results would not be nationally representative, such a study could still provide useful information to OEA and DOL about the impact of different ways of doing business.

Challenge 2: The study could have an impact on customer satisfaction. Related to Challenge 1, customers might become frustrated by the hand-offs that would be required as part of a prioritization or prioritization plus specialization service delivery models. Callers with in-depth issues requiring senior BA assistance could potentially speak to three different people during the course of the phone call: a contractor who conducts random assignment and collects baseline data, a junior BA who determines whether the caller’s question requires in-depth assistance, and a senior BA to provide in-depth assistance. This could result in a decrease in customer satisfaction scores. To mitigate this concern, we suggest that the evaluation team work with OEA/EBSA to adjust expectations about the customer service standard, perhaps removing it temporarily as a performance measure for the program or adjusting the targets.

In conclusion, although EBSA and CEO have indicated that they do not wish to proceed with an evaluation of the prioritization service delivery model at this time, this appendix provides information that would be needed to implement such an evaluation should they decide to do so in the future.

This page has been left blank for double–sided copying.

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