DEMONSTRATION PROJECTS

EVALUATION HANDBOOK

DEE ANN BATTEN, PH.D.
(202-606-2544, DABATTEN@OPM.GOV)
ROBERT GOEHRIG
JOAN JORGENSON

U.S. OFFICE OF PERSONNEL MANAGEMENT
OFFICE OF MERIT SYSTEMS OVERSIGHT & EFFECTIVENESS
OFFICE OF MERIT SYSTEMS EFFECTIVENESS
PROGRAM DEVELOPMENT DIVISION
DEMONSTRATION PROJECT TEAM

TABLE OF CONTENTS

TABLE OF CONTENTS

I. INTRODUCTION

Demonstration projects are ". . . conducted by the Office of Personnel Management, or under its supervision, to determine whether a specified change in personnel management policies or procedures would result in improved Federal personnel management." (5 USC 4701) The law refers to evaluation of demonstration projects in 5 USC 4703 ". . . the Office of Personnel Management may, directly or through agreement or contract with one or more agencies and other public and private organizations, conduct and evaluate demonstration projects."

The goal of a demonstration project is to develop and test new ways of conducting personnel functions or applying human resource systems that are more efficient and effective and thus contribute to the organization's overall mission accomplishment and productivity. A tremendous amount of thought, planning, time and commitment from many people go into the overall demonstration project development, implementation and management. The goal of the demonstration project evaluation is to provide the necessary analytic information upon which conclusions and decisions about the project will be based. These conclusions and decisions include the effectiveness of the project in achieving its goals, the applicability of the project and/or its component innovations Governmentwide, and the resolution (modification, termination, continuation) of the project.

Past experience has shown that evaluations are resource (time, people and money) intensive. The combination of high cost and influential nature of the findings demand that as much care be given to designing and conducting the evaluation as the overall demonstration project. Indeed, the resources devoted to planning, implementing and monitoring the project will have been wasted without a good evaluation from which to base decisions about its future.

The purpose of this document is to provide a summary of the principles, techniques, and methods of conducting a demonstration project evaluation and to convey specific requirements that every evaluation must meet. The intent is not to create expert evaluators or ace statisticians in the few pages of this document, but rather to remove the mystery from the evaluation process, provide enough information to plan the evaluation, and provide a foundation for an effective partnership in designing and conducting a successful demonstration project evaluation.

The *why*, *how*, *what*, *when*, and *who* of a well-designed demonstration project evaluation is provided in the remainder of this section. Section II describes the ***Requirements*** for evaluation and provides guidance on how to meet them. Section III describes the components of the ***Evaluation Plan.*** Section IV includes information for consideration in selecting and working with your ***External Evaluator.*** The ***Conclusion*** contains a summary of the requirements presented in this document.

You may download a Wordperfect, or an Adobe Acrobat version of this document.

TABLE OF CONTENTS

A. WHY - THE IMPORTANCE OF THE EVALUATION

The demonstration project authority requires that each demonstration project be evaluated, directly or through contractors, to determine the demonstration project's "impact on improving public management" (5 USC 4703). The results from the project evaluation serve several purposes. First, the results allow an examination of the effectiveness of the innovations, goals and objectives of the project. Second, the results are used to determine the project's potential applicability elsewhere and to support the decision to pursue legislative changes Governmentwide. Third, the results are the basis for mid-course correction, that is, fine-tuning of the project based on early feedback and knowledge gained through the evaluation. Fourth, evaluation results are communication tools for documenting best practices and sharing lessons learned with employees, stakeholders, and the public. Fifth, the results aid in linking human resource management to organizational and mission outcomes (e.g., the Government Performance and Results Act, GPRA). And sixth, the results are the basis for extension, expansion, or termination of the project for the good of employees or the Government.

B. HOW - PROGRAM EVALUATION

Although the Office of Personnel Management (OPM) does not prescribe one evaluation approach, we do require a sound, defensible research design that protects the integrity of the evaluation process and produces reliable (repeatable) and valid (accurate) results that "measure the impact of the project results in relation to its objectives" (5 CFR 470.317(b)). Because demonstration projects occur in a natural environment, not in a controlled experimental environment (i.e., a laboratory), they are typically exposed to methodological criticisms. Most methodological criticisms concern the inability in a field setting to rule out other, uncontrollable, explanations for change. The use of appropriate program and process evaluation techniques will help rule out many of these uncontrollable explanations for change and allow the strongest (accurate and repeatable) possible conclusion regarding the effect and outcomes of the particular project innovations, as well as the impact of the project on mission accomplishment and organizational effectiveness and productivity.

C. WHAT - REQUIREMENTS AND GUIDANCE

The requirements and guidance described in the next section provide more specific information on the " . . . procedures, methods, and techniques that will be used to show whether the objectives have been met" and the " . . . data collection and analysis procedures to be used to assess the project's success or failure . . ." (5 CFR 470.301). These requirements are designed to ensure an effective evaluation of sufficient quality to meet the needs of various stakeholders. The requirements include technical as well as logistical aspects of evaluation. We have tried to make the technical aspects of evaluation clear and provide references for additional information. We can assist in designing an effective evaluation approach within reasonable resource limits.

TABLE OF CONTENTS

D. WHEN - TIME LINE

Evaluation planning is time-consuming, and is best begun at the inception of the demonstration project development. Evaluation planning is an integral part of the overall development of the project. As innovations are developed, so should the indices and methods used to measure the effect of these innovations. Designing the evaluation and developing the evaluation plan after the final *Federal Register* is published is not good practice and will not produce a quality evaluation with quality information. An effective evaluation must be begun even before project implementation. This in turn means that an evaluation plan should be designed and approved in conjunction with the final project plan well in advance of implementation.

The conduct of the evaluation is an ongoing process throughout the length of the project. Effective evaluation results cannot be based solely on periodic surveys or annually collected data. The best evaluations will take advantage of regularly gathered information, existing automated systems and frequent contact between the evaluator, the project staff and the employees and supervisors taking part in the project. This ongoing concept is similar to that used in agency human resource accountability and strategic planning and organizational performance measurement. Just as effective organizational performance measurement is based on gathering and using critical information on a routine and reasonably frequent basis, so is effective demonstration project evaluation.

E. WHO - ROLES AND RESPONSIBILITIES

The Agency and OPM have different roles in evaluation. OPM sets requirements for evaluation, provides technical assistance, guides and monitors the project and evaluation, and approves the evaluation plan and demonstration project evaluation reports. OPM is also responsible for pulling together results from similar demonstration projects to identify common themes and lessons learned for use in Governmentwide human resource legislation.

The department or agency designs, conducts, and funds the evaluation, and is ultimately responsible for the quality of the evaluation and the results. The agency selects the evaluator and ensures that the work of the evaluator meets OPM evaluation requirements. The evaluator typically provides methodological expertise and resources to gather and analyze data and prepare reports and briefings. There are a variety of ways in which the agency and evaluator may work to ensure a quality evaluation is designed and conducted. The best evaluation results are obtained when there is an open and collaborative relationship between the agency, evaluator and OPM.

Past projects have designated an evaluation team to design and conduct the evaluation and prepare and shepherd the evaluation plan. The evaluation team should include members of the larger demonstration project development team in addition to subject matter and technical experts such as researchers, statisticians, personnelists, systems experts, etc.

TABLE OF CONTENTS

II. TECHNICAL EVALUATION REQUIREMENTS

A. OVERVIEW

The requirements described in this handbook are designed to ensure the highest possible quality of evaluation results that will withstand anticipated methodological scrutiny and meet the information needs of the agency, OPM, and all other stakeholders. Before describing the specific requirements, it is important to provide some brief background material on evaluation. This information is provided as a simple yet strong foundation upon which the requirements in the following sections are built. For this and following sections, the most critical information is contained in the body of the document. Elaborating and/or exemplifying information is provided in appendices.

In their document *Designing Evaluations*, the Government Accounting Office (GAO) defines program evaluation as "The application of scientific research methods to assess program concepts, implementation and effectiveness." (page 88) General principles of evaluation have been developed to assist in designing and conducting projects and evaluations that yield the strongest possible results and are resistant to methodological criticisms. The first principle is *comparison*; contrast the demonstration project group results to those of a comparison group before and after implementation of the project innovations in the demonstration project group. The second principle is *manipulation*; change the factor of interest in the experimental group (in this case, implement the project innovations in the demonstration project group). The third principle is *control*; hold all other factors constant to help rule them out as a possible explanation of the results. The fourth principle is *generalizability*; infer the results to the larger population and determine possibility of application in other environments.

An evaluation example that applies these principles is provided in Appendix A. Further information on the kinds of problems that can affect the results, otherwise known as threats to internal validity are described in Appendix B. Finally, evaluation is most effectively accomplished in a stepwise fashion beginning with the definition of project objectives and hypotheses and ending with final decisions based on the evaluation results. The steps in evaluation are described more fully in Appendix C.

Evaluation requirements include: general issues such as what questions to address and the breadth and depth of evaluation; technical issues such as evaluation design, comparison groups, data collection, appointment of site historians, and statistical analyses; and logistical issues such as the contents of the evaluation plan, and reporting and timing of activities. Requirements are printed in ***bold italics*** to distinguish them from elaborating guidance. Key words and terms are written in *italics*.

TABLE OF CONTENTS

B. Evaluation Questions and Hypotheses

Many issues impact the questions and hypotheses that the demonstration project evaluation must address. These issues include the purpose and goals of the project, the potential impact and generalizability of the project, and the interests of many and varied stakeholders. Experience with human resource demonstration projects and the common questions asked by various stakeholders also influence the questions demonstration projects need to answer. Finally, it is important to track certain types of information in order to ensure the validity of the evaluation.

There are six general types or categories of questions that all demonstration project evaluations are required to address. However, the definition and operationalization (type and source of data collected) of the specific questions or hypotheses under each category will be developed for each demonstration project. The evaluation plan submitted by the agency will describe the specific questions and how data will be collected and analyzed to answer these questions. The section on report requirements provides information regarding the degree to which these questions must be addressed in formal reports, or monitored throughout the course of the project. The six areas all demonstration project evaluations are required to address are:

- ***Did the project accomplish the intended purpose and goals? If not, why not?***
- ***Was the project implemented and operated appropriately and accurately?***
- ***What were the costs, relative to the benefits of the project?***
- ***What was the impact on veterans and other EEO groups?***
- ***Were Merit Systems Principles adhered to and Prohibited Personnel Practices avoided?***
- ***Can the project or portions thereof be generalized to other agencies or Governmentwide?***

C. Breadth and Depth of Evaluation

Each of the six required areas the evaluation must address may be evaluated at varying degrees of breadth and depth. ***In general, the breadth of evaluation must be sufficient to make global assessments of the project's success and recommendations regarding its generalizability.*** The success of the entire project is as important, or perhaps even more important, than the success of each innovation. The degree to which the details of the project affect its success is critical to understanding how the project might be generalized to other agencies or Governmentwide.

The evaluation must be conducted in sufficient depth to assess the degree to which details of the project innovations and systems impact their success. The depth of the evaluation is also influenced by the degree to which the project or its innovations are uniquely designed or applied as compared to similar innovations and projects in previous demonstration projects. For example, it would not be necessary to evaluate the same broadbanding system used for the same types of employees as in previous demonstration projects in as much depth (detail) as a new broadbanding system used on new categories of employees. A well designed evaluation with an appropriate balance of breadth and depth can reduce evaluation costs.

TABLE OF CONTENTS

D. EVALUATION DESIGN

Quasi-experimental design

Many past evaluations have used quasi-experimental designs to evaluate the effects of the demonstration project. The quasi-experimental approach is one used when it is not possible to control assignment of individuals to the experimental (in this case demonstration project) group and/or to the comparison group. The quasi-experimental approach typically incorporates three features: a comparison group; baseline data; and a longitudinal design. Comparison groups allow testing the effects of the project innovations versus other competing explanations for change. Baseline data establishes conditions in the experimental and comparison groups before the demonstration project begins to provide an initial reference point. A longitudinal design uses data collected over time to compare to baseline data in order to determine when, and in what direction, an effect has occurred. These features operationalize the general evaluation principles described in section II.A. The three features are described in more detail shortly.

Implementation and Process Evaluation

It is important to note here that an adequate assessment of the results and outcomes of the demonstration project depends on the degree to which the project is correctly and adequately implemented and conducted throughout the life of the project. This concern for operational accuracy goes beyond a single check at the point of implementation. Because of the long-term nature of demonstration projects, a project can be implemented accurately, then gradually or even more quickly go off track at a later point.

The degree to which the project is not accurately and adequately implemented and conducted affects the *validity* (accuracy) of the results obtained from the assessment of the effects and outcomes of the project. If the project is not implemented correctly or does not operate as planned, then it is not the original project that is being evaluated, but potentially an altogether different project. This is what is known in evaluation as type III error. Thus the evaluation would not be a valid test of the results of the original project, but rather only of the project as it was incorrectly implemented or operated.

Although OPM performs reviews and assists agencies in developing accountability systems, it is ultimately the agencies' responsibility to be sure their human resource functions are operating accurately and according to law. Agencies and evaluators should work together to obtain and report information to provide for similar assurance for a demonstration project. ***Therefore, every evaluation must track the degree and accuracy of the implementation and operation of the project as described in the overall project plan.***

TABLE OF CONTENTS

Critical coverage areas

Assessing the effects and outcomes of the project is the most critical part of the evaluation. However, demonstration projects occur in a realistic environment in which legal, ethical and political considerations are also very important to the agency, OPM and other stakeholders. These considerations include the merit system, equal opportunity, and impact on mission.

Merit is the most important value upon which Federal human resource law, regulations and policy are based. In our system, merit is defined by the Merit Systems Principles (MSPs) described in section 2301(b) of title 5 USC. In addition, the law proscribes certain actions specifically contrary to these principles. These Prohibited Personnel Practices (PPPs) are contained in section 2302(b) title 5 USC. An adapted list of the MSPs and PPPs is contained in Appendix D. There has always been recognition that there is more than one way to define a merit based system. Otherwise, demonstration projects would not have been and continue to be supported by OPM and its stakeholders. While the law prevents demos from waiving a Merit System Principle, there is always the possibility of inadvertent negative impact on one or more MSPs or PPPs during the operation of a demo. No one can predict the impact of every new demo innovation. ***Therefore, each evaluation will assess the impact of the project on the Merit Systems Principles, and Prohibited Personnel Practices.*** This does not have to be at the in-depth legal level of a "compliance" review. We believe this can be done through periodic, easy, and unobtrusive means. However, it must be done in order to ensure the project and its innovations, if effective, are not so at the expense of merit or the perception of merit.

Special consideration also must be given to the impact of the project or its innovations on veterans preference and equal employment opportunity (EEO). It is especially critical to track this information in a demonstration project because they are after all experiments or tests of new human resource innovations. While we take every precaution in project design to ensure no additional risk for these groups, one might be logically concerned that veterans and EEO groups are more at risk in this setting. We require non-demo title 5 agencies to track information on veterans preference and EEO, even though it is a thoroughly tested system. It is therefore reasonable for ***every demonstration project evaluation to assess the impact of the project and its innovations on veterans preference and other equal opportunity groups.*** In addition, tracking this information is important because systematic and persistent adverse impact would be a concern impacting decisions about the project for the good of the government, or its employees. It is important for this information to be reported by the external evaluator as part of the overall evaluation because the results are more likely to withstand stakeholder scrutiny should there be a problem. This data is already being collected in every agency. We simply require summary information be reported on a regular basis by the evaluator.

Also, human resource management does not occur in a vacuum, but rather is an integral part of agency goals and mission. People are a strategic resource for all agencies. Indeed, the largest

budgetary allocation for most agencies is salaries and benefits. The passage of the Government Performance and Results Act in 1993, commonly referred to as GPRA, is having a substantial impact on the perception of human resources and the Federal human resource system in particular. This act requires all executive agencies to prepare strategic and annual performance plans, collect organizational performance measures, and use these measures to improve organizational performance. These measures will also be used by Congress beginning in the year 2000 to determine agency budgets. Demonstration projects are designed to improve the process and outcomes of human resource management. It is not sufficient for demonstration project evaluation to focus only on the efficiency of innovations, the link must be made to effectiveness of innovations and entire projects on the ultimate outcome of improving agency performance. ***Therefore, evaluations will assess the impact of the demonstration project on Government Performance and Results Act (GPRA) issues such as organizational effectiveness and productivity and mission accomplishment (5 USC 4705).***

The application of GPRA principles and development of GPRA measures is ongoing across the Government. The development of measures for strategic human resources is also an ongoing process. Several measures related to human resources can be found in the *Accountability System Development Guide*. Agency personnel responsible for strategic and long term planning are another good sources for the most recent information on appropriate agency-based GPRA (and other similar) measures and data for demonstration project and comparison locations.

E. COMPARISON GROUPS

In order to make statements concerning the effect of the project and its innovations, statistical comparison must be made to the existing system. There are several approaches available for making such comparisons. The strongest methodological approach is to define an intact comparison group to which the demonstration group is compared. The comparison group approach is described in this section because it is the standard against which other approaches are measured. ***Each project will determine the strongest possible approach for comparison, provide supporting rationale, and be fully prepared to execute the approach prior to project implementation.***

An *experimental group* is the group or site at which the treatment has been applied or intervention has taken place. In the case of a demonstration project, this group is often referred to as the "Demo" group. *Comparison groups* are similar to the experimental group in an important way(s), but do not receive demonstration project interventions. We assess the effect of the project by contrasting the change in the experimental group to the change in the comparison group. If a change occurs in the experimental group only, then a case can be made that the project intervention had an effect. If similar changes occur in both groups, it is impossible to know if the project intervention had an effect because intervening variables could have affected both groups.

TABLE OF CONTENTS

Comparison groups or sites should be selected to be as similar as possible to demonstration project sites. Some criteria to be considered in the selection of *comparison* groups are: 1) mission or function; 2) organizational structure (regional versus headquarters environment); 3) workforce factors (occupational mix, grade structure, diversity and veterans) and; 4) change momentum versus traditional environment. The sites need not be identical but it is important to note how the sites differ as well as how they are similar. Thus, baseline measurement of both the experimental and *comparison* sites is critical. The noting of differences is done to account for competing explanations of change (see Appendix B). Some of the *comparison* groups used in other demonstration project evaluations are described in Appendix E.

It should be noted that a *comparison* group is not the same as a *control* group used in a "controlled" experiment. In a demonstration project, the construction of *control* groups is generally not possible because participants cannot be randomly assigned to the experimental or "control" groups. True *control* groups are formed by the process of randomization. That is, participants are assigned to either the control or experimental group using a randomized procedure. Groups constructed of randomly assigned participants will be roughly equivalent in all important respects. The comparison of experimental and *control* groups in a controlled experimental environment allows for very strong statements that a specific innovation *caused* the resulting change in the experimental group. Because we cannot have random assignment in a demonstration project, we cannot make strong statements of causality. However, the use of *comparison* groups does allow for strong and useful statements of effect based on sound scientific methodology.

F. DATA COLLECTION REQUIREMENTS

Type of Data

The type of data collected depends on the project innovations tested, outcomes anticipated and the availability of information. It is important to collect quantitative (countable, numeric) and qualitative (e.g. focus group, case studies) data. It is also important to collect objective (factual, unbiased occurrences such as number of applicants or hires, salary rates) and perceptual/ attitudinal (supervisor's rating of applicant quality) data. In addition, data should be a credible measure of the effects of interest. The ***strongest possible objective data, as well as perceptual/attitudinal data are required*** to assess outcomes and/or effects of the project. Without objective data, the results will only convey how the participants perceive or feel about the demonstration project. This information is important, but is not sufficiently compelling to assess the effects of the project, inform policy, or address the interests of the many stakeholders. Finally, it is most important to focus data collection resources on the most important, innovative, and least previously tested innovations and objectives of the project.

TABLE OF CONTENTS

Timing of Data Collection

Baseline objective and perceptual/attitudinal data collected prior to project implementation from the experimental and comparison groups are required in order to determine the initial, pre-implementation status of the experimental and comparison groups or sites. Collect baseline (especially objective) data that completely describes the groups under the current system and adequately assesses items that are predicted to change during the course of the project innovations. The same data is then collected periodically (longitudinally) from the experimental and comparison sites after the implementation of the project. Contrasts of baseline and longitudinal data within the experimental sites, as well as contrasts of data to comparison sites, provide for the strongest possible statements of the project effects or outcomes. ***Longitudinal objective and perceptual/attitudinal data are required from the experimental and comparison groups.*** While the frequency of longitudinal data collection is not specified here, data should be collected often enough to distinguish important changes in the project's effects and/or operation. It may be important to collect some data quite frequently and other data less frequently. It is also important to collect information from the demonstration project and comparison groups at approximately the same time.

Source/Method of Data Collection

We encourage agencies to use existing data and data systems where appropriate. Not all data needs to be newly gathered. Indeed, agencies that have operational human resource accountability systems and organizational performance measurement systems can obtain data more easily and at reduced cost. Evaluation data can and should be obtained from a variety of sources including existing systems, personnel files, personnel offices, focus groups and surveys. A list of examples of data and their sources is contained in Appendix F. Objective data is collected primarily from personnel files (workforce data) and/or aggregated by personnel office staffs. Demographic data and information such as pay, grade, length of service, and other data obtained from individual personnel files is most easily collected from agency- based automation systems. Information from the Central Personnel Data File (CPDF) maintained by OPM may also be available, however, this information may not be as current as agency based data. The specific items contained in the CPDF are listed in Appendix G. Aggregate workforce and personnel office data such as quality of applicants, timeliness of hiring actions, etc., are additional sources of objective information about the effect of the project innovations. However, obtaining workforce and personnel office data can be costly and time-consuming because collection of this data can require new, specifically designed computer automation systems, labor intensive manual searches through files, or maintaining on-site logs of personnel actions. In general, considerable resources (time, people, training) are required to obtain good objective information. It is important to focus data collection resources on the objective data necessary to assessing the effects of the most important and unique aspects of the project and outcomes.

TABLE OF CONTENTS

Focus groups and case studies are useful sources of qualitative (sometimes quantitative) perceptual and/or attitudinal data. The degree to which this qualitative information agrees with the information learned from objective sources provides stronger support in assessing the effectiveness of the project. Qualitative data sources are particularly useful to alert interested parties to misperceptions and potential problems with the implementation or operation of the demonstration project that need further investigation with objective data. However, significant training is required to ensure that the collection and analysis of data from qualitative sources is not biased by the reviewer/interviewer or other participating group members with strong opinions. The cost of gathering this type of data needs to be weighed against the potential benefits to determine what type of information about what specific project innovations and outcomes can best be collected from these qualitative sources. ***Whatever methods of data collection are chosen, thorough training of data collectors is required prior to collection of the data.***

Surveys are a useful method of obtaining quantitative perceptual or attitudinal information from participants. For example, a baseline attitude survey is often administered before the implementation of the project at both the experimental and comparison sites. A comparable form (same questions asked the same way, but perhaps fewer topics) of the survey is administered longitudinally, at various points after the implementation of the project. Comparisons between the baseline and longitudinal survey responses indicate the degree to which attitudes and perceptions of the innovations change throughout the course of the project. Surveys include background data such as demographics and questions critical to assessing the specific innovations, goals and expected outcomes of the demonstration project. Surveys must be developed to test specific objectives and/or hypotheses because surveys that are too long can result in low response rates and lower quality data. Other surveys are a source of tested and validated survey questions that can serve as the foundation for your survey. Survey questions from the private sector are often copyrighted and cannot be used without permission. Surveys in the public domain are usually free (unless copyrighted) and can be very cost effective. Survey modules that contain items from many public sector surveys on more than 30 personnel and organizational change topics are available upon request and are listed in Appendix H.

Finally, it is important to understand that data can be collected from a *census* of all employees (or other measurement unit) or from a selected group, or *random sample*, of employees. The *census* has the advantage of providing information from all employees on a specific topic. The *random sample* usually costs less, but an appropriate random sample can be difficult to operationalize at the project and comparison locations. There are also choices to be made about how much of the various types of data to collect. Generally speaking more data is better. However, the larger the amount of data, the easier it is to show that even trivial (relatively unimportant, or unmeaningful) effects are statistically significant. In this case, the results need to be reported in terms of *effect size* in order to convey the meaningfulness of the statistically significant results. Alternatively, one can have so little data that it is impossible to statistically show a difference exists even when it really does exist. ***Demonstration project evaluation must collect sufficient data to be able to adequately test the statistical significance of results, and report effect sizes when necessary.***

TABLE OF CONTENTS

G. SITE HISTORIAN

In evaluation research *history* can be defined as an historic event that occurs during the time of the project that can provide a rival explanation for a project outcome. It is critical that the effects of *history* be accounted for when evaluating the demonstration project. It is necessary to maintain a written log of events when the project extends over a long period of time and when the project covers multiple sites that are subject to different influences. For example, one project location may be affected by a change in the local labor market that does not affect the other locations; or a change in policy or operating procedure may not be remembered several years down the road.

The *site historian* is a person specifically charged with documenting *historic* events and changes in both the experimental and comparison sites that have, or may have, an impact on the results of the demonstration project. ***Every demonstration project will have at least one site historian for the demonstration project group and one site historian for the comparison group. Site historians will be selected and trained prior to project implementation and will function for the duration of the project.*** The geographical dispersion, complexity and size of the project will determine how many historians are necessary and where they will be located.

The *site historian's* role is to recognize and record the historic events that are important but which are not necessarily captured elsewhere. This is not a time or resource intensive task but it does require an observant individual. The *site historian* is not required to make a judgement about the importance or expected effect of the event; that is the evaluator's job. It is only necessary to recognize that the event might have an affect and to record the event. The most difficult job of the *site historian* is deciding what to record and what to ignore. It is better to err on the side of caution by over-documenting. The events documented are then sent to the evaluator and used to interpret the results of the evaluation, and the event histories are summarized in evaluation reports. For example, a lower than expected number of hires may be due to a hiring freeze, rather than a problem with a project innovation. Examples of events to record include:

- A change in the environment -- a major employer (e.g., a defense contractor or another federal agency) in the local area lays off a large number of employees.
- A change in the organization -- an reorganization, merger, a reduction in force, a hiring freeze, a hiring freeze is lifted, a shift in agency or installation mission, or a consolidation of personnel to a central location.
- A change in resources -- budget cutbacks or budget growth, staffing allocations.

- A change in procedures -- a reengineering effort is undertaken or there is a major change in work policies and/or procedures.

The *site historian's* journal or log will be used to record the events and changes that may provide a rival explanation for changes in evaluation measures. A standardized reporting format is not practical since every event and site is different. However, each entry should: 1) record the date(s) the event occurred; 2) indicate the names, titles, parties, and/or organizations affected; and 3) include a narrative description of the event or change. A sample site historian's log is contained in Appendix I.

H. STATISTICAL ANALYSIS

Once the data are collected, statistical procedures are used to collate, describe, and analyze the data in order to determine the degree to which the project purpose and goals were accomplished and the potential generalizability of results to larger groups. There are many possible statistical procedures from which to choose. The choice of appropriate procedures is determined by the project questions and hypotheses, evaluation design, and type of data collected. Most statistical procedures can be divided into two types; *descriptive* and *inferential*.

Descriptive procedures help convey basic and relative information on the experimental and comparison groups. Usually this information is presented in tabular or graphical format. For example, a vertical bar chart might display the Grade Point Average (GPA) of new applicants in the experimental and comparison groups during the course of the demonstration project. ***The choice of descriptive statistical techniques must yield a complete, accurate, clear and concise presentation of the data related to critical project questions or hypotheses.*** Graphical analyses, if used appropriately and carefully, can convey a lot of information in a limited space. Graphs should be easily interpretable and comparable.

Descriptive statistics provide indications of the differences, or lack thereof, in our observed samples of individuals. However, the primary goal is to *infer* or generalize these differences to the *populations* represented by our observed *samples*. For example, we want to say that a new pay system works for all people in these occupations, not just those who work in this particular facility on this particular mission. *Inferential* statistics are used to assess the degree to which differences that appear in the descriptive analyses reflect significant (in probabilistic, statistical, or mathematical terms) differences in the *populations* represented by the *sample*. Without some *inferential* analyses it will be impossible to make any generalizations about the effect of the project innovations or outcomes. *Descriptive* statistics only describe the data, *inferential* statistics are used to determine if something really is different, and whether the results are generalizable to other groups. ***Appropriate inferential procedures are required to statistically assess the degree to which the project purpose, goals and outcomes are met, and the***

generalizability of the project to other groups. The presentation of only descriptive information is not sufficient to make statements about the impact of the project innovations and outcomes.

The choice of appropriate inferential procedure is dependent on many issues that are too complex to review here. The goal is to choose a procedure that fits the data and will answer the project questions. The statistical treatment of the data should be considered prior to collection of data. The evaluation plan should include enough information about planned statistical analyses to convey that the agency and external evaluator have a clear understanding of the issues involved in analyzing the data. We do not prescribe specific inferential methods. Rather, we expect the evaluator to make appropriate, accurate and effective use of state-of-the-art inferential methods.

TABLE OF CONTENTS

III. EVALUATION DOCUMENTATION REQUIREMENTS

A. OVERVIEW

The content of the demonstration project and its innovations are documented for communication, operational use and as a major part of the historical record of the project. Documentation of the evaluation is also necessary to convey information to the project staff, employees and the public, serve as the basis for more specific operational information, and function as a public record. Evaluation documentation comes in several forms including the complete evaluation plan, the evaluation section of the *Federal Register* notice, and evaluation reports and briefings. Each of these types of documentation are described along with their requirements in this section.

B. EVALUATION PLAN

Contents

The demonstration project evaluation plan provides complete documentation of the nature, rigor and intent of the evaluation. The evaluation plan must be as fully developed a description of the project evaluation as the project plan is of the innovations. According to 5 USC 4703, the demonstration project plan must include "... the methodology and criteria for evaluation." This broad mandate is given a bit more definition in regulation. 5 CFR 470.301 says the evaluation plan must include:

- measurable goals **or** objectives;
- acceptable expected results or outcomes;
- a description of the procedures, methods, and techniques that will be used to show whether the objectives have been achieved;
- a description of the data collection and analysis procedures to be used to assess the project's success or failure from a qualitative or quantitative standpoint; and
- an itemization of costs and benefits associated with the project to the agency, the Government, and the community.

The plan is required to clearly communicate how the agency will address the issues above and the technical requirements for evaluation that were described in Section II. The evaluation plan may contain a model of the overall evaluation approach and text that addresses specific details not contained in the model. We do not prescribe a format for the evaluation plan; this is best determined by the agency and evaluator. Rather, we prefer to focus on the types and

purpose of information required. The plan must cover all evaluation activities from initial planning through data analysis and reporting. The plan must clearly convey how the evaluation will operationalize and answer each project hypothesis or research question.

As a general rule, the inclusion of more information will provide better understanding of the project evaluation and hopefully promote a smoother evaluation process that yields better results. Minor adjustments in the plan may need to be made during the course of the project in order to address unanticipated issues and maintain the quality of the evaluation. The plan should anticipate such adjustments where possible by indicating flexibility, or communicating options for predicted events that may require adjustment. ***The plan must be complete and contain sufficient detail to ensure the agency is prepared to conduct a quality evaluation.*** Note that while we do not require every operational level detail in the evaluation plan submitted to OPM, these details will have to be determined and communicated to all evaluation participants in order for the evaluation to be effectively conducted and yield valid, reliable and useful information.

We also want to make clear here that the timeliness of evaluation activities is absolutely critical to an effective evaluation that will yield quality results. Evaluation activities must begin even before implementation in order to obtain unbiased baseline information. In addition, experience and accepted research practice has shown that evaluation activities conducted in the absence of an overall evaluation structure and approach are much more likely to yield data that isn't useable, miss the chance to obtain data critical for answering important questions, and increase costs of the evaluation. Retrospective data are often simply not valid, reliable or useful for answering demonstration project questions. In order to minimize this problem, ***the evaluation plan must be submitted to and approved by OPM prior to project implementation.***

Evaluation Models

Modeling is a means of simplifying a complex array of information. Agencies are encouraged to develop a model of their evaluation plan. Models are an aid in making the links between research question and measurable indicators of effects that help ensure the plan is executable. Models are also an excellent means of communicating evaluation information to interested parties.

A good evaluation model will include information on the project innovations and/or the expected effects and general outcomes of the innovations. These will then be tied to observable measures or data for each expected effect. Finally, the source of the observable measures will be listed. Models may focus on the innovations and/or the general outcomes. Multiple models may be used to present the different levels of intended analyses. Whether or not models are chosen and the type of modeling are not as important as ensuring that complete information is conveyed in the evaluation plan and in the evaluation section of the *Federal Register*. Three different evaluation models are presented as examples in Appendix J.

TABLE OF CONTENTS

C. EVALUATION SECTION OF THE *FEDERAL REGISTER* NOTICE

The evaluation section of the *Federal Register* need not contain the entire evaluation plan, but must contain enough evaluation plan details to ensure the public, participating employees, and other stakeholders of the intent, rigor, and nature of the evaluation plan. This can usually be accomplished by including an appropriate model and text addressing other critical issues. The text may be a summary of what is included in the formal evaluation plan presented to OPM. It is most important to convey the degree of coverage of the evaluation and recognition of and intent to address critical methodological issues in the *Federal Register*. The information contained in the evaluation sections of most previous *Federal Register* notices do not contain sufficient information to serve as the evaluation plan.

D. EVALUATION REPORTS

General Issues

The OPM, pursuant to law, shall "establish and maintain a program for the collection and public dissemination of information relating to personnel management research and for encouraging and facilitating the exchange of information among interested persons and entities ..." (5 USC 4702(3)). And, "Upon request of the Director of the Office of Personnel Management, agencies shall cooperate with and assist the Office, to the extent practicable, in any evaluation undertaken under subsection (h) of this section and provide the Office with requested information and reports relating to the conducting of demonstration projects in their respective agencies" (5 USC, 4703).

In addition to the fact that reporting of the evaluation is required by law, it is equally important to provide a complete, accurate, and fair record of the demonstration to all stakeholders and the public. And, most important for the agency, evaluation reports are the basis from which decisions about the resolution of the demonstration project will be made. That is, the decision to terminate, extend, make permanent, or seek Governmentwide legislation is based largely on the quality, timeliness, and content of the evaluation reports. Reports are the ultimate decision documents and public record of the project. They must meet the needs of a variety of stakeholders.

The reports described here, as well as any additional report with the potential to be released to the public, will be submitted to OPM for review and approval. We do not review and approve reports to control the interpretation of information by the agency or the evaluator, but rather to ensure the interpretation is objective, based on good data, and complete. We may occasionally ask for additional analyses to clarify confusing or controversial issues. On the rare occasion when we might disagree with the interpretation of results, we reserve the right to have a letter containing our concerns and additional interpretation in the report. In addition, report approval ensures OPM is fully informed of the results and their interpretation, enables us to support the agency in sharing results with stakeholders, and helps ensure the "story" of the demo is told in a

consistent, timely and meaningful way. OPM will also, within certain guidelines, review draft reports with the goal of fostering the development of timely, quality reports that meet the needs of stakeholders. Approval of demonstration project evaluation reports has been standard practice in OPM for many years. This requirement is not new, simply more clearly spelled out in this Handbook.

Of course, regular contact with your OPM demonstration project manager on evaluation issues (as with all other demonstration project issues) is assumed. And, immediate notification of any severe operational difficulty with any part of the project, whether learned through the evaluation or another source, that results in potential harm to an individual or group, or which may impact the validity of the evaluation results is expected. In addition, past experience has shown that critical, high visibility questions about a project, or more commonly one of its innovations, are sometimes asked by key stakeholders. It is difficult to anticipate what issues or concerns may be involved, but it is wise to be aware that quick response analyses may be required from time to time. Maintaining open and clear communication throughout the course of the evaluation will help alleviate any negative impact of these requests.

The specific contents and timeliness of required formal reports are described below. In general, these are scientific reports that will serve as decision documents and as the public record of the demonstration project. Scientific reports should convey accurate, reliable, valid and complete information about the project in as clear and objective a way as possible. The reports should be written in plain language, for a wide audience and should explain technical terms and avoid the use of jargon. Statements and conclusions made in the report should be based on data contained in the report, or specifically referenced to another source. The report should flow well with the textual and tabular or graphical material linked appropriately. Each report should include an executive summary that covers the main purpose, results and conclusions. All required reports should be formatted to be made available on the World Wide Web. ***Finally, OPM requires an electronic version of the evaluation data for purposes of conducting comparative demonstration project evaluation used to inform legislation.*** Format requirements for this data will be dependent on current computer technology and will be provided upon request.

Types and Schedules

Reporting on the evaluation of the demonstration project can take a variety of forms. Formal reports are required at certain times as described later in this section. However, it will also be necessary to update and report some information about the project during the periods between formal reporting. These intervening reports or updates can take a variety of less formal forms, including letter reports, briefings, or other similar form. ***Updates are required between periods of formal reporting on the status of the project and the evaluation, information pertaining to the accuracy with which the project is implemented and operating, and the impact on equal employment opportunity, veterans, the Merit Systems Principles, and the Prohibited Personnel Practices.***

TABLE OF CONTENTS

The types of evaluation formal reports and the schedule for receiving the reports will be individually developed for each demonstration project and agreed to before the project is implemented. ***Required formal reports include: 1) a baseline/implementation report early in the demonstration project; 2) an interim report prior to termination of the project from which decisions about the project's future will be made; and 3) a summative report within one year after the end of the project. Additional, earlier interim reports will be required in order to make any substantive mid-course corrections in the project.***

Baseline/Implementation Report. This report, ***submitted in final form no later than 18 months*** after project implementation, contains project information and data through the first 12 months of the project. It serves as the public record of early project activities and the reference point upon which later comparisons will be made to determine the effect of the project. This report should include:

- An overview of the project describing purpose, goals, agency environment, demonstration project and comparison groups, and project innovations.
- Evaluation methodology and plan.
- Presentation and analysis of objective and perceptual/attitudinal data from the baseline through the first year of the project from both demonstration project and comparison groups. It is critical to include all necessary objective data as the reference point upon which later comparison will be made.
- Summary of site historian information to date.
- Description of implementation activities, including preparation of operations manuals, training provided for demonstration project operation and data collection, communication activities; problems/solutions identified in implementation or early operation of the project; lessons learned; and issues to watch during the remaining portions of the project.
- Appendices containing measurement instruments, data collection methods and protocols, and other supportive technical information.

Interim Report(s). ***The interim report consists of two documents: A Management Report and a Technical Report***. These reports ***present data through at least the first 3 (three) and one***

half years of the project and must be delivered in final form no later than 9 (nine) months prior to the original expiration date of the project. The Management Report will be a stand-alone report no longer than 30 pages written for various stakeholders including those who will determine the outcome of the project. This report presents summary information and critical results (referenced to the Technical report as necessary) to be used to decide if the project will be made permanent, extended for another period of time, continued until the original expiration date, or terminated early. The Technical report will contain complete information as the public record of the project to this point, and supporting information for decision makers. In general, the reports must address the following issues, with the Management report necessarily containing briefer summaries and the most critical decision related information:

- Brief review of the project and implementation taken from the baseline report.
- Reports of longitudinal objective and perceptual/attitudinal data collected since the baseline.
- Report on the accuracy of operation of the project innovations and the resolution of any discovered problems.
- Site historian summaries, especially new information since the baseline/implementation report.
- Descriptive and inferential statistical analyses of data from the first 3 and one half years of the project compared to the baseline data to determine if the project accomplished its purpose goals and objectives, if not, why not.
- Thorough assessment of the impact of the project on EEO, Veterans, Merit Systems Principles, and Prohibited Personnel Practices.
- Assessment of impact of project on mission and organizational outcomes.
- Conclusions on the efficacy of the innovations and project as a whole.
- Recommendations for resolution of the project (extend, expand, terminate) based on evaluation results.
- Appendices containing measurement instruments, data collection protocols, and other supportive technical information, not contained in previous reports.

Additional (usually earlier) interim reports may be required as documentation for mid-course corrections to the demonstration project. A determination of the need for one of these reports and specification of the contents will be made if and when problems occur that require a mid-course correction to the demonstration project.

Summative Report(s). Similar to the interim report requirement, ***the summative report will***

consist of two documents: A Management Report and a Technical Report delivered in final form no later than one year after the originally scheduled end of the project. The results of the entire project are fully documented in this set of final reports to show the demonstration project's successes, best practices, lessons learned, and the potential for expansion to other agencies or departments or Governmentwide. The summative reports are critical as they publicize what has been accomplished and provide stakeholders (e.g., other agencies, policy makers) with the necessary information to implement similar project innovations. The Management report will contain a review of project purpose and goals, summary of implementation results as necessary, major results in the form of visual charts, major conclusions, and discussion of the appropriateness for expanded or Governmentwide application of the project. The Management report will be a 20-30 page stand-alone report that may reference sections of the technical reports as appropriate and necessary. The Summative Technical report will contain a complete review of the project, operational information collected during the demonstration project, measurement instruments, data analysis techniques and procedures, complete presentation of summative results from all data sources, and thoroughly discussed conclusions. The following issues should be addressed in this report:

- Brief review of the project innovations and implementation information.
- Complete information on accuracy and degree of project operations throughout the life of the project, review of problems encountered and their resolution.
- Complete summary of site historian events and their affect on the project.
- Reports of objective and perceptual/attitudinal data collected throughout the project.
- Descriptive and inferential statistical analyses of data to determine degree to which the project purpose, goals, and objectives were met, if not, why not; and effects of the project as a whole.
- Thorough assessment of the impact of the project on EEO, Veterans, Merit Systems Principles, and Prohibited Personnel Practices.
- Assessment of impact of project on mission and organizational outcomes.
- Conclusions of the effects of the innovations and project as whole.
- Recommendations for limited or Governmentwide expansion.
- Description of best practices, lessons learned and where to go from here.
- Appendices containing measurement instruments, data collection protocols, and other supportive technical information not contained in previous reports.

TABLE OF CONTENTS

IV. EXTERNAL EVALUATOR

A. SOME SUGGESTED ISSUES

This section is included to summarize previous information and include other issues agencies may consider when selecting and/or working with an external evaluator. This section is not intended to direct the agency in its method or criteria of selection of an evaluator. Rather, the intent is to provide in one place some important information that may assist the agencies in talking with potential or selected external evaluators about evaluation. This information is provided because the selection of a competent external evaluator with which the agency can establish a good working relationship is critical to obtaining a quality evaluation of the demonstration project. The issues presented below, along with the issues presented more formally elsewhere in this document, will provide an initial list of areas to address with a potential or selected evaluator. This is not intended to be an all inclusive list. OPM demonstration project managers can provide additional expertise and share the latest issues that can impact the demonstration project evaluation.

- Planning
  - How much experience in evaluating Federal personnel programs?
  - How much experience in program evaluation in general?
  - How much experience in evaluating other personnel programs?
- Data Collection
  - Who collects data to provide to contractor -- contractor or agency?
  - Who designs/prepares data collection protocols?
  - How much commitment required from the participants in the field?
  - How much can be automated?
  - What about anonymity and protection of participant identity?
  - Focus groups; how many, where?
  - Who owns final data?
  - Copy for OPM.
- Surveys
  - Designed for specific issues?
  - Low-cost options?
  - Who distributes and collects?
  - Sampling versus census?
  - Who owns survey?
- Training
  - Site historian, focus group leaders and interviewers?
  - Personnel office data collectors?
- Data Analysis

  - What statistical procedures?
  - Use of new designs or statistical methodology?
  - Will the results answer the important questions?
  - Can results be clearly communicated?

- Reporting
  - Schedules?
  - Contents and review mechanisms?

B. STATEMENT OF WORK

A statement of work (SOW) is typically used in the contracting process to describe the major tasks and sub-tasks to be performed, the methodologies to be used, and the responsibilities of those involved in the evaluation. An SOW usually begins with some background information including the purpose of the project, description of the demonstration project authority and information about the innovations. It will usually address many of the items listed above in Section A, as well as details about the evaluation plan such as methodological approach, anticipated methodological difficulties, and specific analytic skills and experience required of the evaluator. The SOW will also describe the contents and timing of required deliverable products and the method of collaboration (if any) that may be used to complete the work. It is most important that the statement of work be as complete and clear as possible. A well-written statement of work will result in contractor proposals that address the needs of the agency and the many other stakeholders interested in the evaluation.

C. SOURCES OF EXTERNAL EVALUATORS

Agencies may use outside consultants, another Federal agency, a college or university, a profit or non-profit organization, other agency components with evaluation expertise, or internal, yet organizationally independent, evaluation teams to assist in the design of the plan and/or to conduct the evaluation. ***The external evaluator is required to be organizationally independent from the organization participating in the demonstration project.*** This is to avoid a clear or potential conflict of interest that may taint the project results. ***Finally, an evaluator must be on board prior to implementation of the project.***

Previous contractors have been private sector, universities, non-profits, and other Federal entities. A list of evaluators who have previously been contracted to conduct demonstration project evaluations is included in Appendix K. The inclusion of this list is not an endorsement of these contractors by OPM, but is here only to illustrate the range of evaluators that have conducted demonstration project evaluations in the past.

TABLE OF CONTENTS

V. CONCLUSION

A successful evaluation depends on all parties working together to ensure an evaluation that will yield valid (accurate) and reliable (repeatable) results, that withstand methodological scrutiny, and meet the information needs of the agency, OPM, and other stakeholders. Without a quality evaluation, the effort devoted to the demonstration project will have been wasted because there will be no basis from which to make decisions about the resolution of the current project or the expansion of the project innovations to other environments or Governmentwide.

The intent of this document has been to provide fundamental information on the principles, techniques, and methods of evaluation and to convey the requirements for evaluation of a demonstration project. Hopefully, this document has provided sufficient initial information for a productive and collaborative evaluation endeavor between OPM, the agency, and their external evaluator. Further, more detailed information about various aspects of evaluation can be obtained from the references listed in the bibliography.

A summary of the requirements for evaluation described in this document are organized below in terms of the general stages of evaluation. A checklist containing these requirements is contained in Appendix L.

General Issues

- Evaluation assesses degree to which project purpose and goals are met, and if not, why not
- Evaluation assesses cost
- Evaluation tracks project implementation and operation
- Evaluation assesses impact on veterans and other EEO groups
- Evaluation assesses impact on Merit System Principles and Prohibited Personnel Practices
- Evaluation assesses degree of generalization of project to other groups or Governmentwide
- Sufficient evaluation breadth for global assessments and recommendations
- Sufficient evaluation depth to determine effect of project details

Planning - Pre-implementation

- OPM approves complete evaluation plan
- External evaluator (independent of participating agencies) on board
- Determine, support and prepare to execute strongest possible comparison approach
- Training of data collectors prior to data collection
- Selection and training of site historians for demonstration project and comparison groups
- Collection of baseline objective and perceptual/attitudinal data from demonstration project and comparison groups

TABLE OF CONTENTS

Execution

- Collection of strongest objective and perceptual/attitudinal data
- Longitudinal objective and perceptual/attitudinal data from demonstration project and comparison groups
- Track degree and accuracy of implementation/operation of project
- Collect data on adherence to Merit Systems Principles, avoidance of Prohibited Personnel Practices, impact on Veterans and other EEO groups
- Collect data on GRPA-type issues such as mission accomplishment and organizational productivity and effectiveness
- Collect enough data to statistically identify important effects, and report effect sizes when necessary
- Maintain Site Historians

Analysis

- Descriptive statistical Techniques that result in complete, accurate, clear, and concise presentation of data linked to critical questions or hypotheses
- Appropriate inferential procedures required to assess impact of the project

Reports

- All reports with potential to be released to the public approved by OPM
- Baseline/Implementation report no later than 18 months after project Implementation
- Management and Technical Interim Decision Reports delivered no later than nine (9) months prior to project expiration
- Management and Technical Summative Reports delivered no later than one year after project expiration
- Intervening updates on status of evaluation, project operation, impact on Veterans, EEO, Merit Systems Principles and Prohibited Personnel Practices
- Electronic version of data provided to OPM/OMSOE

TABLE OF CONTENTS

SELECTED REFERENCES

General Evaluation Theory, Design, and Analysis

Bingham, Richard D. and Claire L. Felbinger (1989). *Evaluation in Practice: A Methodological Approach.* White Plains, NY: Longman, Inc.

Campbell, Donald T. and Julian C. Stanley (1963). *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally Publishing Company.

Cook, Thomas and William Shadish (1986 & 1994). *Annual Review of Psychology.*

Langbein, Laura I. (1980). *Discovering Whether Programs Work: A Guide to Statistical Methods for Program Evaluation.* Glenview, Il: Scott, Foresman.

Nachmias, David (1979). *Public Policy Evaluation: Approaches and Methods.* New York: St. Martin's Press.

Rossi, Peter H., Howard E. Freeman, and Lipsey, Mark W. (1999). *Evaluation: A Systematic Approach.* 6th Edition. Newbury Park, CA: Sage Publications.

Shadish, William (1991). *Foundations of Program Evaluation.* Newbury Park, CA: Sage Publications.

Stecher, Brian M. and W. Alan Davis (1990). *How to Focus an Evaluation.* 2nd Edition. Newbury Park, CA: Sage Publications.

U.S. General Accounting Office (1991). *Designing Evaluations.* Transfer paper 10.1.4 (PEMD-10.1.4). Washington, DC. (This is a FREE publication that can be obtained by calling GAO on 202-512-6000 or on http://www.gao.gov).

Case Study Research

U.S. General Accounting Office. *Case Study Evaluations.* Transfer paper 10.1.9. Washington, DC. (This is a FREE publication that can be obtained by calling GAO on 202-512-6000 or on http://www.gao.gov).

Yin, Robert K. (1989). *Case Study Research: Design and Methods.* Newbury Park, CA: Sage Publications.

TABLE OF CONTENTS

Survey Research

Converse, J.M. and Stanley Presser (1986). *Survey Questions: Handcrafting the Standardized Questionnaire.* Newbury Park, CA: Sage Publications.

U.S. General Accounting Office (GAO). *Developing and Using Questionnaires.* Transfer paper 10.1.7. Washington, DC. (This is a FREE publication that can be obtained by calling GAO at (202) 512-6000 or on http://www.gao.gov).

Focus Group Research

Krueger, Richard (1994). Focus *Groups: A Practical Guide for Applied Research* (2nd Ed.) Thousand Oaks, CA, Sage Publications

Analysis - General and Specific Methodologies

U.S. General Accounting Office (1989). *Content Analysis: A Methodology for Structuring and Analyzing Written Material.* Transfer paper 10.1.3. Washington, DC. (This is a FREE publication that can be obtained by calling GAO 202-512-6000 or on http://www.gao.gov).

Wolf, F.M. (1986). *Meta-Analysis: Quantitative Methods for Research Synthesis.* Newbury Park, CA: Sage Publications.

Other

Office of Personnel Management (1999). *HRM Accountability Systems Development Guide*

TABLE OF CONTENTS

APPENDICES

A: Principles Of Evaluation

Principles Of Evaluation

1. *Comparison*: Compare the demonstration group to comparison group. Compare groups before and after project implementation.

2. *Manipulation*: Make a change in the factor of interest (i.e., implement project innovations) and hold all other factors constant.

3. *Control*: Rule out all other factors as an explanation for any relationship that is found. These other factors are the threats to internal and external validity listed in Appendix B.

4. *Generalizability*: The ability to infer from the results of the demonstration project evaluation to the larger population and assess possible application Governmentwide.

Practicum: Experimentation In Action or "How to Cure Scurvy"

To help you put these principles into practical terms, we have included the following evaluation example:

Scurvy is a disease afflicting long distance mariners (sailors) since the beginning of time. The earliest known mention of scurvy was made by the Egyptians in 1500 B.C. The ancient Greek and Roman mariners were plagued by the disease. The disease became a menace to the 17th century British Navy which reportedly lost 10,000 sailors to scurvy between 1580 and 1600. The British were desperate for a cure. On a trip from England to America the ship's doctor aboard HMS Salisbury, Dr. James Lind, decided to experiment with various cures for scurvy.

Dr. Lind's experiment included all the key principles of evaluation that were discussed in section II. The four essential elements are reviewed in the context of the scurvy experiments below.

1. Comparison

Dr. Lind was not sure of the cause of scurvy but felt it had something to do with diet. His experiment used six comparison groups. Each group's ration was supplemented by victuals (dietary supplements) that Dr. Lind hoped would cure scurvy. In this manner, Dr. Lind could monitor the progress of each group, and since each was given the same daily food ration, he felt any improvement could be attributable to the dietary supplement. Notice that all comparison groups had some form of treatment; that is there was no comparison group. Having no comparison group is sometimes necessary when withholding the treatment could be life-threatening. Because the sailors used were already sick, all sailors received some sort of treatment.

TABLE OF CONTENTS

Each groups daily ration was supplemented with the following:

Group 1 - 1 qt apple cider per day
Group 2 - 25 drops elixir vitriol (mixture of sulfuric acid and aromatic herbs)
Group 3 - 6 spoonfuls of vinegar on an empty stomach
Group 4 - an herb and spice drink
Group 5 - ½ pint sea water per day
Group 6 - 2 oranges and 1 lemon per day

2. Manipulation

Dr. Lind attempted to hold all other factors constant except for the hypothesized causal factor, the supplement. The sailors' lifestyle was essentially the same since all the sailors were ill and confined to sick bay. Since the doctor believed food to be an important part of the cure, it was necessary for him to manipulate the diet of the sailors. All sailors were given the same daily rations to ensure that the causal variable was the only factor that changed. The yummy menu is displayed below:

Breakfast - water gruel sweetened with sugar
Lunch - fresh mutton broth or boiled biscuits and sugar
Dinner - barley and raisins with wine

3. Control

The principle of control requires that all other factors be ruled out as the explanation for any observed relationship between scurvy and diet. These other factors are the threats to internal validity discussed in Appendix B. Threats to internal validity are controlled by random assignment of subjects into treatment and comparison groups. This is exactly what Dr. Lind did; he randomly assigned sailors into one of the six treatment groups.

4. Generalizability

This is the extent to which the results from the experiment can be generalized to the larger population. In this case, the larger population would be the entire human race. Since Dr. Lind conducted a controlled, randomized experiment, he felt the results from the trial could be applicable to the entire population. Notice that none of the threats to validity (see Appendix B) were violated in this trial making for generalizable results.

As I'm sure you are already aware or have been able to surmise, the experiment was a success. The sailors whose diet was supplemented by oranges and lemons were cured in a matter of days. Since apples contain a small amount of vitamin C, the group of sailors whose diet was supplemented with apple cider recovered somewhat but not enough to return to work. Luckily for the sailors aboard HMS Salisbury, Dr. Lind did not need to wait for a lengthy report to confirm what he had observed. All the sailors in Dr. Lind's care were then given citrus immediately and all made a prompt recovery.

TABLE OF CONTENTS

Although the true cause of scurvy was not to be discovered until 1932, Dr. Lind's experiment showed that consuming citrus fruits both prevented and cured scurvy. As a result of this experiment, the British Navy began to included lemon and lime juice in the daily ration of its sailors. A final bit of humor, from this study British sailors were to become known as limeys.

TABLE OF CONTENTS

B: Threats to Validity

This section provides information on the threats to validity (accuracy) of the evaluation. These are many of the things that can occur in an evaluation that lead to problems in interpreting the results.

1. Internal Validity -- the extent to which effect is attributable to the treatment and no other factors - the degree to which you can rule out alternative explanations for impact.

    - Threats to internal validity -- to rule out all other factors as explanation - eight factors which can produce effects that can be confused for experiment -- i.e., competing explanations for cause of change.

        - *History* -- Historic events that occur during the time of the program that can provide rival explanations.

        - *Maturation* -- Changes occurring in the subject over time -- changes in the subject itself not outcomes.

        - *Testing* -- The effects of taking a test on the scores of a second test -- subjects may remember questions and discuss; test may sensitize into subject area. Example: Hawthorne (placebo) effect.

        - *Instrumentation* -- Changes in the calibration or a change in the observers or scorers - a.k.a. instrument decay.

        - *Regression Artifact* -- a.k.a. regression to the mean -- cases are selected for inclusion based on their extreme scores but a propensity for a group over time to score more consistently with the group average.

        - *Selection Bias* -- effect due to uncontrolled differences in the people in the experimental and comparison groups.

        - *Experimental Mortality* -- Differential loss of subject from the experimental or comparison group.

TABLE OF CONTENTS

2. External validity -- Generalizability of results to other organizations, and/or government-wide.

    - Threats to external validity -

        - *Biased Sample* - - The sample of target units should be an unbiased sample of the targets that are or will actually be clients/covered by the program.

        - *Inconsistent Program Reproductions* -- the project should be a faithful reproduction of programs as they actually are or would be implemented in reality.

        - *Testing in an Inappropriate or Unrepresentative Setting* -- The setting or range of settings should closely resemble the actual settings.

TABLE OF CONTENTS

C: Merit System Principles and Prohibited Personnel Practices

MERIT SYSTEMS PRINCIPLES: Adapted from § 2301 (b) of the title 5 USC

1. Recruit, select, and advance on merit after fair and open competition.
2. Treat employees and applicants fairly and equitably.
3. Provide equal pay for equal work and reward excellent performance.
4. Maintain high standards of integrity, conduct, and concern for the public interest.
5. Manage employees efficiently and effectively.
6. Retain or separate employees on the basis of their performance.
7. Educate and train employees if it will result in better organizational or individual performance.
8. Protect employees from improper political influence.
9. Protect employees against reprisal for the lawful disclosure of information in "whistleblower" situations.

PROHIBITED PERSONNEL PRINCIPLES: Adapted from § 2302 (b) of the title 5 USC

1. Illegally discriminate for or against any employee/applicant.
2. Solicit or consider improper employment recommendations.
3. Coerce an employee's political activity.
4. Obstruct a person's right to compete for employment.
5. Influence any person to withdraw from competition for a position.
6. Give unauthorized preference or improper advantage.
7. Employ or promote a relative.
8. Retaliate against a whistleblower, whether an employee or applicant.
9. Retaliate against employees or applicants for filing an appeal.
10. Unlawfully discriminate for off duty conduct.
11. Violate any law, rule, or regulation which implements or directly concerns the merit principles.
12. Knowingly violate veterans' preference requirements.

TABLE OF CONTENTS

D: Steps in an Evaluation

An evaluation usually consists of these basic steps:

1. Clarify project objectives/define purpose through organizational assessments -- This is an important and thoughtful step that can be quite time consuming. The value of the project and the results depend on the degree to which the objectives and purpose, research questions and thus hypotheses, can be fully defined and operationalized.

2. Formulate an evaluation plan -- Note that these steps depend on clear and precise objectives and purposes.

   - Develop measurable indicators - operationalize.
   - Choose an evaluation strategy - design for data collection/analysis.
   - Outline procedures - how evaluation will be implemented - operations guidelines.
   - Identify data sources (surveys, focus groups, workforce data).
   - Weigh strengths and weaknesses of the evaluation approach.
   - Consider costs, constraints, and resources.
   - Choose a strategy.

3. Collect baseline data from experimental and comparison group.

4. Implement project innovations, and assess implementation.

5. Obtain periodic data to monitor the project operations and determine need for mid-course adjustments.

6. Obtain summative information and final data.

7. Compare actual final results to expected results.

8. Make decisions regarding the impact of the innovations and the project outcomes.

9. State conclusions, assess generalizability.

TABLE OF CONTENTS

E: Comparison Groups

COMPARISON GROUPS USED IN PAST DEMONSTRATION PROJECTS		
Project	**Key Innovations**	**Description of Comparison Groups**
Agriculture Department	Streamlined Hiring Process - Categorical Ranking	Eighty-one comparison sites were used. The 222 Agricultural Research Service (ARS) and Forest Service (FS) locations were randomly assigned to demonstration and comparison groups. Seventy ARS and 71 FS locations became demonstration sites. Fifty ARS and 31 FS locations became the comparison sites. When the project was extended, all sites became demonstration project sites, and the Bureau of Land Management and the National Institutes of Health were used as comparison groups.
National Institute of Standards and Technology	Pay for Performance in a Broadband Framework	Four criteria were used for selecting comparison groups. They were nature and mission; job series and levels of personnel; geographic location; and practical considerations (e.g., likelihood of cooperation, data availability). Data from three sites were combined to form one comparison group; The Institute for Telecommunications Science, Office of Oceanic and Atmospheric Research, and the Mountain Administrative Support Center.
Navy Department "China Lake"	Pay for Performance in a Broadband Framework	Two comparison groups were chosen for "similarity in mission, size, and staffing." The Naval Air Development Center (Philadelphia) and the Naval Surface Weapons Center (Dahlgren, VA & White Oak, MD) were chosen as comparison groups.
McClellan AF Base "Pacer Share"	Gainsharing	The Air Force has five Air Logistics Centers. One of them was the experimental site so the others were used as controls.
Air Force Department	Contribution-Based Compensation	Uses a "constructed" comparison group from similar occupations in civilian agencies as well as comparing to other DoD Labs.
Commerce	Pay for Performance, Broadbanding	Selected organizations not participating in the demo.
DoD Acquistion Workforce	Pay for Performance, Broadbanding	Selected organizations not participating in the demo.

TABLE OF CONTENTS

F: Data and Sources

DATA AND SOURCES	
I. Individual Data	**Sources**
Workforce Data: Data collected on each employee; e.g., personnel actions, salary, race, gender, performance appraisal, etc.	Agency Personnel Records or Central Personnel Data Files (CPDF)
Employee Attitudes & Customer Satisfaction Surveys: Periodic attitude surveys are used to assess perceptual changes resulting from the personnel changes, e.g., pay satisfaction, job satisfaction.	Employee Survey Customer Satisfaction Survey
II. Aggregate Data	
Consolidated Workforce Data: Analyzed for changes in workforce characteristics and compensation trends; e.g., number of personnel on-board, racial composition of workforce, occupational composition of workforce, pay level, performance appraisal distribution, average salaries, salary cost comparison, etc.	Agency Personnel Records or CPDF
Budget/personnel data: payroll, salary, leave usage, etc.	Agency Budget and Personnel Records
III. Agency Core Outcome Measures	
Performance measures, organizational outcome measures: e.g., Veterans Affairs uses a Balanced Scorecard to track the following GPRA measures; cost per claim, employee development; customer satisfaction; accuracy; and speed.	GPRA, Strategic Plan Tracking Documents, Agency Personnel Records
IV. Qualitative Data	
Focus Groups, Structured Interviews, and Case Studies: These data are especially useful in identifying problems in implementation, why/how innovations are/are not working as planned.	Collected On-Site

TABLE OF CONTENTS

G: Central Personnel Data File (CPDF) Data Elements

Social Security Number
Date of Birth
Gender
Race or National Origin - 16 categories
Handicap Status - 85 categories
Education Level - 22 categories
Veterans Preference - 6 categories
Supervisory Status - 5 categories
Agency/subelement
Duty Station
Personnel Office Identifier
Service Computation Date
Type of Appointment - 17 categories
Federal Labor Standards Act (FLSA) Category (exempt, nonexempt)
Position Occupied (e.g., competitive service, SES) - 4 categories
Work Schedule (e.g., full time, part time) - 11 categories
Occupation - job series number
Professional Administrative, Technical, Clerical Occupations (PATCO) - 7 categories
Functional Classification (e.g., research, planning, construction, etc.) - 20 categories
Pay Plan (e.g., General Schedule) - 100 categories. Each demonstration project testing pay will be given its own code prior to implementation.
Grade - 15 categories
Step - 10 categories
Rating of Record - 7 categories
Basic Pay - the actual dollar amount is provided
Locality Adjustment Flag - records whether employees receive locality pay, (yes, no)
Locality Pay - records the actual amount
Locality Pay Area - 34 categories
Adjusted Basic Pay
Pay Rate Determination (e.g., special rate, retained pay) - 20 categories
Cost of Living Allowance (COLA)
Retention Allowance
Staffing Differential
Supervisory Differential Flag (yes, no)

TABLE OF CONTENTS

H: Public Domain Survey Modules

We have created topical modules of questions from several Governmentwide surveys as a starting point in developing survey questions. Contact your OPM demonstration project manager for more information.

ACCOUNT	Accountability
BACKGRD	Background/demographics
CLASS	Classification
COMMIT	Organizational commitment
COMMUN	Organizational communication
CUSTOMER	Customer service
DIVERSE	Diversity in the workplace (Includes harassment and merit principles)
EMPOWER	Empowerment
FLEX-AWS	Flexitime and Alternative Work Schedule (AWS)
FLXPLACE	Flexiplace
HIRING	Hiring
HRCHANGE	Human resources changing roles
JOBSATIS	Job satisfaction
LABORREL	Labor relations
MERITPRN	Merit principles
ORGCHG	Organizational change
ORGCULT	Organizational culture
ORGPERF	Organizational performance
ORGT&D	Organizational training & development
PARTMGT	Participative management
PAY	Pay (Includes pay satisfaction, pay administration, internal-external equity, pay for performance, FEPCA)
PERFAPPL	Performance appraisal
PERFMGT	Performance management
PMEval	Personnel management evaluation
RATESUPV	Employees rating supervisors
STRATPLN	Strategic planning
SUPV	Supervisory authority/duties
TRAINING	Training
TQM	Total Quality Management (TQM)
TQM/SURV	DoD TQM Survey
TURNOVER	Turnover/retention
WORKUNIT	Workunit performance
WORKGRP	Workgroup performance

TABLE OF CONTENTS

I: Sample Site Historian Log

SAMPLE SITE HISTORIAN LOG	
Date	**Recorded Event**
5/6/97	A reduction in force occurred resulting in 80 demonstration project employees receiving RIF letters
7/9/97	Budget cutbacks resulted in 10% funding loss for the Demo
8/9/97	Hiring freeze until 9/98
10/97	100 transferred employees from HR division moved into demo

TABLE OF CONTENTS

J: Sample Evaluation Models

Outcome Evaluation Model -- VA Demonstration Project			
Project innovation	**Expected Effects**	**Observable Measures**	**Data Sources**
skill-based pay, with a variable pay component, in a team environment	Increase value to taxpayer	Cost per claim	Balanced scorecard
	Improve customer service	Accuracy Speed Customer satisfaction	Balanced scorecard Customer satisfaction survey
	Improve employee pay satisfaction	Employee attitudes toward pay	Employee attitude survey
	Improve employee development	Employee satisfaction Increased opportunities for developing and applying skills and competencies	Employee attitude survey Balanced scorecard Case study
	Maintain workforce diversity	Workforce composition	Workforce data

TABLE OF CONTENTS

Commerce Demonstration Project General Evaluation Model		
Expected Results	**Measures**	**Data Sources**
1-1 The quality of new hires will increase (among occupations of interest)	- new hire GPA's -quality of new hires schools -professional publications -proportion of new hires who receive awards -proportion of new hires with previous scholastic honors -highest degree earned -proportion of new hires in professional societies and hold office in those societies -number of patents -performance ratings of new hires (pre and post demo)	-new hire interviews/surveys/records -annual survey -agency records;
1-2 Hiring officials will perceive an improvement in the quality of new hires.	- supervisory attitudes toward selectees	-annual survey -focus groups
1-3 Hiring officials will see an improvement in the quality of the applicant pool.	- supervisory attitudes toward selectees	-annual survey -focus groups
1-4 Non-supervisors will perceive an improvement in the quality/KSAs of their co-workers.	- employee attitudes toward selectees	-annual survey
1-5 Applicants will be more likely to accept job offers.	-applicant acceptance rate of job offers will increase.	-agency records -focus groups
1-6 Recruiting allowances will be instrumental in attracting new candidates.	- ratio of allowance offers to allowance acceptance and declinations - selectees attitudes toward recruiting allowances	-agency records -focus groups among new hires
1-7 Staffing project innovations will have no adverse impact on diversity	- number of women, minorities, veterans, etc., hired	-agency

TABLE OF CONTENTS

Intervention Impact Model			
Intervention	**Expected Effect**	**Measure**	**Data Source**
Broadbanding	reduced vacancies	vacancy rates	personnel records
Flexible in-hire rates	reduced turnover	turnover rates turnover reasons	workforce data exit interview/survey
Training and development	increased organizational commitment	organizational commitment reduced turnover	attitude survey workforce data

TABLE OF CONTENTS

K: Demonstration Project Evaluators Used in the Past

Demonstration Project	Key Innovations	Evaluator	Years Evaluated
Department of Agriculture	Streamlined Hiring Process Categorical Ranking System of Hiring	Pennsylvania State University Institute for Policy Research & Evaluation Center for Applied Behavioral Sciences	Years 1-5
Department of Air Force	Contribution-Based Compensation System	Office of Personnel Management, Personnel Resource Division; Systems Research and Applications Corporation	Years 1-5
National Institute of Standards and Technology	Pay for Performance within a Broadbanding Framework	University Research Corporation (URC)	Years 1-2
		HumRRO	Years 3-5
		Office of Personnel Management	Years 6+
Department of Navy "China Lake"	Pay for Performance within a Broadbanding Framework	(Evaluation plan/measures developed by University of Southern California, Coopers & Lybrand and OPM)	
		Coopers & Lybrand	Year 1
		Office of Personnel Management	Years 2-6+
Department of Air Force "Pacer Share"	Gainsharing, Team, Partnership with Unions	RAND Corporation	Years 1-3
		Navy Personnel Research and Development Center (NPRDC)	Years 4-5
Federal Bureau of Investigation	Recruitment and Retention Incentives	Office of Personnel Management	Years 1-5
Federal Aviation Administration	Recruitment and Retention Incentives	Research Management Consultants, Inc. and HumRRO International, Inc.	Year 1
Commerce	Pay for Performance, Broadbanding	Booz, Allen, Hamilton	Year 1-5
DoD Acquisition Workforce	Pay for Performance		Year 1-5

Note: Inclusion of this information does not indicate an endorsement of these organizations.

TABLE OF CONTENTS

L: Demonstration Project Evaluation Checklist

DEMONSTRATION PROJECT EVALUATION REQUIREMENTS	✓
GENERAL ISSUES	
Evaluation assesses degree to which project purpose and goals are met, and if not, why not	
Evaluation assesses cost	
Evaluation tracks project implementation and operation	
Evaluation assesses impact on veterans and other EEO groups	
Evaluation assesses impact on Merit System Principles and Prohibited Personnel Practices	
Evaluation assesses degree of generalization of project to other groups or Governmentwide	
Sufficient evaluation breadth for global assessments and recommendations	
Sufficient evaluation depth to determine affect of project details	
PLANNING - PRE-IMPLEMENTATION	
OPM approved complete evaluation plan	
External Evaluator (independent of participating agencies) on board	
Determine, support, be prepared to execute strongest approach for comparison	
Training of Data Collectors prior to data collection	
Selection and Training of Site Historians	
Collection of baseline objective and perceptual/attitudinal data from demonstration project and comparison sites	

TABLE OF CONTENTS

DEMONSTRATION PROJECT EVALUATION REQUIREMENTS (Cont.)	✓
EXECUTION	
Strongest objective and perceptual/attitudinal data for all innovations and anticipated outcomes	
Longitudinal objective and perceptual/attitudinal data	
Track degree and accuracy of implementation/operation of project	
Collect Data on adherence to Merit Systems Principles, avoidance of Prohibited Personnel Practices, impact on Veterans and other EEO groups	
Collect data on GRPA-type issues such mission accomplishment and organizational productivity and effectiveness	
Collect enough data to statistically identify important effects, and report effect sizes when necessary	
Maintain Site Historians	
DATA ANALYSIS	
Descriptive Statistical Techniques that result in complete, accurate, clear, and concise presentation of data linked to critical questions or hypotheses	
Appropriate inferential procedures required to assess impact of the project	
REPORTS	
Baseline/Implementation report no later than 18 months after project Implementation	
Management and Technical Interim Decision Reports delivered no later than 9 (nine) months prior to project expiration	
Management and Technical Summative Reports delivered no later than one year after original project expiration date	
Intervening updates on status of evaluation, project operation, impact on veterans, EEO, Merit Systems Principles and Prohibited Personnel Practices	
Electronic version of data provided to OPM	