Social Innovation Fund Evaluation Reporting Guidance

Feasibility, Implementation, and Impact Study Reports

Effective October 1, 2016

NOTE TO READERS

This is the first draft of this document. Corporation for National and Community Service (CNCS) asks Social Innovation Fund (SIF) grantees and their partners to share any comments, suggested edits and feedback they have on this product to Andrea Robles (arobles@cns.gov) by April 1, 2017.

ACKNOWLEDGEMENTS

This document was completed under supervision and oversight of Lily Zandniapour, Ph.D. (lzandniapour@cns.gov) and Andrea Robles, Ph.D. (arobles@cns.gov) in the Office of Research and Evaluation (ORE) at CNCS. CNCS would like to thank JBS International for the development of this document under Contract Number GS-10F-0285K Order Number CNSHQ11F0050 for the Evaluation Review and Monitoring for the Social Innovation Fund Project. In particular, Nicole Vicinanza, Ph.D., Peter Lovegrove Ph.D., Leith Lombas Ph.D, Rebecca Frazier Ph.D, Claudia Birmingham, and Stacey Houston. were instrumental in the development of the content of this document. Their work was informed by their experience in reviewing study reports submitted for the evaluations commissioned under the Social Innovation Fund Program. Others who contributed to the development of the content include Darcy Strouse, Ph.D., Laura Sosinsky Ph.D., and Heide Jackson Ph.D. Beth Slater Ph.D. from ORE/CNCS, and Pamela Dubitsky from SIF/CNCS reviewed the document and provided edits and comments. Chantaline Tetu formatted the final version of the document. We thank them all for their insights and contributions.

SUGGESTED CITATION

Corporation for National and Community Service. (2016) S*ocial Innovation Fund Evaluation Reporting Guidance: Feasibility, Implementation and Impact Study Reports*. Washington, D.C.: Author.

TABLE OF CONTENTS

TABLE OF CONTENTS

INTRODUCTION

The program models funded by SIF grantees must produce rigorous evaluative evidence that help explain why programs are successful and how they may be improved and/or expanded. The evidence produced by these evaluations also builds the existing evidence base for these and similar programs addressing these issues.

To this end, it is the expectation of CNCS that each grantee or subgrantee will conduct and report on an impact evaluation for each program model targeting a specific level of evidence by the end of the five-year SIF grant period. SIF evaluations typically represent a substantial level of effort and resources, and it is expected that the evaluation reports produced by SIF intermediaries and their subgrantees reflect program learning from the SIF experience, address program impacts in a rigorous way, and contribute to the broader understanding of effective programs addressing community needs.

At the final reporting stage, SIF impact studies are assessed based on a range of criteria including:

- level of study rigor (i.e., final level of internal validity, external validity, and overall quality of study design implementation), and
- Study impact findings on confirmatory (and possibly other) outcomes.

The evaluation reports SIF intermediaries and subgrantees produce are a key component of SIF's efforts to disseminate findings regarding rigorously documented program impacts and sharing lessons learned about effective programming and building evidence. This document is intended to promote a shared understanding among CNCS, SIF grantees, SIF subgrantees, and evaluators about how evaluation reporting can support these purposes. CNCS will provide reviews and request revisions of reports, if needed, to ensure that these reports: 1) address the questions for the approved SIF Evaluation Plan (SEP), 2) provide sufficient information to document whether the targeted level of evidence for the approved SEP was reached, and 3) share SIF findings with stakeholders and broader audiences. Final reports require review and acceptance by both the grantee and CNCS prior to the public distribution of findings representing SIF work.

I. REPORTING GUIDANCE OVERVIEW

The SIF Evaluation Reporting Guidance Document serves as a guide and resource for developing evaluation reports for the SIF program. This guidance is intended to be used by SIF grantees, subgrantees, and evaluators in preparing, reviewing, and finalizing evaluation reports. Since each of these audiences may have a different level of evaluation knowledge, non-technical terms are used where possible, and technical terms are defined throughout the document and in the glossary.

The evaluation reports that SIF grantees and subgrantees produce are critical to both documenting the level of evidence attained by SIF evaluations and sharing the knowledge that SIF builds. CNCS understands that evaluation reporting is not a one-time effort, but occurs throughout the evaluation process over the course of several years. This document provides SIF grantees, subgrantees, and evaluators with support in developing evaluation reports that align with their SIF Evaluation Plans (SEPs). It also provides the checklist of criteria that CNCS will use to review each report for completeness and quality. The checklists are included in Appendix A1.

CNCS sees the process of ongoing program evaluation and knowledge building as a key aspect of the SIF that can improve grantee and subgrantee programs, while also benefiting other organizations throughout the nonprofit and public sectors. The agency is committed to supporting grantees in their efforts to increase the evidence of their programs' effectiveness within their SIF Fund portfolios. CNCS works closely with grantees by providing them with technical assistance on documenting and disseminating the evidence of effectiveness for each program model within the SIF.

The SIF Reporting Guidance Document is comprised of five main parts.

Part	Topic	Component
Part I	Reporting Guidance Overview	- Describes the goals and sections of this document
Part II	Types of SIF Evaluation Studies	- Describes the three main types of evaluation studies - Provides suggested report outlines for each type of study
Part II	Types of SIF Evaluation Reports	- Presents an overview of the two different SIF report categories—interim reports and final reports
Part IV	Detailed Content Guidance for Reports	- Key reporting requirements for major sections of evaluation reports
Part V	Appendices	
	Appendix A: Templates and Tools	- A full report checklist - Guidance on reporting evaluations results in tables and figures
	Appendix B: Assessing Outgoing Level of Evidence	- Sample materials used by CNCS external reviewers to assess a study's achieved level of evidence
	Appendix C: Reference and Resource List	- Detailed list of references and resources supporting the SIF evaluation design, implementation, and reporting process.

Two additional companion documents to this document include the:

1. SIF Evaluation Plan (SEP) Guidance Document, which provides information on how to develop a rigorous evaluation plan; and,
2. SIF Evaluation Research Glossary, which provides definitions of key technical and SIF-specific terms.

While each of these documents were primarily developed SIF-supported organizations and partners, they have wider application. These resources describe the components involved in rigorous evaluation planning and reporting and as such, can be used by the broader social and nonprofit sectors.

II. TYPES OF SIF EVALUATION STUDIES

This section briefly describes the purpose of different types of SIF evaluation studies --implementation, feasibility, and impact -- and provides detailed outlines for reporting on each type of study. Part III of this guidance document outlines key information that should be provided when reporting on each type of evaluation study.

Implementation Study

Purpose

The purpose of an implementation evaluation is to assess if and the degree to which a program is delivered as intended. In particular, implementation studies discern how closely the program theory and intended procedures align with actual program practice. Aspects of program implementation that may be studied are program enrollment as well as specific dimensions of fidelity such as dosage, quality, differentiation, and responsiveness. Unlike outcome and impact evaluations which focus on the results, implementation evaluations focus on the process by which a program uses its resources to provide services to the target population and accomplish program objectives.

In assessing the alignment between program theory and program delivery, implementation studies capture how well the program reaches the appropriate target population. The evaluation includes specific methods and measures to quantify and assess the degree to which each aspect of the program was delivered by program staff and received by clients. If a comparison or control group is included in the evaluation plan, it is important that the implementation evaluation assess whether or not the comparison or control group received portions of the program or similar services offered by other programs in the area.

Implementation evaluation reports also describe the data collection process, with regard to program services provided to both program participants and the control or comparison group members, if applicable. The report should include more specifics regarding the impact evaluation such as the types and sources of data such as how they will be collected (e.g., surveys, interviews, focus groups) and how the data collected will be analyzed and used.

Suggested Outline

Below is a basic report outline for an Implementation Evaluation Report. Part III details the content that should be reported for main sections of the outline.

Executive Summary

1. Introduction
    - A. Program Background and Problem Definition
    - B. Overview of Prior Research
    - C. Overview of Implementation Study as detailed in the SEP (including design/approach and methodology used)
    - D. Research Questions
        1. Impact questions and findings to date
            - a) Confirmatory
            - b) Exploratory
        2. Implementation questions addressed in this report
    - E. Contribution of the Study
        1. Level of Evidence Targeted by the Impact Study
        2. Strengths and Limitations of the Implementation study
        3. Connection of this Implementation Study to Future Research

2. Study Approach and Methods
    - A. Implementation Study Design
    - B. Sampling (if applicable), Measures, and Data Collection by Dimensions of Implementation
        1. Fidelity to Program Design
        2. Program Exposure (Dosage)
        3. Program Quality
        4. Program Participant Responsiveness
        5. Program Differentiation
        6. Participant Satisfaction

3. Analysis Method for Assessing Implementation

(For each dimension of implementation, describe analysis method, e.g., t-tests/chi-square, correlation, multiple regression)

- A. Fidelity to Program Design
- B. Program Exposure (Dosage)
- C. Program Quality
- D. Program Participant Responsiveness
- E. Program Differentiation
- F. Participant Satisfaction

4. Findings, Lessons Learned, and Next Steps
    - A. Findings by Research Questions (including outputs and preliminary information on outcomes, if applicable; include findings for each implementation dimension)

5. Study Logistics Updates (May be included as an appendix or an accompanying memo.)
    - A. Protection of Human Subjects
    - B. Budget and Timeline
    - C. Evaluation and Program Staff Involvement

Feasibility Study

Purpose

A feasibility study may be conducted prior to carrying out a more rigorous evaluation. A feasibility study provides information on what types of evaluation strategies might work well with the program as it currently operates and any identified barriers that might need to be overcome for the impact evaluation to be conducted, such as time, human resources, and/or budget constraints. Feasibility studies may also include assessments of other resources that are available to maintain or improve the program implementation, and the ability of the program to successfully engage in a rigorous implementation and/or impact evaluation.

Suggested Outline

Below is a basic report outline for a Feasibility Study Report. Part III details the content that should be reported for main sections of the outline.

Executive Summary

1. Introduction
    - A. Program Background and Problem Definition
    - B. Overview of Prior Research
    - C. Overview of Impact Study as detailed in the SEP (including design/approach and methodology used)
    - D. Research Questions
        1. Impact questions (and findings to date, if applicable)
            - a) Confirmatory
            - b) Exploratory
        2. Implementation questions (and findings to date, if applicable)
    - E. Goals of the Feasibility Study
        1. Evaluability Assessment (research questions addressed in this report)
        2. How the Findings will Support the Impact Study
    - F. Contribution of the Study
        1. Level of Evidence Targeted by the Impact Study
        2. Strengths and Limitations of the Study
        3. Connection of the Study to Future Research

2. Study Approach and Methods
    - A. Feasibility Study Design
    - B. Sampling, Measures, and Data Collection
        1. Sampling (assignment to treatment and counterfactual groups, if applicable)
        2. Measures and Instruments
        3. Data Collection Activities
        4. Suitability of (or modifications to) the Planned Sample, Measures, and Data Collection Activities

3. Data Analysis
   - A. Unit of Assignment and Analysis
   - B. Analysis Approach as applicable (e.g., Process, Outcome, Impact Intent-to-Treat, Impact Treatment-on-Treated)
   - C. Formation of Matched Groups (if applicable)
   - D. Treatment of Missing Data (if applicable)
   - E. Analysis Model/Type (specific statistical approaches and assumptions if applicable)

4. Findings, Lessons Learned, and Next Steps
   - A. Assessment of the Feasibility of the Planned Impact Evaluation
   - B. Findings by Research Question
   - C. Preliminary Outputs and Outcomes
   - D. Barriers to Achieving High Levels of Internal and External Validity
   - E. Recommended Modifications to Measures and Procedures

5. Study Logistics Updates (May be included as an appendix.)
   - A. Protection of Human Subjects
   - B. Budget and Timeline
   - C. Evaluation and Program Staff Involvement

Impact Study

Purpose

An impact evaluation provides an understanding of how program components are related to any changes in the condition(s) the program seeks to change among its beneficiaries. Ideally, an impact evaluation provides evidence about whether the observed changes in the treated condition might be credited to the program.

The evidence attained by an impact evaluation is determined by the extent to which it maximizes both internal and external validity. Internal validity refers to the ability of the evaluation findings to accurately reflect the impact of the program on participants or beneficiaries. Research designs are thought to have good internal validity if they incorporate design features that are effective in ruling out other plausible explanations for the changes in the outcome(s) targeted by the program. External validity refers to the degree to which a study's findings can be generalized to a diverse target population; external validity also pertains to diversity across time (i.e., the program is effective over several years) and geographic location (i.e., the program is effective in different places).

The stronger the level of evidence, the greater the degree of confidence that the impact findings can be attributed to the intervention, and the greater the extent to which the evaluation results can be applied to groups other than those in the evaluation.

The choice of an Impact Evaluation Design is key to developing an understanding of how the program selected for the evaluation maximizes the internal and external validity of the study.

Suggested Outline

Below is a basic report outline for an Impact Evaluation Final Report. Part III details the content that should be reported for main sections of the outline.

Executive Summary

1. Introduction
   - A. Program Background and Problem Definition
   - B. Overview of Prior Research
   - C. Overview of Impact Study (including design/approach and methodology used)
   - D. Research Questions
     1. Impact questions and findings
        - a) Confirmatory
        - b) Exploratory
     2. Implementation questions and findings
   - E. Contribution of the Study
     1. Level of Evidence Generated by the Study
     2. Strengths and Limitations of the Study
     3. Connection of this Study to Future Research

2. Study Approach and Methods
   - A. Implementation Study Design (brief summary)
   - B. Impact Study Design
     Description (e.g., Randomized Between Groups (Experimental) Design, Between-Groups Design- Formed by Matching Design, Between-Groups Design- Formed by Cut-off Score (RDD), Single Group Design, Interrupted Time Series Design, Non-Experimental Design), including strengths and limitations, internal and external validity
   - C. Sampling, Measures, and Data Collection (For each design above included in the report)
   - 1. Sampling
   - 2. Measures and Instruments
   - 3. Data Collection Activities(e.g., timing, processes for each data source)

3. Statistical Analysis of Impacts
   For each research question addressed in the current report describe:
   - A. Unit of Assignment and Analysis
   - B. Analysis Approach (e.g., Process, Outcome, Impact Intent-to-Treat, Impact Treatment-on-Treated)
   - C. Formation of Matched Groups
   - D. Treatment of Missing Data
   - E. Analysis Model/Type (i.e., specific statistical approaches and assumptions)
     1. Tests for Statistical Significance
     2. Adjustment for Multiple Comparisons (if applicable)
     3. Assessment of Effect Sizes

4. Findings, Lessons Learned, and Next Steps

5. Study Logistics Updates (May be included as an appendix.)
   - A. Protection of Human Subjects
   - B. Budget and Timeline
   - C. Evaluation and Program Staff Involvement

III. TYPES OF SIF REVIEWS

SIF evaluation studies should produce reports addressing evaluation progress and findings to date every 12 months after beginning data collection and upon completion of an evaluation study. This provides CNCS with information about the intervention and whether the evaluation is on track to attain, or has attained, its targeted level of evidence. There are two main categories of SIF reports: interim and final.

Interim Reports

Interim reports may occur annually and timing may vary from year to year depending on the needs of the evaluation. All evaluations should provide at least one report per year that addresses findings (including any implementation or early impact findings) from the evaluation to date, any challenges or lessons learned regarding the evaluation itself, and any lessons learned regarding the intervention being evaluated. This report should provide sufficient information to allow CNCS to determine whether or not the evaluation is on track to attain its targeted level of evidence. Appendix C of this document contains the Progress Review Form that CNCS uses to review reports to determine whether the evaluation is on track to attain, or has attained, its target level of evidence. The Report Review Form for interim and annual reports captures and provides feedback on progress in key study areas and it is organized by the type study being reported on (Feasibility, Implementation, Impact).

Final Reports

Once the study has been completed, SIF evaluation studies should produce final evaluation reports documenting the purpose, background, methods, conduct, and findings for each completed impact, implementation, and feasibility study. A final evaluation report should provide sufficient information to allow CNCS to determine how well the study has been conducted, lessons learned, and, for impact studies, whether the program model has achieved its targeted level of evidence[1]. The subsequent sections of this document outline the content that should be included in final reports for completed SIF evaluation studies.

[1] Final Evaluation Report Review Forms that CNCS uses to review final reports for completed impact, implementation, and feasibility studies are available as separate documents.

IV. DETAILED CONTENT GUIDANCE FOR INTERIM AND FINAL REPORTS

This section identifies key information needed when reporting on SIF evaluation studies in both interim and final reports. Based on the type of study being reported on and the type of design employed, not all sections of this guidance will be relevant for a given report.

The report content guidance is organized by major sections of an evaluation report. The required content for each section maps to the items that CNCS will use to review and assess the overall quality of the study.

Throughout this part, the sidebars contain two different sources of information:

- Checklist items that CNCS will use to review and assess the quality of the report and,
- Additional resources related to that section.

Report Review Checklist: Executive Summary

The following items are briefly described:

- The names of the Grantee, Subgrantee, and evaluation contractor, and the years during which the study took place
- The program and intended outcomes/impacts
- Relevant prior research
- The targeted level of evidence
- The evaluation design, including comparison/control group approach
- The measures/instruments
- The analysis approaches used
- The research questions addressed and key findings
- Key updates related to evaluation timing/timeline and budget.
- Key changes to the program or evaluation team
- Key next steps for the evaluation and/or program

Executive Summary

Description

The Executive Summary of your report should give the reader an overview of the program, the context of the evaluation approach you used, the research questions you addressed, and your key findings. As with any Executive Summary, its purpose is to highlight what is in the full document, either to brief very busy readers or to serve as an overview to better prepare readers for the full document.

Section Content

The Executive Summary should contain the following details in the order presented.

Identify the names of the Grantee, Subgrantee (if applicable), SIF Cohort (years during which study took place) and evaluation contractor. Include names of the organizations and program sites involved in the evaluation.

Provide a summary of the program and intended outcomes/impacts. Include a one-paragraph synopsis of the program and intervention, target and actual population served (including total numbers served by the program, number of sites, etc.) and what it intends to change/impact.

Identify relevant prior research. Provide a brief synopsis (one to two sentences) of the prior research done on the program.

Identify the targeted level of evidence. Identify the targeted level of evidence (e.g. preliminary, moderate, and strong) and briefly describe why the level was targeted and how the current study advances the evidence base for the intervention.

Introduce the evaluation design, including comparison/control group approach. Provide a brief summary of the evaluation design and methods as implemented, and the final/analytic sample size numbers of both treatment and comparison, control, or other counterfactual groups.

Identify the measures/instruments. Describe the data sources, measures/instruments used, and the types and amount of data collected from them.

Summarize the analysis approaches used. Give a brief description of the analysis approach (es) used to answer the research questions the current report addresses.

List the research questions addressed and key findings. Highlight evaluation findings (expected and unexpected) and high level answers to the research questions the report addresses. Briefly address the implications, any recommendations, and lessons learned.

Identify key updates related to the evaluation timeline, budget, program or evaluation team. If applicable, provide a summary of any challenges related to or deviations from the key timeline elements/dates (e.g., dates for participant recruitment and data collection, analysis and reporting), evaluation budget, evaluation team, or program team.

Provide key next steps for the evaluation and/or program. Describe how the evaluation and/or program will proceed following this reporting period.

Report Introduction

Description

The introduction to the report should establish the context for the evaluation report. The introduction can help frame what might be expected from this and future evaluations of the program. Although the level of detail needed in the introduction may vary based on the type of report and intended audience, it should be complete enough to allow a reader who has not read the SEP or any earlier reports to have sufficient context to understand the findings and implications of the report. The introduction should also note when there have been major changes to the evaluation or the program from the SEP.

The introduction should note the type of evaluation(s) undertaken (implementation, feasibility, and/or impact), type of report (interim or final evaluation report)and intended audience(s), provide a brief summary of the program's theory of change, prior research, a description of the problem, intervention and participants, the research questions (implementation and impact), and the level of evidence generated by the findings.

Report Review Checklist: Introduction

- The type of evaluation, type of report, and intended audience are identified.
- The theory of change and prior research are briefly discussed, including previous level of evidence.
- The program model is briefly described, including key information such as the number of participants, inputs, components/ activities, and key outcomes.
- The targeted level of evidence for the current study is described with specific justification.
- Program implementation questions are clearly stated.
- Program impact questions (confirmatory and exploratory) are clearly stated.
- Changes to the SEP are reported.

Section Content

The Introduction section should include several details:

- Identify the type of evaluation, type of report, and intended audience. Specify what type of report is being submitted, an interim or final report. Indicate the type of evaluation(s) undertaken (an implementation, feasibility, and/or impact evaluation), and detail the intended audience(s) for the report (e.g., program staff, funders, the public).

- Discuss the theory of change and prior research, including previous level of evidence. Briefly discuss the problem the intervention is designed to address, how the program was developed, the theory of change, and any prior research relevant to the report, including the previous level of evidence for the intervention.

Describe the program model briefly, including the number of participants, inputs, components/ activities, and key outcomes:

- Provide an overview of the intervention activities and their components, as well as the inputs (i.e., program staff, funding, other resources) and activities that support the program and how they are delivered to program participants. Also, define the program unit (e.g., sessions, classes, visits).
- Briefly discuss the outputs, outcomes, and impacts the program targets and note which ones the current report addresses.
- Describe the target community with respect to the need the intervention is designed to address, as well as the population in need of support and relevant demographic characteristics. If the program was previously evaluated, be sure to include a discussion of any differences between the populations being served in the current study compared to previous studies.
- Note the number of: a) clients being served by the program and who are participating in the evaluation (i.e., the number of individuals receiving intervention services in the treatment group(s); and, b) individuals who are participating in the study across all study groups, both treatment and counterfactual.

Describe and justify the targeted level of evidence for the current study. Note the targeted level of evidence (e.g., preliminary, moderate, strong) and describe how the current study advances the evidence base for the intervention. Provide specific justification as to how the target level of evidence has been, or will be, achieved using the current study design.

Clearly identify the program research questions (i.e., implementation and impact questions). The Introduction should align the current report with the approved research questions (implementation and impact) as detailed in the SEP. In a numbered list, bulleted, or table format, document the specific research questions the study is attempting to answer, or has already answered, and how it is hoped the evaluation findings will be used following the study.

Clearly state whether questions are **implementation** questions, **confirmatory impact** questions, and **exploratory impact** questions.

If the report is not intended to address all the research questions from the SEP (e.g., it is a report on implementation or impact findings only), note the questions that the current report intends to address, but also list other questions that have been, or will be, addressed in past or future reports. Findings for research questions answered in earlier reports should be briefly summarized in the introduction to provide context.

Document whether there have been any changes to the SEP. Note if the program model or evaluation design has undergone any changes since the evaluation plan was approved, the justification for the changes, and how these changes influence the evaluation. If there have been no changes, please state that no changes have occurred.

Implementation Study Reporting - Design, Methods, and Findings

An Implementation Evaluation Report, or section in a comprehensive Final Report, should include the following details in the order listed in the Report Review Checklist; if the Implementation Evaluation Report is already complete as a stand-alone report, the Final Impact Study Report should include a summary of the Implementation Study approach, methods, and findings sufficient for the reader to understand program implementation and how Implementation Study results are related to effects in the Impact Study.

Report Review Checklist: Implementation Evaluation

- Study design and procedures for measuring program implementation in the treatment group are presented.
- Details are provided for how each implementation dimension was measured, including target levels.
- Measures are clearly described (or included in an appendix), including a description of the construction and validation of all measures.
- Analysis method and procedures for assessing implementation are described.
- Any preliminary or final implementation analysis findings are detailed.
- Lessons learned from implementation results are discussed.
- Changes to the SEP are reported.

Describe the study design and procedures for measuring program implementation in the program group. Present the study design, approach, and methodology used to assess implementation. This may focus on assessing whether the program, in practice, matches the theory of change that was generated to develop the program and address the issue of concern, the program components clients engage in through participation in the intervention, the dosage program clients receive, or other dimensions of implementation.

Detail on how each implementation dimension was measured, including target levels, if appropriate. Provide details of how fidelity and each related dimension (enrollment, exposure/dosage, quality, responsiveness, differentiation, and satisfaction) was measured. Discuss the sources of data, amount of data collected, such as the number of surveys completed, the number of interviews completed, and, if using focus groups, the number of groups and the number of participants in each group. In addition, it is important to include a description of the sample. Be sure to include details such as the size of the sample, rationale for targeting the sample, and the demographic composition of the sample, particularly with respect to variables that are being used to measure the characteristics of the treatment and comparison groups. If ongoing, include the current and targeted sample sizes. The report should also include a description of the data collection procedures, such as who collected the data, when it was collected, and how the data was collected, processed, stored, and analyzed. Lastly, this section of the report should include the measures used for each dimension of the implementation study, and include target levels, where appropriate.

Clearly describe measures (or include in an appendix), including a description of the construction and validation of all measures. Provide information on appropriateness of the measures for the study population, including how validity and reliability of the measures were tested and the results of those tests. If instruments were constructed for the study, describe how they were developed and piloted.

Describe the analysis method(s) and procedures for assessing implementation. Identify the research study design(s), approach, and methodology, the types of qualitative analyses and/or quantitative statistical analyses the study employed (e.g., t-tests, chi-square, correlation, multiple regression), and the procedures the evaluation team used to perform the analyses.

Detail any preliminary or final implementation analysis findings. Address the program delivery context and include a full program and study timeline that traces the program delivery process, including the components and activities that take place during each step in the implementation process. Also, the report should list the program outputs and whether or not the program reached its output goals.

If applicable, include any preliminary information on outcomes and impacts, and relate how the implementation results are being used to support the impact study. It is extremely important to note any changes to the impact study design that resulted from implementation study findings, including any changes to the control or comparison group.

Discuss lessons learned from implementation results. Include key findings and lessons learned from the implementation process. It should also note any lessons learned through implementation that might strengthen the evaluation.

Document whether there have been changes to the SEP. Any major changes made to the implementation evaluation proposed in the SEP should be discussed. If there have been no changes, please state that no changes have occurred.

Feasibility Study Reporting - Design, Methods, and Findings

Reports for feasibility studies should address implementation research questions posed in the SEP that are designed to assess how the program actually works on the ground in comparison with the program theory and logic model.

Feasibility reports should include the following details outlined in the Report Review Checklist.

Describe barriers to proposing a design with the potential to contribute to strong or moderate evidence. For example, consider:

- Whether program implementation has sufficient fidelity, including how a program was implemented if the implementation differed from the original program theory and logic model, to provide useful information for understanding the feasibility of future evaluation efforts.
- Whether the program has sufficient effective service delivery (the program can be delivered to enough people, with sufficient intensity) for future evaluations to be feasible.
- Ability of the program to recruit and serve a population that is relatively large and representative of the program's target population.
- Ability of the evaluation to recruit an appropriate comparison group of sufficient size.
- The ability of measures and data collection instruments to assess the intended outcomes with the program and the population. (This may include the feasibility of collecting data on preliminary and/or intermediate outcomes or impacts, as well as final outcomes or impacts.)

Report Review Checklist: Feasibility Study

- Barriers to proposing a design with the potential to contribute to strong or moderate evidence are described.
- Full study design is clearly and comprehensively explained.
- Description of the treatment and counterfactual groups are included.
- Where appropriate, assignment of study participants to groups is described.
- The instruments or processes tested are described.
- How this study will lead to an impact evaluation yielding moderate to strong evidence during the SIF timeframe is described.
- Changes to the SEP are reported.

Clearly and comprehensively explain the full study design. Include a full description of the current study design and how it assesses the feasibility of future evaluation efforts. Specify any barriers to achieving the targeted level of evidence and how subsequent evaluation efforts will support reaching evidence goals during the SIF timeframe.

Describe the treatment and counterfactual groups. Describe if data collection took place as outlined in the SEP. Include information on whether the data were collected from the proposed treatment population indicated in the SEP, or if the data included in the study came from a different population. For example, an SEP may have proposed collecting data on children aged three to five years, but the evaluation team may have found that six-year-olds were included in the program, which would expand the sample population. Also, include information regarding the representativeness of the treatment group from which data were collected. Indicate if

Additional Resources

For more information on various designs for feasibility studies, see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2859314/

the sample differed from the population targeted by the program and any effect that may have had on how the results of the feasibility study will apply to the full study.

If appropriate, also describe any potential counterfactual group(s); include such information as the size of the group, characteristics of that group and potential confounds/nonequivalence when lined up with the program group, any overflow of the program to the control/comparison group, and similar services that the comparison participants could receive from other local programs separate from the SIF funded program. Additional information about the program and comparison samples may include, but is not limited to, demographic and socioeconomic data, information on participants' histories or past utilization of services, and specifics particular to the program's services that might inform an impact analysis, such as previous health conditions or educational test scores.

Where appropriate, describe assignment of study participants to groups. If applicable, describe how individuals or groups were assigned to the treatment and counterfactual groups. If selection criteria were used to determine eligibility for program services, it is important to describe what criteria worked best and why.

Describe the study instruments or processes tested. Provide information on the development or alteration of any data collection instruments and techniques used in the feasibility study. Describe who completed the data collection activities (i.e., program staff, evaluation team staff, others), how data were collected, at what points in time, and over what period of time. If secondary or administrative data were used in the study, describe how those data were collected, by whom, at what points in time, and over what period of time. Indicate if data were collected only for the individuals who received program services, or if non-program participant data were also collected.

In addition to describing the measures and data collection techniques, indicate if the measures and techniques proved suitable for use in the planned impact evaluation. Relate any problems encountered in the process that would prevent these measures and activities from being used in future evaluations that could yield moderate or strong evidence.

Discuss how this study will lead to an impact evaluation yielding moderate to strong evidence during the SIF timeframe. Include findings from the feasibility study. The findings should reflect the research questions posed in the SEP. Include any (preliminary) outputs or outcomes for the sample that received services and for the comparison group, if applicable. Specify how the findings from the feasibility study support future evaluation efforts and form the basis for undertaking an evaluation with the potential to yield moderate or strong evidence during the SIF grant timeframe

The report should provide an assessment of the overall feasibility of the planned impact evaluation for the program, including aspects of the feasibility study that proved promising and effective (i.e., measures, data collection, analysis techniques). Ultimately, the feasibility study should report on the program's ability to achieve high levels of internal and external validity through the proposed impact evaluation design.

Changes to the SEP are reported. Explain any major changes that were made to the feasibility study proposed in the SEP. For example, study enrollment must be representative of the population. If there have been no changes, please state that no changes have occurred.

Impact Study—Approach and Methods

An Impact Study Report must include a discussion of the study approach (research design) and methods (sampling, measures, and data collection). For any impact design type, the report should offer a brief description of the design, its strengths and limitations, and how the project unfolded with the chosen design, particularly noting any challenges to implementing the plan as proposed in the SEP. The design type and specifics of the design itself should be stated clearly. The findings should include statistical evidence of how well a program works and what effect it has on participants. Each report should note the target level of evidence for the approved SEP and the level achieved by the current report.

Report Review Checklist: Impact Evaluation Design Selection

- The report clearly identifies the study design selected.
- The report justifies the target level of evidence based on a discussion of internal and external study validity.

Impact Evaluation Design Selection

This section describes the contents of impact reports using an experimental design (i.e., randomized control trial (RCT)) or the most common types of quasi-experimental designs (QED) used by SIF grantees. This section also describes report contents for non-experimental research designs (such as cost-benefit analyses and feasibility studies) which can be combined with more rigorous designs to assess program impacts. The report should include several components regardless of the design selected:

Clearly identify the study design selected. Provide a clear description of the impact study design being utilized. Readers should have a clear sense of what the counterfactual condition (e.g., control/comparison group) comprises, including what the unit of assignment is for the study (e.g., individual, classroom, and site) and how the design increases the comparability of the treatment and counterfactual groups/conditions (through random assignment, matching, etc.).

Justify the target level of evidence based on a discussion of internal and external study validity. Identify the target level of evidence and describe exactly how the study design selected will be able to reach the targeted level of evidence and will address potential threats to internal and external validity. For example, a randomized control trial with data collection across several sites in different states has the potential to reach a strong level of evidence. As such, the evaluators should explain how the random assignment process addresses several threats to internal validity (such as maturation, regression to the mean, experimenter effects, etc.) and how the presence of multiple sites across addresses potential threats to external validity. In circumstances where quasi-experimental designs are being utilized, it is important to clearly specify which threats are addressed by the evaluation design and how, as well as which threats to internal validity are not addressed.

Random Between-Groups (Experimental) Design/Randomized Control Trial (RCT)

Description

The Randomized Between-Groups Design or Randomized Control Trial (RCT) is the strongest evaluation design available in terms of reducing threats to internal validity as it includes random assignment of program participants (groups of participants, program sites, schools, etc.) to either a program participant group or a control group that is not exposed to the intervention.

Section Content

RCT evaluation reports should include procedures to conduct the random assignment, determine the statistical equivalence of study groups, and assess variations in program implementation across study groups. Specifically, reports should resolve the following information.

Identify the unit of random assignment (and its alignment with unit of analysis). Clearly identify the unit of random assignment, (i.e., an individual person, a site, a classroom, an instructor). That unit should be the same as the unit at which the treatment impact is estimated, the unit of analysis.

Describe procedures used to conduct the random assignment, including who implemented the random assignment, how procedures were implemented, and the procedures used to verify that probability of assignment groups are generated by random numbers. Include descriptions of the method used to select and assign sample participants to a treatment or control condition, including:

- the complete randomization process - include how sample participants were defined, identified, screened, and any eligibility criteria used to recruit participants or limit participation
- measured characteristics used to distinguish individual participants (i.e., age, grade, gender, race, socio-economic status, size)
- the specific timing of random assignment prior to the beginning of program service receipt, and whether random assignment of all participants to groups is simultaneous or rolling
- any computer software or analog algorithm to assure randomization of assignment
- if the sample contains subgroups or strata, a description of how these are defined or formed and,
- a description of how the sample represents or differs from the population.

Report Review Checklist: Random Between-Groups (Experimental) Design

- Unit of random assignment is clearly identified (and aligned with unit of analysis).
- Procedures to conduct the random assignment, including who implemented the random assignment, how procedures were implemented, and the procedures used to verify that probability of assignment groups are described.
- Blocking, stratification, or matching procedures used—to improve precision in the estimate of the program effect or to balance groups on measured characteristic(s)—are described.
- The program group and, to the extent possible, the control group conditions are described.
- Procedures and results of an analysis to confirm equivalence of groups are discussed.
- Changes to the SEP are reported.

Additional Resources

For more information on Random Control Trials and guidelines for reporting on Random Control Trials, see http://www.ebbp.org/course_outlines/randomized_controlled_trials/ - DA.

Describe blocking, stratification, or matching procedures used to improve precision in the estimate of the program effect or to balance groups on measured characteristic(s). Include descriptions of:

- any blocking or stratification of the sample and its rationale
- program and evaluation staff involvement in the random assignment
- steps taken to prevent any intentional or unintended bias in random assignment
- the number of participants assigned to each condition and whether the treatment and control groups are equal sizes and,
- the process for determining and aligning treatment and control group sizes with a statistical power analysis if the sample is different from what was proposed in the SEP, or a description of the statistical power analysis previously conducted if the groups are the same size as specified in the SEP.

Describe the program group and, to the extent possible, the control group conditions. Include a description of the program model and conditions that could influence the results of the study. These include the following:

- program implementation data that includes what constituted treatment or exposure, dosage levels (minimums, maximums, averages, etc.), whether partially treated participants are included in analyses, whether, how many, and how/why any participants crossed over from treatment to control or vice versa, and findings on fidelity, exposure, quality of program delivery, participant responsiveness and engagement, program differentiation, and participant satisfaction
- conditions to which control participants were exposed (i.e., knowledge of the program, not participating in some parts of a program, any similar services received)
- any program conditions, restrictions, standards or requirements, and benchmarks or goals that may have influenced the study, including participation rates, completion rates, attrition rates, or minimum dosages that were set and,
- the role of program staff, the evaluation team or others, and their awareness of treatment and control assignment and any influence this knowledge may have on the outcome.

Discuss procedures and results of an analysis to confirm equivalence of groups. Once groups are formed, the statistical equivalence of the groups on measured characteristics at baseline should be analyzed and verified (e.g., through t-tests, ANOVAs/MANOVAs). If the groups vary on any of these characteristics to a statistically significant degree, the report should describe any steps taken to address and correct for this variation, such as increasing sample sizes, or correcting for the variation statistically.

As the program proceeds, evaluators should continue to monitor and report on statistical equivalence and group composition. Any variation among the groups that might lead to change in group composition or differentiation in the characteristics of participants in each group should be noted and, if necessary, compensated for. For instance, if

Additional Resources

- For information on how to write the Methods Section of a Research Paper, see (http://rc.rcjournal.com/content/49/10/1229.full.pdf).
- For information on establishing baseline equivalence, see page 15 of What Works Clearinghouse (http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf).

one group experiences a substantially higher dropout rate, a skewed dropout rate, or differences in dosage (for instance, attendance rates in a long-term program) such that the groups are no longer statistically equivalent, or have exposure to intervention content through formal or informal mechanisms, the report should include information about this and any steps taken to address the change in group size or quality. This could include statistical adjustments during the analysis phase, or adding additional participants to the study, if feasible.

The report may also include any descriptive information or qualitative analysis documenting why the group characteristics changed (particularly if they do not remain statistically equivalent), if attrition is higher than anticipated, or if group sizes drop below the estimated minimum detectable effect size (MDES).

Document whether there have been changes to the SEP. It is important to highlight in the report(s) any significant changes to the experimental design as outlined in the SEP, why they occurred, and how the revised design will analyze impact. State any changes in targeted level of evidence (e.g., if the changes mean that the study will target moderate instead of strong level of evidence). If there have been no changes, please state that no changes have occurred.

Between Groups (Quasi-Experimental) Design Formed by Matching

Description

Between Groups designs formed by matching are good alternatives when it is not feasible or ethical to randomly assign participants to either the treatment or comparison group. These designs attempt to achieve a high level of similarity between the treatment and comparison groups without random assignment. Matching methods align treatment and comparison groups to minimize potential selection biases caused by variables that might create differences in the outcome other than the treatment (i.e., covariates). Some quasi-experimental designs with groups formed by matching have the potential to reach moderate or strong levels of evidence, depending on how they are carried out.

Report Review Checklist: Between Groups Design Formed by Matching

- Reasons why the comparison group might differ from the treatment group, the method of matching used, and how the method adjusts for differences are discussed.
- Unit of matching is clearly identified and aligned with unit of analysis.
- Standardized mean differences in baseline characteristics between treatment and comparison groups are described.
- Changes to the SEP are reported.

If propensity score matching is used:

- Procedures to carry out the matching to form a comparison group are described, including the method and how the propensity score accommodates the study design.
- A description of the variables used in the matching is included.
- Information documenting the success of the matching process are provided, including which variables were associated with treatment group membership, the distribution of propensity scores in the treatment and comparison groups, the proportion of cases matched, and the standardized mean differences in baseline characteristics between treatment and comparison groups.
- How the propensity score was used in the analysis is described. Matched and unmatched results are provided for comparison.

If no propensity score matching is used:

- Methods used to form the proposed comparison group are described such that the validity of the matching is explained and documented.
- How the propensity score was used in the analysis is described. Matched and unmatched results are provided for comparison.

Section Content

A report that describes an evaluation using a Between Groups Design Formed by Matching must present several details about the matching approach to demonstrate that equivalence of the treatment and comparison groups was actually achieved and assessing equivalence of the groups at baseline. There are additional reporting requirements based on whether or not propensity score matching is used. In general, these reports should:

Discuss reasons why the comparison group might differ from the treatment group, the method of matching used, and how the method adjusts for differences. Include details of any potential sources of selection bias that remain as threats to internal validity because these differences have not been equalized across the two groups (e.g. they are factors completely unique to or confounded with the program or comparison group, or they are factors not included in the impact analysis).

A major threat to the validity of evaluation findings may occur when the comparison group receives program services or the treatment group does not receive services equally or as intended. For example, some members of the comparison group may have received similar services from another agency, or different sites may have differentially delivered program services to treatment group participants. If there is contamination of the samples, describe how the design was impacted and how the analyses adjusted for the contamination.

The type of matching used for the study should be clearly described, including the type of matching (e.g. propensity scores, Malhanoubis distance), the steps to conducting the matching procedure, and the variables included. An explanation should be provided as to how the given procedure can reduce threats to internal validity from selection bias, and citing literature supporting the use of the chosen technique is recommended.

Clearly identify the unit of matching and its alignment with unit of intervention and unit of analysis. The unit of matching is who or what will be matched. Matching may occur at the individual level, where participants are the unit of matching, or at a more macro level (i.e., schools or sites may be matched). The report should present the unit of matching, as well as describe how the unit of matching is aligned with the unit of intervention (e.g., programs delivered at the school-level should use schools as the unit of matching) and the unit of analysis (e.g., statistical analyses using data from individual participants should use participants as the unit of matching).

Describe the standardized mean differences in baseline characteristics between treatment and comparison groups after the match was conducted. Document the baseline equivalence of the program and comparison groups on observed characteristics. Estimation of the statistical significance of group means should be included, with $p < .10$ or .05 being a standard p-value indicating statistical significance. Conventional standards of group non-equivalence can be applied to standardized mean differences, or effect sizes (i.e. Cohen's d), and should be included in the report. The report should discuss any relevant implications for the analysis and the study's targeted level of evidence where any group non-equivalences have been observed post-match.

Document whether there have been changes to the SEP. Explain if the actual design and procedures vary from those proposed in the original evaluation plan. For example, the matching procedure originally proposed may not have been feasible using the actual sample data. Discuss any differences between the actual designs and matching procedures to those described in the plan, and justify why the actual design and procedures are likely to yield reliable and valid results. If there have been no changes, please state that no changes have occurred.

With propensity score matching

Describe procedures to carry out the matching to form a comparison group, including the method and how the propensity score accommodates the study design. Different methods for matching have different levels of effectiveness. The report should present the method used for matching and justification for selecting that method. Propensity score matching is the preferred method as it uses a statistical adjustment based on multiple pre-intervention variables. For propensity score matching, the type of matching (e.g., caliper, nearest neighbor) and how the propensity score accommodates the study design (e.g., weights, clustering, multiple stages of matching) should be described.

Include a description of the variables used in the matching. Describe the variables used for matching, data sources (e.g., administrative data, census data, and medical records), how the variables were selected, and a justification for the quality of the matching variables (e.g., cite research that demonstrates the correlation between the variable and the outcome). It is useful to present the correlation between the matching variables and outcome as measured in the current sample(s), especially when there is limited past research documenting the relationship between the matching variables and outcome.

Provide information documenting the success of the matching process, including which variables were associated with treatment group membership, the distribution of propensity scores in the treatment and comparison groups, the proportion of cases matched, and the standardized mean differences in baseline characteristics between treatment and comparison groups. A critical aspect of this section of the report is demonstrating the success of the matching process by assessing equivalence of the groups at baseline. Evaluation reports should document the strength of the association between the matching variables and treatment group membership, and the proportion of cases that were matched.

When matching has been performed, a better alternative to significance testing is to measure standardized bias, Cohen's d, and/or the percent bias reduction before and after matching. Rubin (2001) established criteria for measuring balance between groups: (a) the standardized mean difference of the propensity score is less than 0.5, (b) the ratio of the variances of the propensity score is close to 1, and (c) the ratio of the variances of the residual errors of the covariates after predicting the propensity scores is close to 1. In addition to these statistics, the report should include the distribution of propensity scores in the treatment and comparison groups, and present both the matched and unmatched results for comparison.

Sometimes, matching does not yield equivalent groups. If this is the case, present the assessment of equivalence to demonstrate how the groups are different. Explain how the differences in groups are handled in the analyses (e.g., controlling for covariates that differ significantly between groups) and describe any limitations the lack of equivalence places on the evaluation findings.

Describe how the propensity score was used in the analysis and provide matched and unmatched results for comparison. For propensity score matching, explain how the propensity score was used in the analysis. For example, the score could be used as a variable in the analytic model or as a weight. If transformation was used, describe how the propensity scores were transformed (e.g., logit transformation). Note what adjustments were made to the size of the analysis sample by matching the program and comparison groups.

Without propensity score matching

Describe methods used to form the proposed comparison group such that the validity of the matching is explained and documented. For other matching methods, the algorithm used for matching should be provided. For all matching methods, include a justification for the matching method with an explanation of how the treatment and comparison groups may differ, how the matching method adjusts for those differences, and how the method addresses threats to internal validity (include citations to relevant documentation). Include a description of the variables used for matching, the data sources (e.g., administrative data, census data, medical records), and how the variables were selected (e.g., cite research that demonstrates the correlation between the matching variable and the outcome, to the extent possible). Present the correlation between the matching variables and outcome as measured in the current sample(s).

Additional Resources

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services and Outcomes Research Methodology, 2(3-4), 169-188.

Regression Discontinuity (Quasi-Experimental) Design (RDD)

Description

Regression Discontinuity Designs (RDD) evaluate the causal effects of an intervention when a Randomized-Control Trial (RCT) is not feasible given real-world limitations on assignment to the treatment and control groups . The RDD uses a cut-off score on a test, or other pretreatment scored assessment or forcing variable (e.g., date of birth), to assign sample units to the treatment or comparison group (Thistlewaite & Campbell, 1960). There must be randomness in the forcing variable, and individuals (e.g., potential treatment recipients, program staff) cannot manipulate treatment status.

Report Review Checklist: Regression Discontinuity Design

- The alignment of the measure, cut-off score, and bandwidth with the unit of analysis is described.
- The methods used to apply the cut-off score are detailed.
- The pre-test measure, distribution of measure, and bandwidth are as expected.
- The treatment and counterfactual conditions are as proposed.
- The details of the model specification and how the model was estimated are provided.
- Changes to the SEP are reported.

Section Content

When reporting the results of a RDD study, there are several issues to address to demonstrate that the results are reliable and valid, along with the information required to generate conclusions regarding your evaluation questions and replicability.

Identify how the measure, cut-off score, and bandwidth selected are aligned with the unit of analysis. Provide a summary of the RDD, define the sampling unit (e.g., students or schools), forcing variable (the variable used to assign units to the treatment or comparison group), cut-off score, and outcome variable. Describe the response rate, overall attrition, and attrition by treatment status to demonstrate that differential attrition is not a problem and that the sample is of sufficient quality to move forward with the RDD analyses.

Detail the methods used to apply the cut-off score. Clearly detail how the score was used to assign study participants to either the treatment or comparison group. It would be useful to provide any available documentation to support that the assignment occurred as planned (e.g. the score was used as planned in the assignment decision, and there were no case exceptions to the planned assignment decision).

[2] In a RCT, each sample unit, such as an individual student or school, has a known, non-zero chance of being assigned to the treatment or control group. Many programs are unable to randomly assign sample units to the treatment due to ethical and practical limitations imposed by the real-world context of program implementation.

Discuss the degree to which the pre-test measure, distribution of measure, and bandwidth are as expected. Discuss the degree to which assignment to the treatment and comparison groups adhered to the cut-off criterion. The cut-off score for the forcing variable must be used to assign units to the treatment and comparison groups. To assess adherence to the cut-off and whether there are a sufficient number of units in the comparison group, present the frequencies (as raw numbers and percentages of total sample size) of units properly receiving treatment or not receiving treatment, crossovers, and defiers in a table or in the report narrative. Also, the report should identify whether the RDD is sharp or fuzzy. A sharp RDD refers to occasions when the treatment is solely based on the forcing variable cut-off criterion (i.e., there are no crossovers or defiers). A fuzzy RDD refers to situations where the treatment reaches participants in the comparison group, or the treatment is not received by participants in the intervention group (i.e., there are crossovers and/or defiers). If the RDD is fuzzy, explain why the cut-off criterion was not strictly followed and what other factors may have influenced the delivery of the treatment.

With regard to any baseline differences, present graphs comparing the average values by the forcing variable, using the same intervals as in the previous graphs, for each of the covariates (not including treatment assignment). For internal validity to be strongest, there should be no discontinuities at the cut-off in the covariate graphs. Discuss and adjust for any problems exposed by these graphs in the report.

Provide histograms showing the distribution of the outcome variable, and include the line of best fit on either side of the cut-off value. The report should also fully specify and explain the model used to estimate the treatment impact for the RDD.

Describe whether the treatment and counterfactual conditions are as proposed. Include details of the distribution of pre-test and (when available) post-test outcome variables, and discuss the degree to which these distributions are concordant with what was anticipated per the SEP (e.g. is there a steady, continuous distribution of pre-test measures, are there sufficient cases around the cut-off score to allow for reliable detection of results or was it necessary to adjust the bandwidth in any way, and are there steady, linear relations with comparable slopes between pre-and post-measures below and above the cut-off score?).

Provide the details of the model specification and how the model was estimated. Describe the necessary details to allow reviews to fully assess any threats to internal validity not addressed by the impact analysis, including the type of procedure, steps to conducting the analysis, the sample size, variables included, and the interpretation of the results. See the analysis section for more details.

Document whether there have been changes to the SEP. Note any major changes from the design proposed in the SEP. If there have been no changes, please state that no changes have occurred.

Single Group (Quasi-Experimental) Design

Description

In a Single Group Design (e.g., Pre/Post Tests, Single Subject Designs), participants are given the treatment and changes are observed. No randomization is employed in this design and there is no control group (some variations do include a comparison group, such as an Interrupted Time Series design with a comparison group). At a minimum, there will be a pre-test and a post-test administered, with the intervention taking place between the tests.

Report Review Checklist: Single Group Design

- Each intervention phase of the design, including the baseline condition, is clearly described.
- Number of measures during each measurement phase is detailed.
- Number of measures during each measurement phase to establish trend and rule out rival explanations is described.
- Appropriateness of timing of measures pre/post interruption to the intervention is explained.
- Changes to the SEP are reported.

Section Content

A report for a study with a single group design report should:

Describe each intervention phase of the design, including the baseline condition. Clearly describe each intervention phase, including the baseline condition, timing and duration of data collection, data collection procedures, sample, and the measures used.

Detail the number of measures used during each measurement phase. Explain the number of measures and timing with respect to the intervention.

Discuss the extent to which the number of measures during each measurement phase is sufficient to establish trend and rule out rival explanations. Describe, or show, how the number of measures during each phase was sufficient to establish a trend and rule out rival explanations. Latent growth curve models require measurement of a minimum of three time points, while it is strongly recommended that interrupted time series models have a minimum of three time points pre- and post-intervention. These requirements allow for adequate modeling of the growth patterns for outcome variables.

Describe how the timing of measures pre/post interruption is appropriate to the intervention. Clearly describe the baseline condition, each intervention phase, and the timing of the measures in regard to the intervention. Importantly, any staggering of the timing of program implementation across the study sample should be accommodated by the model, and data collection should occur before any program services have been delivered. The capacity of the model to address additional threats to the plausibility of causal assertions, such as history effects, attrition, statistical regression, and experimental and subject effects are of special concern for single case designs (e.g., interrupted time series) and should also be discussed in the report. Some supplementary analysis or information (e.g. secondary data that rules out history effects) may be useful to include.

Document whether there have been changes to the SEP. Document any major changes in the treatment or implementation from what was proposed in the SEP, including the reasons for the changes. If there have been no changes, please state that no changes have occurred.

Interrupted Time Series (Quasi-Experimental) Design

Description

The Interrupted Time Series Design is a quasi-experimental or experimental design that makes repeated measures on the outcome of interest before and after an intervention is imposed. A well-constructed Interrupted Time Series has an advantage over other designs because provides information on whether the changes are intermittent or delayed, permanent or temporary, and control for confounding variables and the tendency of data to regress to the mean. However, this type of design is especially vulnerable to attrition effects, particularly differential attrition between groups, and this point should be addressed in the report and the results justified. It is strongly recommended that interrupted time series models have a minimum of three time points pre- and post-intervention. These requirements allow for adequate modeling of the growth patterns for outcome variables. This design can be conducted with a comparison group.

Report Review Checklist: Interrupted Time Series

- The number of measures in the pre- and post-intervention to establish a trend and rule out rival explanations is described.
- The appropriateness of the timing of measure pre- and post-intervention is explained.
- Comparison cases are clearly described.
- Treatment and counterfactual conditions are as proposed in the SEP.
- Changes to the SEP are reported.

Section Content

The report should provide a table that includes means of the outcome at each time-point for each group (program and comparison, where applicable), and the number of subjects at each time point in each group.

It is also helpful to create one or more figures of the model estimated means at each time point for each group, compared to the sample means at each time point to show that the trends are appropriately modeled.

In addition, the report should:

- Detail the extent to which the timing of measure pre- and post-intervention is appropriate to the intervention. It is recommended that as many data points as possible are gathered, especially if there are seasonal or cyclical effects (like the effects of winter break for education studies). The timing between measures must allow for variation between the time points, or the effort in gathering more data will be less effective. The interval between pre- and post-intervention data collection should be small enough for no viable alternative explanations for the program impact to appear feasible.

- Describe comparison cases. Although a counterfactual group is not technically necessary to do an Interrupted Time Series, it adds considerable support to the study's internal validity and is strongly recommended when targeting moderate levels of evidence. As with any quasi-experimental approach, it is beneficial for the study to have empirical evidence that the groups are equivalent at baseline. In a Comparative Interrupted Time Series, it is additionally beneficial to provide evidence that the secular trends before the intervention are equivalent between the

groups. In any case, the comparison group should be clearly described, and threats to internal validity in that group, such as contamination, should be noted.

- Discuss the extent to which the treatment and counterfactual conditions are as proposed in the SEP. Describe unexpected changes that affected implementation of the intervention, the intervention participants, or the comparison group. Explain how changes to the treatment or counterfactual conditions were addressed to limit threats to internal validity of the study.

- Document whether there have been changes to the SEP. Any changes to the SEP should be described, including the explanation for the changes. If there have been no changes, please state that no changes have occurred.

Report Review Checklist: Non-Experimental Design

- Barriers to proposing a design with the potential to contribute to strong or moderate evidence categories are described.
- Full study design is clearly and comprehensively explained.
- Descriptions of the treatment and counterfactual groups (if any) are included.
- Where appropriate, assignment of study participants to groups is described.
- Additional threats to the internal validity of the design are discussed.
- Ways future evaluations can be designed to rule out these threats are described.

Non-Experimental Design

Description

Non-Experimental Design studies encompass a broad range of studies that often do not include a counterfactual and do not include pre- and post- intervention measurements. This includes single group studies with measurement at only one time point (i.e., no pre/post intervention measurement and no counterfactual), implementation and feasibility studies, some cost studies, case studies, and systems change studies.

Subgrantees reporting on non-experimental design feasibility or implementation studies should refer to the relevant previous sections for specific guidance in addition to this section.

Section Content

A report for a study with a non-experimental design should:

- Describe barriers to implementing a design with the potential to contribute to strong or moderate evidence categories.
- Discuss the progress of the non-experimental study and how this study might support development of a full Experimental or Quasi-Experimental Design.

Additional Resources

For more information on non-experimental research, see http://2012books.lardbucket.org/books/psychology-research-methods-core-skills-and-concepts/s11-nonexperimental-research.html.

[3] Once the non-experimental study is completed, it is required that an updated SEP be submitted to CNCS outlining a full study, including an impact evaluation yielding moderate or strong evidence during the SIF timeframe.

-- First, describe the challenges to developing an impact study. Briefly summarize the barriers to identifying a counterfactual and any testing measures and problems when developing an instrument, focusing only on those addressed in the study.

-- Second, include the proposed options for a full impact study, as well as their advantages and disadvantages. Describe which options for a full study and impact evaluation are vetted and whether they are excluded or supported by the study, or how preliminary results will contribute to the development of the full study and impact evaluation. Discuss threats to internal, and if relevant, external validity in the impact evaluation options presented.

Explain the full study design. Clearly and comprehensively describe the instruments, measures, processes or methods tested (i.e., pilot testing, validity, reliability) as part of the non-experimental study, the results and conclusions from these tests, as well as the implications for a full impact study. Include a description of how the instruments, measures, processes or methods were tested, how often, how long, and how frequently they were tested, and the implications for a full study.

Include descriptions of the treatment and counterfactual groups (if any). Describe the possible treatment and counterfactual groups, and the process for assigning participants to the groups. The identification of a representative sample with an appropriate counterfactual group is a key element of an impact study (i.e., experimental or quasi-experimental design).

Describe each of the populations under study, potential sampling techniques, and the treatment and counterfactual groups under consideration. For instance, if a study involves mothers and young children, studying the mothers may have certain advantages because of the ethical concerns involved in studying children, but studying the children may provide more direct measurement of impact. A study of both the mothers and children may also be a possibility. The evaluation team should describe the threats to internal and external validity presented by the chosen option.

Where appropriate, describe assignment of study participants to groups. If the non-experimental study includes preliminary testing using a treatment and counterfactual group, fully describe the process for assigning study participants to these groups. Describe any challenges to group formation or equivalence and how these inform the impact evaluation.

Discuss additional threats to the internal validity of the design. Given the lack of a clear comparison group, most non-experimental designs are unable to rule out significant threats to internal validity. As such, discuss the extent to which the study results may potentially be impacted by selection bias, selection additive effects (differential reactions to treatment due to selection characteristics), regression to the mean, history (events unrelated to treatment), maturation, novelty, testing (experience with the pre-measures can influence post-measures), and expectancy effects (from experimenter or tester expectations).

Describe ways future evaluations can be designed to rule out these threats. Include a description of how the conclusions of the non-experimental study will lead to a full impact study that will address the threats to internal validity.

Document whether there have been changes to the SEP. Note any changes from the design as proposed in the SEP, including the explanation for the changes. If there have been no changes, please state that no changes have occurred.

Study Participant/Sample Flow

Study Participant/Sample Flow

Description

Clearly describe recruitment procedures, response or enrollment rate, sample sizes for all groups at each major study time point, including the numbers/percentages of potential participants screened out as ineligible, who did not enroll due to non-consent/refusal, or who could not be located or could not complete a full enrollment process (Table 1), and describe the characteristics of sample participants to assure the reader that study participants are representative of the target population and program impacts are not merely a function of selection bias.

Additionally, it is important to report on rates of sample:

- retention
- attrition
- non-response, and
- missing data.

This enables the reader to determine if study findings are based on the entire original sample or on a particular subset of the sample. While some sample attrition and missing data are typical, findings from evaluations with substantial attrition or missing data (especially differential attrition and non-random missing data) must be interpreted with caution as these issues may threaten the validity of the evaluation.

Report Review Checklist: Sample Description

- The study participant flow is described in the text as well as a table or diagram and includes the number of each study sample (intervention and comparison groups) at different time points.
- The composition of the sample is described, including demographics and other characteristics relevant to the study.
- Changes to the SEP are reported.

Sample Recruitment and Retention

- There is a description of recruitment and retention strategies and efforts.
- Sample retention rate is reported.
- How overall and differential attrition was assessed is detailed.
- Any differential attrition findings are reported.
- Changes to the SEP are reported.

Non -Response Bias and Missing Data

- Description and results of assessment and adjustment for potential biases (due to non-consent and data non-response) are included.
- Statistical procedures used to adjust for missing data are discussed.
- Changes to the SEP are reported.

In particular, there are three main aspects of the study participant/sample flow that should be reported:

a. sample description (size and composition)
b. sample retention and attrition and,
c. non-response bias and missing data.

For each study group (e.g., intervention and control/comparison), the numbers of participants assigned to each group, received the intended treatment, and were analyzed for the primary outcome(s) should be reported. If available, the number of people assessed for eligibility should also be reported. Although this number is relevant to external validity only and is arguably less important than the other counts, it is a useful indicator of whether trial participants were likely to be representative of all eligible participants.

Section Content

Participant Flow Description

Describe flow of participants through the study, including the sample size for each set of data collected at different time points. Specify the number of participants who were enrolled in both intervention and comparison groups at baseline and then present sample sizes at each subsequent measurement point. If sample sizes differ across measures (e.g., response rates differ by data source or data collection method either by design or due to nonresponse), specify the sample sizes for each primary measure to illustrate the extent of missing data in each outcome.

The participant flow should also be documented via a participant flow diagram similar to the diagram used by CONSORT (Consolidated Standards of Reporting Trials) for the transparent reporting of randomized controlled trials: http://www.consort-statement.org/consort-statement/flow-diagram. A table can also be used instead of a flow diagram to share participant flow numbers.

The key study participant numbers and time points to be included in the flow chart (or a "participant flow" table) are described in Table 1. These counts include the number of people included and not included for: 1) enrollment, 2) assignment to study groups, 3) intervention allocation (by study group), 4) follow up (by study group), and 5) analysis (by study group).

Table 1. Study Participant Flow - Types of Participant Counts to Report

Study Timepoint	Number of People* Included	Number of People* Not Included	Notes
1-Enrollment (e.g., assessment for eligibility)	People evaluated for potential enrollment	People who did not meet inclusion criteria or met the inclusion criteria but declined participation	Reasons for exclusion should be itemized and reported.
2-Assignment to Study Groups (if a multi-group study design)	Participants randomly assigned or assigned by other procedures (e.g., matching)		If a single group study, report the number of people receiving the intervention.
3-Intervention Allocation	Participants who received treatment as allocated, by study group	Participants who did not receive treatment as allocated, by study group	Reasons for not receiving the intervention should be itemized and reported. —If a single group study, report the number of people receiving the intervention.
4-Follow up	Participants who completed the intervention as allocated, by study group	Participants who did not complete treatment as allocated, by study group	Important counts for assessment of internal validity and interpretation of results. —Reasons for not completing treatment should be itemized and reported.
	Participants who completed EACH follow up as planned, by study group	Participants who did not complete follow-up(s) as planned, by study group	Important counts for assessment of internal validity and interpretation of results. —Reasons for not completing follow-up(s) should be itemized and reported.
5-Analysis	Participants included in main analysis, by study group	Participants excluded from main analysis, by study group	Crucial count for assessing whether a trial has been analyzed by intention to treat; reasons for excluding participants should be given.

* Adjust unit accordingly if the "unit" is a group vs. individuals (e.g., schools, communities, etc.) or a study employs a single group design.

Describe the composition of the sample, including demographics and other characteristics relevant to the study. In order to determine whether the study sample is representative of the target population and whether intervention and comparison groups are equivalent at baseline, it is important to clearly describe the composition of the sample. This description should include basic demographics (such as gender, race, ethnicity, age, education, socioeconomic status, etc.) and characteristics that may be relevant to the study (such as prior exposure to the program or to program-like services, baseline levels of the outcome variable, etc.).

Additionally, if the study experiences substantial attrition, then interim and final reports should also comment on the composition of the treatment and comparison groups after attrition and note any significant differences in the composition of study participants from baseline to follow-up.

Document whether there have been changes to the SEP. Any changes from the SEP should be explained, including justifications for any changes. If there have been no changes, please state that no changes have occurred.

Sample Recruitment, Retention and Attrition

Sample retention is important because attrition among research participants affects the evaluation's internal validity, statistical power, and potentially, external validity.

Describe recruitment and retention strategies and efforts. The report should describe how the evaluation recruited, engaged, and retained participants and tracked their data. It should also discuss any troubleshooting that occurred to support participant retention and challenges that occurred in securing the current and final target sample size

Report the retention rate for the study sample. It is important for a study to monitor the number of participants that are retained in each study group (treatment and comparison groups) during the duration of the study. For studies that engage participants for more than a short period of time before the final measurement occurs, and for any studies that incorporate pre- and post-measures on participants, attrition should be assessed, reported, and its effects considered.

Report attrition numbers (i.e., overall and differential) at each follow-up and detail how attrition was assessed. There are many approaches that can be used to assess and adjust for attrition and non-response bias. Non-response and/or attrition of greater than 20 percent should always be checked to determine whether different individuals or groups are dropping out more, or responding less to data collection.

Report whether there was differential attrition. Differential attrition between the treatment and comparison or control group is important to assess and document. If differential attrition is high, it is a potential source of bias, and can lead to a change or lack of change being attributed to the treatment when it is actually caused by participant characteristics. Participants who stay throughout the study may be systematically different from those who leave the study.

Characteristics associated with differential attrition should be assessed, monitored, and reported on (e.g., by using logistic regression to determine which characteristics predict the propensity to drop out). Where attrition is high, evaluators should report on the statistical procedures (and their findings) used to check results for robustness under different conditions should be employed (e.g.,

"what if" scenarios to check how the results might change if a group with higher attrition had responded at the same rate as one with lower attrition).

Document whether there have been changes to the SEP. Any changes from the SEP should be explained, including justification for any changes. If there have been no changes, please state that no changes have occurred.

Non-response Bias and Missing Data

Include the description and results of assessment and adjustment for potential biases -- due to non-consent and data non-response. Evaluators should describe procedures used to assess non-response bias (i.e., when participants are retained in a study, but do not have complete data, can also affect study results). Study participants who do not answer certain questions may be systematically different than those who do.

For example, a study measuring the effects of an integrated behavioral health model may find that patients with more severe health problems at baseline are less likely to respond to patient follow-up surveys than those who are in fair health.

- If incomplete data is due to the inability of some participants to complete the item or assessment due to physical capacity, this should be reported clearly and separately from other types of nonresponse. For example, some participants may not be able to complete certain physical tasks such as tasks assessing grip strength or pulmonary function tasks assessing breathing capacity. These are examples of legitimate missing data because the participant cannot perform the task, not because the participant chose not to respond or did not understand the item or instructions. In such cases, the nonresponse rates should be calculated separately for each reason and analyses should account for these appropriately.

- If incomplete data and non-response is due to method variance or error, report this clearly and specifically. For example, participants may not answer certain measures or individual items due to the format of the measure (e.g., too negatively worded, too long, too difficult to read, to intrusive or personal). If the potential nonresponse was anticipated, report how the potential for nonresponse was assessed and managed.

- If the nonresponse became apparent during data collection, such as from implementation or satisfaction surveys, the extent of the nonresponse, the reasons for it (if known), and any systematic variation or differential completion rates by participant characteristics should be noted. In addition, any techniques to adjust measures or data collection procedures during the evaluation should be described clearly (e.g., substituted new measure, had interviewers read surveys aloud as needed).

Discuss statistical procedures used to adjust for missing data. Discuss any procedures used to adjust or address statistically significant differences in item or instrument non-response data. Approaches include: getting more data, weighting the data you have, and adjusting and imputing

Additional Resources

- Differential attrition bias: What Works Clearinghouse Procedures and Standards Handbook version 2.1, page 13.

- Practical Tools for Non-response Bias offers a description of multiple approaches to address attrition or non-response bias.

where possible. If getting additional data, weighting data, or adjusting data are not possible, make sure that the report documents, describes, and accounts for missing data in analyses.

Document whether there have been changes to the SEP. Highlight any changes from the original evaluation plan in regard to sample size, sample retention strategies, and attrition. Include an explanation for any changes. Post-hoc power analysis should be conducted if it is suspected that a smaller than planned sample size resulted in statistically non-significant, but directionally positive findings. If there have been no changes, please state that no changes have occurred.

Report Review Checklist: Measures

- How each variable is used to measure outcomes and impacts in the study is detailed.
- Content and timing are described.
- Updates to or findings on measure construction, reliability, and validity are described.
- Changes to the SEP are reported.

Measures

Description

Every report should have a full Measures section detailing measures used in the current report, and any ongoing updates on other measures to be used in the final evaluation. This section documents the quality of the measures used in the evaluation, since using unreliable or invalid measures can reduce readers' faith in a study's findings, as well as the internal validity of the study itself. To successfully detect the effect of an intervention, a study's measures should be, at a minimum, empirically reliable and valid and appropriate for the population and construct being studied.

Section Content

Detail how each variable is used to measure outcomes and impacts in the study. Each measure used should be identified as either addressing confirmatory or exploratory research questions. Clearly indicate how each measure addresses the outcomes and impacts identified in the logic model.

Describe the content and timing of measurement. The timing of the administration of the instruments should be transparent. Additionally, the report should include the following details about the measures:

- number of respondents
- administration method
- number of questions included
- administration time
- organization and wording of the questions
- response categories used (if appropriate)
- potential score/response ranges and,
- distributions of the responses for model assumption purposes.

Describe updates to or findings on measure construction, reliability, and validity. For each instrument, include: (a) measures of reliability, (b) measures of validity, and (c) a description of the measure to determine appropriateness.

While it is important to reveal the origin and the original psychometric properties of a commercially available measure, it is imperative that the characteristics of the data gathered are also reliable, valid, and appropriate. Thus, the reliability/validity measures of the SIF evaluation sample should be reported in addition to any previous work in the creation of the measure.

When reporting on pilot test results, it is important to document what answers you received to the questions/items on your instruments and how your participants reacted to the instrument (e.g., provide a summary of the results of any cognitive debriefing you conducted).

Document whether there have been changes to the SEP. Finally, any necessary changes to measures from the SEP should be highlighted and justified. If there have been no changes, please state that no changes have occurred.

Report Review Checklist: Data Collection Activities

- A description of data collection activities for baseline measures/ statistical controls is provided.
- A description of who collected the data is included.
- A description of the role of staff members delivering the intervention with regard to data collection is described.
- The timing of data collection, relative to delivery of the program is explained.
- Discussion of whether the mode of data collection is the same for the intervention and control groups is included.
- Changes to the SEP are reported.

Data Collection Activities

Primary data refers to data collected for the evaluation, usually by the evaluation team or program staff (e.g. surveys, observations). Secondary data has been collected by someone else for another purpose, but will be used for the evaluation (e.g., existing administrative records).

Regardless of the type of data collected, the report should describe how the data were collected. Specify if data were collected through administrative records, program systems, or through instruments specifically created for the evaluation. For example, participants may have completed surveys, program staff may have maintained records such as performance assessments, or administrative data may have been used, such as academic transcripts.

Primary Data

Description

Use of valid and reliable measures is critical to supporting internal study validity, and appropriate administration of these measures helps protect your respondents and ensure reliability of your data.

Key primary data collection elements that should be monitored and reported on as they can affect both the reliability and validity of your study results include:

- Who collected the data?
- What data was collected?
- Was data collected from existing sources (e.g., school district test scores, health or wage records) and/or directly from participants of the study.
- When was data collected (both relative to the start of the study itself and to participants' enrollment in the study)?

- How often was data collected?
- How was data collected (e.g., online, paper and pencil, by phone, face-to-face)?
- How were data collectors trained? If relevant, what were the inter-rater reliability procedures and assessment results?
- What processes were used to protect respondents or support response?
- How was data handled following collection (e.g., transfer, cleaning, coding, storage prior to analysis)?

Section Content

Provide a full description of the data collection activities in the body of a report, or in a technical appendix to support the study's findings.

Provide a description of data collection activities for baseline measures/statistical controls. Baseline data for program participants and counterfactual group members establish the pre-intervention status of the sample and the equivalence of subgroups (e.g., treatment and counterfactual groups, lagged cohorts, different sites). The data are also the foundation for assessing change in participants over time.

The report should describe baseline data collection, including timing with regard to the program intervention and baseline data collection among subgroups, specifically:

- Describe the measures used and indicate if there were any changes from the measures listed in the SEP.
- Report on whether the same baseline measures were used for all participants and groups, or how they were comparable.
- If, for any reason, baseline data collection differed among groups, describe how baseline measures differed.

Baseline analysis should not only include impact measures, but should also assess the equivalence—or non-equivalence—of the program and comparison groups on key characteristics through the use of statistical controls. Even if sampling was conducted using a method intended to form equivalent groups, testing group equivalence through analysis of statistical controls at baseline is recommended. If groups were not equivalent at baseline, report on the steps taken as part of the analysis to adjust for baseline differences between treatment and comparison groups (see also the QED with groups formed by matching section).

Include a description of who collected the data. Describe who collected the data; it may have been the evaluation team, program staff, or some other party. If administrative data were used, specify the source(s) and availability of the data, the validity of the data with regard to the study, how reliability of the data was assured or assessed, and the evaluation team's experience working with that type of information.

Provide a description of the role of staff members delivering the intervention with regard to data collection. If applicable, describe the exact role of program staff members who are delivering the intervention in relation to the data collection process. Because there is a potential for bias, describe how this threat to the internal validity of the study was addressed.

Explain the timing of data collection, relative to delivery of the program. Clearly explain the timing of data collection, indicating when baseline data collection started and ended, when randomization or group assignment occurred (if relevant), when the intervention started, and any other key dates in the data collection process. Describe the intervals between data points and the timing relative to the intervention (i.e., before, during, after).

Discuss whether the mode of data collection is the same for the intervention and control groups. If data collection differed among observed groups, describe those differences, including sources, means of data collection, and who collected the data.

Document whether there have been changes to the SEP. If data collection activities were not conducted as outlined in the SEP, explain how they were modified (e.g., data sources, timing, procedures, and any unanticipated differences in the way treatment and comparison data were collected). If there have been no changes, please state that no changes have occurred.

Secondary/Administrative Data

Description

The results of evaluation studies depend heavily on the quality of the data collected. In order to assure the reader that appropriate inferences are made, it is imperative that the data collected is described in detail. This is especially true when the researchers are relying on an outside, secondary, source to collect their data. When reporting on the use of secondary/administrative data (including abstracted medical data):

- The secondary data source and its description should be transparent enough for the reader to replicate the process of collecting data (the logistics of receiving and storing the data should be explained).
- It is essential that the documentation of the results includes information about the data cleaning process, how new variables were constructed, and how datasets were merged to assure the reader that appropriate steps were taken to provide high quality data.
- It is also beneficial to note the Memorandum of Understanding (MOU) between institutions and disclose any obstacles to obtaining the data (e.g., obtaining Institutional Review Board [IRB] approval).
- Changes to the secondary datasets should be discussed.

Report Review Checklist: Secondary/Administrative Data

- Reasons for using secondary/administrative data are provided.
- The data and source are described.
- The steps to receiving and storing the data are described.
- Overlap between, or coordination of, SIF study data collection and secondary/administrative data can be determined. Details of any characteristics unique to either the treatment or comparison group are provided.
- Data construction, cleaning, and merging procedures are provided, including any recalibration of data structure and weights.
- A full description of the final analysis dataset is provided, including details of variables included and generalizability.
- Any problems with agreements for data access, storage, use, and reporting, as well as details of any changes to MOUs are provided.
- Details of MOU, and any strategies/relationships that will facilitate data delivery and/or use are provided.
- Changes to the SEP are reported.

Additional Resources

The Data Quality Assessment Tool for Administrative Data

http://www.bls.gov/osmr/datatool.pdf

Section Content

Provide the reasons for using secondary/administrative data. The evaluation report should make a qualitative case that the use of the secondary datasets is appropriate for the study. For example, perfect confounds inherent in certain datasets (e.g., only two schools in the study, one for each study group) should be disclosed at this point as a potential limitation to inferences that might be made about the study's findings. In addition, describe how the evaluation team coordinated any direct SIF study data collection with use of secondary or administrative data, including how and/or to what extent the treatment and comparison group obtained through existing data differ on key characteristics.

Describe the data and source(s). Be explicit about the type of data used in the study (e.g., city/county data, school administrative data); describing in detail the contents of such data, the security measures to protect the data, and the protocols researchers engaged to make appropriate use of the data.

Describe the steps for receiving and storing the data. Describe the details of the logistics of how the data were delivered and stored (e.g., the data were stored on a storage device such as a flash drive, physically delivered to the principal investigator).

Discuss the existence of overlap between, or coordination of, SIF study data collection and secondary/administrative data. Detail any characteristics unique to either the treatment or comparison group. Specify which data sources are primary, which ones are secondary, and clearly describe the timing and coordination of data collection for both data sources. Indicate any unique characteristics, likely due to sampling or data collection (e.g., geographic location, or means of data collection) that separates the program from comparison samples and cannot be accounted for in the impact analysis. Any resulting threats to internal validity should be discussed, potentially with supplementary information that explains how these threats can be addressed in future evaluations.

Provide data construction, cleaning, and merging procedures, including any recalibration of data structure and weights. The software used to screen, clean, and analyze the data should be disclosed (e.g., all data is analyzed in SPSS and any data value that is outside the theoretical limits is set to "missing"). Also, all new variables and weights that are constructed from the secondary dataset should be discussed.

Provide a full description of the final analysis dataset, including details of variables included and generalizability. The final analysis should be transparent, including variables used, whether the results are generalizable, and any problems inherent in the dataset.

Discuss any problems with agreements for data access, storage, use, and reporting, as well as details of any changes to MOUs. Obtaining secondary data (especially from school districts and national databases) can sometimes take longer than researchers anticipate, so interim reports should provide regular updates on the status of these agreements. Furthermore, all issues in regard to the MOU, data access, storage, and use should be disclosed in the evaluation report and also note the effects of any such issues on the timeline.

[4] SPSS is an acronym for IBM's Statistical Package for the Social Sciences.

Provide details of the MOU, and any strategies/relationships that facilitated data delivery and/or use. Provide the details of the MOU, and all steps and parties that helped to facilitate data delivery or use (or both). Discuss the steps in legally obtaining the data (i.e., IRB approval).

Document whether there have been changes to the SEP. Note any changes as proposed in the SEP including an explanation for the changes. If there have been no changes, please state that no changes have occurred.

Impact Study—Analysis and Results

Description

To ensure the strongest possible evidence from the evaluation, the correct statistical analysis technique must be employed and reported on. The statistical technique chosen depends on the type of research questions specified in the research design, and on the type(s) and quantity of data available for the analysis.

Statistical analysis elements to be reported on include:

- Use of an Intent-to-Treat (ITT) analysis
- Use of a Treatment on Treated (TOT) framework
- Analysis Assumptions and Alignment of Analysis with Data and Research Questions
- Covariant Adjustments
- Standard Errors
- Tests of Significance
- Multiple Comparisons
- Alignment of Unit of Analysis
- Effect Sizes
- Use of Difference in Differences (DID)

Section Content

Provide a clear description of the steps of the analysis. The Statistical Analysis section of the report should provide a clear and complete description of how the study's data was analyzed to arrive at the findings. Each report produced, whether the analysis is preliminary, midterm, or final, should include a full description of the analysis procedures used. If the report is intended for use by a non-technical audience, technical details of the analysis can be provided in an appendix. The description should include the following details:

- The approach taken to analyze the data (ITT, TOT, other)
- The number of cases available at each major step in the analysis, describing when and why cases were excluded in any step and,
- The software used to analyze data.

Each statistical package has its own idiosyncrasies and being transparent about the chosen package and any adjustments to the data or analysis approach necessary as a result of the software helps the reader determine any limitations of the results.

Report Review Checklist: Statistical Analysis of Impacts

- A clear description of the steps of the analysis is provided.
- How the statistical analysis of the data is aligned with the research questions is explained.
- How the statistical analysis is aligned such that the unit of analysis corresponds to the unit of assignment is described.
- Model estimation procedures are included.
- If applicable, covariate adjustments to estimated program effects, estimation of standard errors, and corrections for multiple comparisons are described.
- If sample sizes do not align with target levels in the SEP, supplementary power analyses and results are described.
- Preliminary or final impact analysis findings are detailed.
- Changes to the SEP are reported.

Additional Resources

- More information on and examples of statistical analysis can be found on the UCLA statistical computing website: http://www.ats.ucla.edu/stat/examples/default.htm .
- More information on the commonly used statistical tests and models, the data they require, and the questions they can answer can be found in

Appendix B4.

Explain how the statistical analysis of the data is aligned with the research questions. Clearly describe and note the statistical tests used, ensuring that these are appropriate both to the data collected and the evaluation question. Detail the statistical assumptions made and note any violations of the assumptions and the potential impacts these may have on the results.

Describe how the statistical analysis is aligned such that the unit of analysis corresponds to the unit of assignment. Clearly state the level of analysis (e.g., individual, group), describe how the level of analysis aligns with the process used to assign individuals to treatment or comparison groups, and account for any clustering in the data. Examples of a unit of analysis—the entity for which the treatment impact is estimated in the statistical model—includes individuals, classrooms, schools or districts, and geographic units. The unit of assignment is the defined entity for which assignment into the program, comparison, or control group was performed. If a program assigns whole classrooms to use or not use a new curriculum, the unit of analysis must be the classroom and the impact estimate for the new curriculum must be estimated at the classroom level, as opposed to the student level. The outcomes can be measured at the student level, (e.g., by student tests), but a multi-level model including a classroom-level estimate of the program effect must be used to analyze the data.

Include model estimation procedures. Describe and illustrate the model(s) used. Show the model used through formulas or (in the case of structural equation modeling) diagrams.

If applicable, describe covariate adjustments to estimated program effects, estimation of standard errors, and corrections for multiple comparisons. Detail any adjustments or decisions made in the analysis that may influence the results. In every data analysis, there are unexpected challenges. Choices made to resolve these challenges should be transparently described.

If applicable, include sample sizes for each analysis. If sample sizes differ across measures, report the sample size for each group in each analysis (e.g., analyses based on data from administrative data sources, likely to be complete or near-complete, vs. analyses based on survey data from a subset of participants either by design or due to incomplete survey data/ nonresponse).

If sample sizes do not align with target levels from the SEP, describe the results of supplementary power analyses. Detail any follow-up power analyses conducted due to a mis-alignment of sample sizes with the target sample sizes identified during the evaluation planning stage (i.e. described in the SEP).

Detail any preliminary or final impact analysis findings, including non-significant findings, organized by confirmatory and exploratory research question. Provide the results of the analysis for each specific impact study research question, stating whether the question is confirmatory or exploratory and including both results of statistical tests and any relevant effect sizes. Note the direction of the changes (e.g., positive or negative, increases or decreases), in relation to the counterfactual group (if applicable).

Each analysis (especially subgroup analysis) should be described separately in the document to allow the reader to understand the covariates used and other analytical decisions made.

- In addition to statistical significance results, effect sizes should be reported to provide important information on the practical significance of impact study results.

-Report effect sizes for, at a minimum, each confirmatory outcome result.
- When reporting effect sizes provide the name and equation for the metric used to calculate the effect sizes (e.g., Cohen's d, Glass's delta, Hedge's g).
- Resources for additional information on effect sizes can be found in Appendix C (Reference and Resource List).

• Appendix A2 (Tables and Figures) provides important guidance for presenting and formatting figures and tables in your report.

Document whether there have been changes to the SEP. Finally, the details of the analytical strategy should match what was proposed in the approved SEP. If changes in the analytic strategy were made to what is proposed in the SEP, a rationale should be offered explaining the change(s), and any limitations to the new analytical strategy should be discussed. If there have been no changes, please state that no changes have occurred. Post-hoc analyses to address questions that emerge during the study should be identified as such; these questions and analyses are considered exploratory and do not preclude or supplant conducting analyses and reporting results to address the primary confirmatory research questions in the SEP.

Conclusion—Summary of Findings, Lessons Learned, and Next Steps

Description

The Conclusion is one of the most critical sections of an evaluation report. The report should provide a clearly written conclusion for the evaluation that provides a summary of key findings, presentation of any key themes, apparent contradictions, reflections on the study's hypotheses, and discussion of lessons learned about the program and evaluation process. Further, it provides a clear explanation of suggested next steps for the program and its evaluation, as well as implications for other comparable programs.

The report should also describe how lessons learned during the evaluation process inform future program implementation and evaluations. These lessons learned are also useful for other comparable programs, including expanding the program to other settings and/or populations.

Report Review Checklist: Conclusion: Findings, Lessons Learned, and Next Steps

- A summary of Implementation findings is provided.
- A summary of Impact/Outcome findings is provided.
- An explanation of the level of evidence to which this study contributes.
- A discussion of lessons learned about the program and the process of evaluation is provided.
- A description of next steps for the program and for comparable programs is provided.
- Changes to the SEP are reported.

Section Content

Summary of Implementation Findings

Provide a summary of Implementation findings. These findings are important in assessing if, and the degree to which, a program performs as intended. Thus, the report should:

- Summarize findings along any themes that occur to the program and its stakeholders, such as which findings appeared surprising, or were seemingly inconsistent.
- Include any indication of how the Implementation Evaluation findings may explain the results seen in the Impact Evaluation. Describe whether the implementation study methods allowed the program to study the fully served population at the specified program dosage(s).

Also include a Summary of Findings related to any of the dimensions of implementation analyses addressed in the study. These include fidelity, enrollment, exposure, quality of program delivery, participant responsiveness and engagement, program differentiation, and participant satisfaction.

Summarize whether there have been changes to the SEP. The Implementation Findings section should summarize any changes from the SEP related to the Implementation Evaluation. Specifically, it should explain the extent to which all aspects of the implementation analysis match what was proposed in the SEP and reasons why some aspects (if any) were changed.

Summary of Outcome/Impact Findings

Provide a summary of Impact/Outcome findings. Summarize how the program affected changes in its participants and provide evidence about the extent to which the study can support conclusions that the program causes the observed changes in terms of evaluation design and execution (e.g., how well the study minimized threats to internal validity). Note the direction of the changes (e.g., positive

or negative, increases or decreases), in relation to the counterfactual group (if applicable). The report conclusions related to Impact Questions should include several components:

- Briefly summarize the research design used for the program.
- Provide findings from baseline, preliminary, midterm, and final analyses available at the time of reporting, detailing patterns in the findings as the study progressed.
- Note how analysis methods, samples, and measures changed from the approved SEP and over the course of the study, if applicable.
- Note the extent to which findings with high degrees of internal and external validity might support making causal inferences with respect to the specified unit of analyses.

Provide an explanation of the level of evidence to which this study contributes. Describe the level of evidence to which the findings in this report contribute based on the selected design type and ability to minimize threats to internal and external validity. Include statistical significance and effect sizes, briefly explaining which statistic is being reported, and discuss the results in comparison to previous evaluations of this or similar programs.

If a preliminary level of evidence is achieved, describe barriers to achieving moderate or strong levels of evidence and how these barriers affect the Impact Findings being reported. Describe how the preliminary evidence the study attains can be used to support studies targeting moderate or strong evidence in the future. Additionally, describe how the current evaluation builds the knowledge base concerning the program or intervention.

Summarize whether there have been changes to the SEP. The report should summarize any changes from the SEP related to the Impact Evaluations. Specifically, it should explain the extent to which all aspects of the impact analysis match what was proposed in the SEP and reasons why some aspects (if any) were changed.

Lessons Learned, Study Limitations, and Next Steps

Provide a discussion of lessons learned about the program and the process of evaluation. In the Lessons Learned section, provide a summary of what was learned about this program during the process of implementation and evaluation. Be sure to include key lessons, implications, and recommendations that could inform the field of the program from Implementation, Outcome, and Impact findings, as well as from the evaluation process itself. Also include a summary of study limitations, such as threats to internal validity for impact studies and describe lessons learned from involvement with SIF and federally funded initiatives in general.

Provide a description of next steps for the program and for comparable programs. Lastly, describe the next steps for the program. Be sure that these next steps are based on the findings and lessons learned detailed in the previous sections. Also, provide next steps or implications for other comparable programs that may have similar implementation or analysis processes. If there are no next steps, explain reasons why any suggestions were not included.

Other Aspects of Study Logistics and Feasibility

Description

Unexpected logistical changes and challenges may limit a study's capability to produce moderate or strong levels of evidence, and may, in fact, stop a study from proceeding or reporting at all.

These changes and challenges should be reported on and may include:

- problems securing Institutional Review Board (IRB) approval or waiver
- problems securing data from administrative sources
- problems with study execution
- changes to the intervention program during the study
- changes to the study timeline or budget
- changes to subgrantees or study sites
- changes to the evaluator and/or,
- changes to the program's role in the evaluation.

Report Review Checklist: Other Aspects of Feasibility

- Any problems with securing IRB approval are detailed, and impacts on the study timeline are discussed.
- Any changes to the timeline for data collection, follow-up, analysis, and reporting are included.
- Any changes to the evaluator/ subgrantee personnel and/or roles are reported.
- Any changes to the budget are reported.

Even when these issues are ultimately addressed, they may have far reaching impacts for the study and may contribute to lessons learned, including:

- limiting sample size (and thus study power and detectable effect size)
- endangering participants and limiting the extent to which data can be ethically used
- limiting evaluator knowledge of the full study and the extent to which limitations to data can be noted, addressed, and overcome
- limiting the extent to which the study can address all research questions and,
- causing disruptions in the quality and completeness of the study data that limit the extent to which study findings are reported and used.

Section Content

Explain any problems with securing IRB approval, changes to the study's schedule and budget, and modifications to the evaluator and/or Subgrantee's roles in the evaluation. If all of these aspects proceeded as planned, report that there were no changes or issues.

Discuss any problems with securing IRB approval, and impacts on the study timeline. To ensure that a study meets standards for human subjects' protection, IRB approval or a waiver is needed. Identify the IRB that was consulted and indicate whether and when the approval or a waiver was attained. Describe any additional approvals required and obtained (e.g., school district approval).

Discuss any problems with securing initial IRB approval, if applicable, and any issues that may have arisen during the annual review. Explain why IRB approval was initially denied and whether major modifications were necessary; for example, the IRB may have required changes to the study design, sampling methods, data collection methods, instruments, or analysis plan. Discuss how the delayed approval process affected the study's timeline (e.g., data collection activities were rescheduled).

If any changes to the IRB process or application were necessary, the report should include full details of the new approval, attaching any relevant supporting documentation.

Include any changes to the timeline for data collection, follow-up, analysis, and reporting. The study's timeline should include major events in the evaluation (i.e., design, instrument identification/development, sampling, data collection, analysis, reporting), subcomponents (e.g., baseline, intermediate, and follow-up data collection), and their start and end dates.

Note any changes to the timeline, such as deletions and additions of major tasks, delays, or early completion of major tasks. (Changes in the timeline that affect collection of baseline or follow-up data with respect to participant (e.g., enrollment, treatment) should be fully described in the Data Collection section of the report.)

Explain the reasons for the unexpected changes, how challenges were addressed or mitigated, and how the changes affected the study's progress and budget.

Report any changes to the evaluator/subgrantee personnel and/or role. Report any changes in the evaluation team personnel or role that affected the study's timeline, budget, or potential quality. For example, if the evaluator or evaluation team changed during the evaluation, describe what change was made, when and why, and note the new evaluation team's skills and capacity to conduct the evaluation.

Discuss any changes to the Subgrantee role in the evaluation. This may include a major change in personnel or level of responsibility for tasks. Explain why the changes occurred and how they affected the study's progress. For example, there may have been gaps in staff skills required to complete tasks and the planned technical assistance to build the Subgrantee's capacity was not sufficient. Discuss how changes were addressed and mitigated (e.g., new personnel were hired, the evaluator took on additional tasks).

Report any changes to the budget. The budget reflects the time and cost to achieve every major component of the evaluation. It should include hours and cost by person/position (evaluator and staff), and direct and indirect expenses. Detail any changes that were made to the budget, such as a reduction in projected funding, unanticipated evaluation costs, or cost overruns. Note when these occurred in the timeline of the study, and indicate how/whether the change was addressed or mitigated. Discuss the impact on the study and whether the impact was major or minimal. For example, did a reduction in funds result in a reduced sample size, or did it limit the evaluator's hours allotted to present findings to stakeholders?

Appendix A: Templates and Tools

Appendix A1: Full Evaluation Report Review Checklist

Executive Summary

- The names of the Grantee, Subgrantee (if applicable), and evaluation contractor;
- The program and intended outcomes/impacts;
- Relevant prior research;
- The targeted level of evidence;
- The evaluation design, including comparison/control group approach;
- The measures/instruments;
- The research questions addressed and key findings;
- The analysis approaches used;
- Key updates related to evaluation timing/timeline and budget;
- Key changes to the program or evaluation team; and
- Key next steps for the evaluation and/or program.

Introduction

- The type of evaluation, type of report, and intended audience are identified.
- The theory of change and prior research are briefly discussed, including previous level of evidence.
- The program model is briefly described, including key information such as the number of participants, inputs, components/activities, and key outcomes.
- The targeted level of evidence for the current study is described with specific justification.
- Program implementation questions are clearly stated.
- Program impact questions are clearly stated.
- Changes to the SEP are reported.

Implementation Evaluation

- Study design and procedures for measuring program implementation in the program group are presented.
- Details are provided for how each implementation dimension was measured, including target levels if appropriate.
- Analysis method for assessing implementation is described, including procedures.
- Any preliminary or final implementation analysis findings are detailed.
- Measures are clearly described (or included in an appendix—including a description of the construction and validation of all measures.
- Lessons learned from implementation results are discussed.
- Changes to the SEP are reported.

Feasibility Study

- Barriers to proposing a design with the potential to contribute to strong or moderate evidence are described.
- Full study design is clearly and comprehensively explained.
- Description of the treatment and counterfactual groups are included.
- Where appropriate, assignment of study participants to groups is described.
- The instruments or processes tested are described.
- How this study leads to a study yielding moderate to strong evidence during the SIF timeframe is described.
- Changes to the SEP are reported

Impact Evaluation Design Selection

- The report clearly identifies the study design selected.
- The report justifies the target level of evidence based on a discussion of internal and external validity.

Random Between-groups (Experimental) Design

- Unit of random assignment is clearly identified (and aligned with unit of analysis).
- Procedures to conduct the random assignment, including who implemented the random assignment, how procedures were implemented, and procedures used to verify probability of assignment groups, are described and generated by random numbers.
- Blocking, stratification, or matching procedures used—to improve precision in the estimate of the program effect or to balance groups on measured characteristic(s)—are described.
- The program group and, to the extent possible, the control group conditions are described.
- Procedures and results of an analysis to confirm equivalence of groups are discussed.
- Changes to the SEP are reported

Between Groups Design Formed by Matching

- Reasons why the comparison group might differ from the treatment group, the method of matching used, and how method adjusts for differences are discussed.
- Unit of matching is clearly identified and aligned with unit of analysis.
  Standardized mean differences in baseline characteristics between treatment and comparison groups are described.
- Changes to the SEP are reported

If propensity score matching is used

- Procedures to carry out the matching to form a comparison group are described, including the method and how the propensity score accommodates the study design.
- A description of the variables used in the matching is included.
- Information documenting the success of the matching process is provided, including which variables were associated with treatment group membership, the distribution of propensity scores in the treatment and comparison groups, the proportion of cases matched, and the standardized mean differences in baseline characteristics between treatment and comparison groups.
- How the propensity score was used in the analysis is described. Matched and unmatched results are provided for comparison.

If no propensity score matching is used

- Methods used to form the proposed comparison group are described such that the validity of the matching is explained and documented.

Regression Discontinuity Design

- The measure, cut-off score, and bandwidth selected are aligned with the unit of analysis.
- The methods used to apply the cut-off score are detailed.
- The pretest measure, distribution of measure, and bandwidth are as expected.
- The treatment and counterfactual conditions are as proposed.
- The details of the model specification and how the model was estimated are provided.
- Changes to the SEP are reported.

Single Group Design

- Each intervention phase of the design, including the baseline condition, is clearly described.
- Number of measures during each measurement phase is detailed.
- Number of measures during each measurement phase is sufficient to establish trend and rule out rival explanations.
- Timing of measures pre/post interruption is appropriate to the intervention.
- Changes to the SEP are reported

Interrupted Time Series

- The number of measures in the pre and post-intervention is sufficient to establish a trend and rule out rival explanations.
- The timing of measure pre and post-intervention is appropriate to the intervention.
- Comparison cases are clearly described.
- Treatment and counterfactual conditions are as proposed in the SEP.
- Changes to the SEP are reported

Non-Experimental Design

- Barriers to proposing a design with the potential to contribute to strong or moderate evidence categories are described.
- Full study design is clearly and comprehensively explained
- Description of the treatment and counterfactual groups (if any) are included.
- Where appropriate, assignment of study participants to groups is described.
- Additional threats to the internal validity of the design are discussed.
- Ways future evaluations can be designed to rule out these threats are described.
- Changes to the SEP are reported.

Study Participant/Sample Flow (Impact Studies)

Participant Flow Description

- The flow of participants through the study is described, including the sample size for each set of data collected from intervention and comparison groups at different time points.
- The composition of the sample is described, including demographics and other characteristics relevant to the study.
- Changes to the SEP are reported.
- Sample Recruitment and Retention
- There is a description of recruitment and retention strategies and efforts.
- Sample retention rate is reported.
- How overall and differential attrition was assessed is detailed.
- Any differential attrition findings are reported.
- Changes to the SEP are reported.
- Non-Response Bias and Missing Data
- Results of assessment and adjustment for potential biases (due to non-consent and data non-response) are included.
- Statistical procedures used to adjust for missing data are discussed.
- Changes to the SEP are reported.

Measures

- How each variable used to measure outcomes and impacts in the study is detailed.
- Content and timing are described.
- Updates to or findings on measure construction, reliability, and validity are described.
- Changes to the SEP are reported.

Data Collection Activities

- A description of who collected the data is included.
- A description of the role of staff members delivering the intervention with regard to data collection is described.
- The timing of data collection, relative to delivery of the program, is explained.
- Discussion of whether the mode of data collection is the same for the intervention and control groups is included.
- Changes to the SEP are reported.

Secondary/Administrative Data

- Reasons for using secondary/administrative data are provided.
- The data and source(s) are described.
- The steps to receiving and storing the data are described.
- Overlap between, or coordination of, SIF study data collection and secondary/administrative data can be determined. Details of any characteristics unique to either the treatment or comparison group are provided.
- Data construction, cleaning, and merging procedures are provided, including any recalibration of data structure and weights.
- A full description of the final analysis dataset is provided, including details of variables included and generalizability.
- Any problems with agreements for data access, storage, use, and reporting, as well as details of any changes to MOU are provided.
- Details of MOU, and any strategies/relationships that will facilitate data delivery and/or use are provided.
- Changes to the SEP are reported.

Statistical Analysis of Impacts

- A clear description of the steps of the analysis is provided.
- How the statistical analysis of the data is aligned with the research questions is explained.
- How the statistical analysis is aligned such that the unit of analysis corresponds to the unit of assignment is described.
- Model estimation procedures are included.
- If applicable, covariate adjustments to estimated program effects, estimation of standard errors, and corrections for multiple comparisons are described.
- If sample sizes do not align with target levels established in the SEP, then supplementary power analyses and results are described.
- Organized by research question, any preliminary or final impact analysis findings are detailed.
- Lessons learned from impact results are discussed.
- Changes to the SEP are reported.

Conclusions - Findings, Lessons Learned, and Next Steps

- Summary of Implementation findings is provided.
- Summary of Impact/Outcome findings is provided.
- An explanation of the level of evidence to which this study contributes is provided.
- An explanation of the level of evidence that will be targeted in future reports and how this differs from the SEP is included.
- A discussion of lessons learned about the program and the process of evaluation is provided.
- A description of next steps for the program and for comparable programs is provided.
- Changes to the SEP are reported.

Other Aspects of Study Logistics and Feasibility

- Any problems with securing IRB approval is detailed, and impacts on the study timeline are discussed.
- Any changes to the timeline for data collection, follow-up, analysis, and reporting are included.
- Any changes to the evaluator/subgrantee personnel and/or their roles are reported.
- Any changes to the budget are reported.

Appendix A2: Reporting Study Results in the Report Text, Tables, and Figures

There are a few guidelines for presenting study results in the report text, tables, and figures in a clear and concise manner. These guidelines have been adapted from the Publication Manual of the American Psychological Associations (Paiz, et al. 2014).

Examples of properly formatted tables and figures are available on pages xx –xx.

Overview of Reporting Guidelines for Tables and Figures
1. Titles, abbreviations, headings, and formatting should be consistent across all tables and figures in the report.

2. All results discussed in the text of the report should also be contained in an appropriate table or figure. All results shown in tables or figures should be discussed in the report text.

Example: As shown in Table 2, the treatment group experienced five more events than the control group.

3. For all confirmatory outcomes, the report should indicate whether findings are statistically significant and provide effect sizes. Regardless of whether a finding is statistically significant, the effect size should be provided.

Example: As shown in Table 3, 50% of the treatment group experienced an improvement in the outcome compared to 20% of the control group. A two sided t-test showed the difference in outcome improvement between the treatment and control group is statistically significant ($p<.01$) and substantively large; the effect size (using Cohen's d) is .55.

4. A reader should be able to glean all relevant information from a table without any redundancy.

5. The type of data, model, and/or analysis should be presented in the title of the table or caption of the figure.

6. All figures and tables should be sequentially numbered throughout the report. Any appendix tables and figures should continue the numbering from the main text.

Example: Table 1,2…5; Figure 1, 2…5.

7. All tables and figures should clearly indicate the sample size.

8. We recommend reporting the appropriate measure of variance for all tables and figures eg. SD, SE, etc.

9. The statistical significance of p-values should be reported as follows: † p<.10 *p<.05 **p<.01 ***p<.001

10. Notes should be used to provide additional relevant information on the data source, analytic sample, and method of analysis for all tables and figure. Notes may be appropriate if a sub-sample analysis is being conducted, if data is multiply imputed, if standardized coefficients are shown, if robust standard errors are reported, if p-values have been adjusted for multiple comparisons.

11. For public reports, it is common for small sample sizes be omitted or rounded. Generally, it is recommend to not report cells less than 10.

12. Examples of Properly Formatted Tables and Figures
Examples 1a and 1b:

Table 1a
Means and Standard Deviations on the Post-Treatment Measure of Common Core Standards Score by Group Assignment

		Common Core Standards Test Score	
Group Assign-ment	N	Mean	SD
Treatment Group	45	85.9	3.1
Control Group	45	76.4	2.7
Total	90		2.9

Table 1b
Results of t-test and descriptive statistics on the Post-Treatment Measure of Common Core Standards Test Score by Group Assignment

	Group Assignment								
	Treatment Group (n=45)				Control Group (n=45)				
	Mean	SD	n		Mean	SD	n	t	df
Common Core Standards Score	85.9	3.1	45		76.4	2.7	45	15.50***	88
† p<.10 *p<.05 **p<.01 ***p<.001									

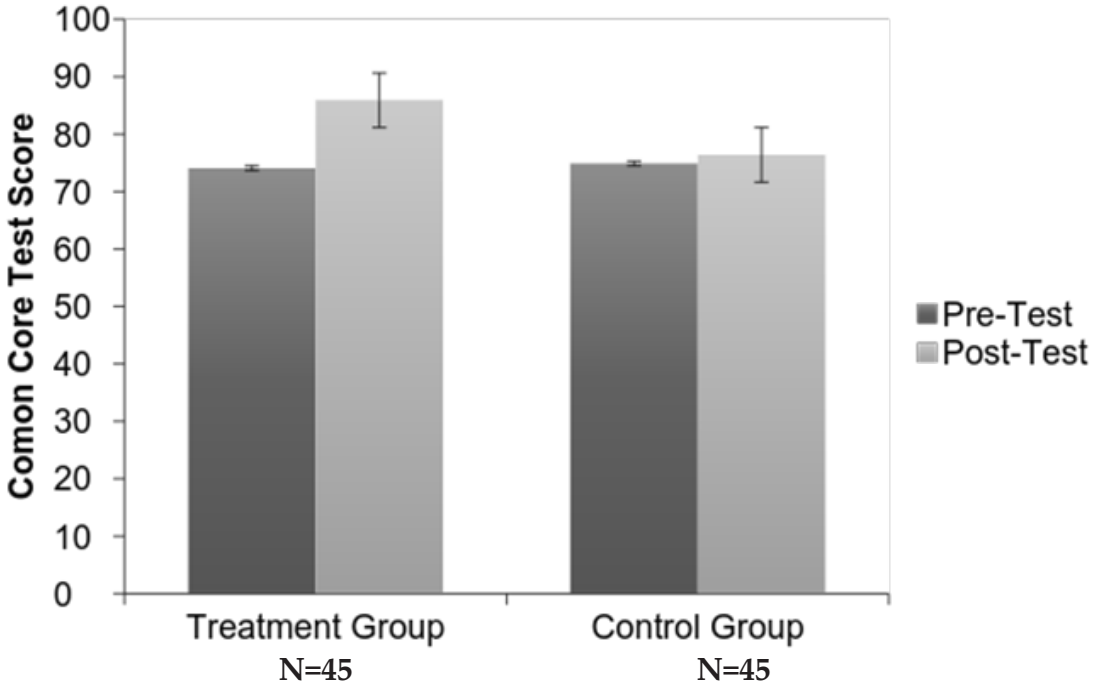

Figure 1. Student pre- and post-test Common Core test scores by group assignment.

Example 3:
Table 2
Summary of Multiple Regression Analysis for Variables Predicting Common Core Standards Scores (N=90)

	Model 1			Model 2		
Variable	B	SE B	β	B	SE B	β
Child Age				0.13	0.25	0.06
Mother's Education				0.21	0.07	.13**
Treatment Group (1=Treatment, 0=Control)	0.58	0.27	.12*	-0.98	0.56	-0.08
N	90			90		
R2	0.05			0.16		
F for change in R2	2.67			9.96**		
Notes: Child age and mother's education are centered at their means. B shows coefficients from OLS model; β shows XY standardized coefficients from OLS model.						
† p<.10 *p<.05 **p<.01 ***p<.001						

Appendix B: Assessing Outgoing Level of Evidence

At the final reporting stage, completed SIF impact studies are assessed based on a range of criteria in three main areas:

- Level of Study Rigor (based on final level of internal validity, external validity, and overall quality of study design execution),
- Quality of Program Model Implementation, and
- Strength of Study Impact Findings on Confirmatory Outcomes.

These criteria are evaluated on the information captured about the completed impact study by the SIF Final Evaluation Report Review Form and Final Report Acceptance Form. Copies of these forms can be found on the knowledge network or from your Program Officer.
The criteria assess study rigor dimensions that are critical for: a) ensuring that the study produced rigorous, scientifically-valid estimates of program impact and b) determining the final, outgoing evidence level tier (strong, moderate, preliminary) for the completed impact study.

Below is a checklist summary of how final reports are assessed to determine the achieved level of evidence.

Criteria	Mets?		Rating/Description/Note
	Y	N	
1. Overall Study Quality: Average Final Report Review Rating of >= 4.5			
2. Study Design Type: RCT or QED (Between Groups with Matching, Between Groups w/Cutoff (RDD), Single Group (Interrupted Time Series)			
3. Study Design Rigor (Quality of Design & Implementation) - Research Design Rating>= 4			
4. Fidelity-Implementation Study Rating of >=4 (and no attrition issue, differential attrition <= 20%)			
5. Internal Validity Threats: - <=2 threats to IV and Mortality (attrition) not an issue (see #3)			
6. External Validity - External Validity Item #1 and/or #2: = Strengthened a. If Multisite, study must have >= 2 sites or >= 2 diff. population			
7. Outcomes – Effective Evidence (all 3 criteria below must be met) 7.1-At least one, positive, significant finding for at least one confirmatory outcome OR if no positive, significant confirmatory outcomes, at least one positive, significant finding for an outcome identified in the study's logic model. i. If multi-site, for either of these conditions, positive, significant findings must be found for 2 or more sites.			
7.2-No negative intervention effects on confirmatory outcomes (No demonstrated significant negative intervention effects).			
7.3-Practical significance of a moderate or large effect on at least one confirmatory outcome OR another outcome identified in the study's logic model.			
8. Contingencies - NO contingencies that introduced threats to the scientific validity of the study.			

Final Outgoing LOE Rating: _____Strong _____Moderate _____Preliminary

Appendix C: References and Resources

DATA COLLECTION ACTIVITIES

What Works Clearinghouse. (2008). WWC procedures and standards handbook (Version 3.0). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf.

U.S. Department of Health and Human Services, Office of Research Integrity. Data Management. Retrieved from http://ori.hhs.gov/data-management-0 .

Evaluation and Feasibility Studies
Bowen, D. J., Kreuter, M., Spring, B., Cofta-Woerpel, L., Linnan, L., Weiner, D., . . . Fabrizio, C. (2009). How we design feasibility studies. American journal of preventive medicine, 36(5), 452-457. Retrieved from http://ac.els-cdn.com/S0749379709000968/1-s2.0-S0749379709000968-main.pdf?_tid=e0028068-353f-11e6-9a82-00000aab0f6b&acdnat=1466246093_7c8a390d1659ec66794ed0709504a61e

Leeuw, F. L., & Vaessen, J. (2009). Impact evaluations and development: NONIE guidance on impact evaluation: Network of networks on impact evaluation.

Scriven, M. (2012). Evaluating evaluations: A meta-evaluation checklist.

Trevisan, M. S., & Huang, Y. M. (2003). Evaluability assessment: a primer. Practical Assessment, Research & Evaluation, 8(20), 2-9.
Human Subjects Protection/IRB

National Institutes of Health. Human Subjects Protection and Inclusion of Women, Minorities, and Children. Retrieved from http://archives.nih.gov/asites/grants/05-29-2015/Grants/peer/guidelines_general/Human_Subjects_Protection_and_Inclusion.pdf

US Department of Health Human Services. Code of Federal Regulations, 45CFR46. Protection of Human Subjects.

Implementation Studies

Bickman, L., Riemer, M., Brown, J. L., Jones, S. M., Flay, B. R., Li, K.-K., . . . Massetti, G. M. (2009). Approaches to measuring implementation fidelity in school-based program evaluations. Journal of Research in Character Education, 7(2), 75.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. Health education research, 18(2), 237-256.
Kovaleski, J. F., Gickling, E. E., Morrow, H., & Swank, P. R. (1999). High Versus Low Implementation of Instructional Support Teams A Case for Maintaining Program Fidelity. Remedial and Special Education, 20(3), 170-183.
Interrupted Time Series
Ramsay, C. R., Matowe, L., Grilli, R., Grimshaw, J. M., & Thomas, R. E. (2003). Interrupted time series

designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. International journal of technology assessment in health care, 19(04), 613-623.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference: Houghton, Mifflin and Company.
Latent Variable Analysis
Heckman, J. J., Tobias, J. L., & Vytlacil, E. (2000). Simple estimators for treatment parameters in a latent variable framework with an application to estimating the returns to schooling. Retrieved from http://www.nber.org/papers/w7950

Jung, T., & Wickrama, K. (2008). An introduction to latent class growth analysis and growth mixture modeling. Social and Personality Psychology Compass, 2(1), 302-317.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. Psychometrika, 55(1), 107-122.

Preacher, K. J. (2008). Latent growth curve modeling: Sage.
Measures
Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of applied psychology, 78(1), 98.

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. Journal of extension, 37(2), 1-5.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74(1), 107-120. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2792363/pdf/11336_2008_Article_9101.pdf
Missing Data
Allison, P. D. (2002). Missing data series: Quantitative applications in the social sciences. Thousand Oaks. CA: Sa ee.

Enders, C. K. (2010). Applied missing data analysis: Guilford Press.

Laird, N. M. (1988). Missing data in longitudinal studies. Statistics in medicine, 7(1-2), 305-315.

Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data: John Wiley & Sons.
Multilevel Modeling
Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models: Cambridge University Press.

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2013). Multilevel and longitudinal modeling with IBM SPSS: Routledge.

Kreft, I. G., Kreft, I., & de Leeuw, J. (1998). Introducing multilevel modeling: Sage.

Non-Experimental Design
Price, P. C., & Jhangiani, R. (2013). Research Methods in Psychology: Core Concepts and Skills.

Power Analysis

Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results: Cambridge University Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior research methods, 41(4), 1149-1160.

Propensity Score Matching
Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of economic surveys, 22(1), 31-72.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. The Review of Economic Studies, 65(2), 261-294.

Peikes, D. N., Moreno, L., & Orzol, S. M. (2012). Propensity score matching. The American Statistician.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services and Outcomes Research Methodology, 2(3-4), 169-188.

P-Values, Multiple Comparisons, and Effect Sizes

Cao, J., & Zhang, S. (2014). Multiple comparison procedures. JAMA, 312(5), 543-544. Retrieved from http://jama.jamanetwork.com/data/Journals/JAMA/930618/jgm140005.pdf

Cohen, J. (1995). The earth is round (p<. 05): Rejoinder.
Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results: Cambridge University Press.

Goulden, K. J. (2006). Effect sizes for research: A broad practical approach: LWW.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (2016). What If There Were No Significance Tests?: Classic Edition: Routledge.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. Child development, 71(1), 173-180.

Toothaker, L. E. (1993). Multiple comparison procedures: Sage.

Qualitative Methodology
Miles, M. B., Huberman, A. M., & Saldana, J. (2013). Qualitative data analysis: A methods sourcebook: SAGE Publications, Incorporated.
RCT
Falaye, F. (2009). Issues in mounting randomized experiments in educational research and evaluation. Global Journal of Educational Research, 8(1/2), 21.

Kane, R. L., Wang, J., & Garrard, J. (2007). Reporting in randomized clinical trials improved after adoption of the CONSORT statement. Journal of clinical epidemiology, 60(3), 241-249.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., . . . Murphy, S. A. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14), 1355-1360.

Moher, D., Schulz, K. F., Altman, D. G., & Group, C. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet, 357(9263), 1191-1194.

Towne, L., & Hilton, M. (2004). Implementing randomized field trials in education: Report of a workshop: National Academies Press.
RDD

Bloom, H. S. (2012). Modern regression discontinuity analysis. Journal of Research on Educational Effectiveness, 5(1), 43-82.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of econometrics, 142(2), 615-635. Retrieved from http://ac.els-cdn.com/S0304407607001091/1-s2.0-S0304407607001091-main.pdf?_tid=e5deb164-353f-11e6-aaeb-00000aab0f26&acdnat=1466246103_d898d318854b7b257deb223a56fab5e3

Jacob, R. T., Zhu, P., Somers, M.-A., & Bloom, H. S. (2012). A practical guide to regression discontinuity: Citeseer.

Lee, D. S., & Lemieux, T. (2009). Regression discontinuity designs in economics. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.2624&rep=rep1&type=pdf

Lee, H. B. (2008). Analysing data from a regression discontinuity study: A research note. Journal of Research Methods and Methodological Issues, 2(1), 1-9.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J., Porter, J., & Smith, J. (2010). Standards for Regression Discontinuity Designs. What Works Clearinghouse.

Research Synthesis/Meta-analysis

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., & Light, R. J. (1994). Meta-Analysis for Explanation.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). The handbook of research synthesis and meta-analysis: Russell Sage Foundation.

Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results: Cambridge University Press.

Sampling

University of Reading Statistical Services Centre. (2000). Some Basic Ideas of Sampling. Retrieved from http://www.reading.ac.uk/ssc/resources/Docs/Some_Basic_Ideas_Of_Sampling.pdf

Single Group Design

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. Psychological Bulletin, 82(4), 511.

Kratochwill, T., Hitchcock, J., Horner, R., Levin, J. R., Odom, S., Rindskopf, D., & Shadish, W. (2010). Single-case designs technical documentation. What Works Clearinghouse. Retrieved from http://files.eric.ed.gov/fulltext/ED510743.pdf

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference: Houghton, Mifflin and Company.

Statistical Analysis of Impacts

UCLA: IDRE Research Technology Group. Retrieved from http://www.ats.ucla.edu/stat/

Allison, P. D. (1999). Multiple regression: A primer: Pine Forge Press.

Hancock, G. R., & Mueller, R. O. (2010). The reviewer's guide to quantitative methods in the social sciences: Routledge.

Wooldridge, J. M. (2015). Introductory econometrics: A modern approach: Nelson Education.

Structural Equation Modeling/Path Analysis

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. Journal of the academy of marketing science, 16(1), 74-94.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. International journal of Research in Marketing, 13(2), 139-161.

Bollen, K. A. (2014). Structural equations with latent variables: John Wiley & Sons.
Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. The Journal of educational research, 99(6), 323-338.

Survey Methodology

Blair, J., Czaja, R. F., & Blair, E. A. (2013). Designing surveys: A guide to decisions and procedures: Sage Publications.

Burgess, T. F. (2001). Guide to the Design of Questionnaires. A general introduction to the design of questionnaires for survey research.

Cui, W. W. (2003). Reducing error in mail surveys. Practical Assessment, Research & Evaluation, 8(18), 1-5.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009).

Survey methodology (Vol. 561): John Wiley & Sons.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. Public opinion quarterly, 68(1), 109-130.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. Annual review of sociology, 65-88.
Table and Figure Guidance
Acock, A. Tables for NCFR Journals. Retrieved from http://people.oregonstate.edu/~acock/tables/

American Psychological Association. (2009). Publication Manual of the American Psychological Association (6th ed.).

American Psychological Association. (2014). The basics of APA style. Accessed August. Retrieved from http://www.apastyle.org/learn/tutorials/basics-tutorial.aspx

OWL Purdue Online Writing Lab. (2014). APA Formatting and Style Guide. Retrieved from https://owl.english.purdue.edu/owl/section/2/10/
Wilkinson, L. (2006). The grammar of graphics: Springer Science & Business Media.

Corporation for National and Community Service
250 E Street SW
Washington, D.C. 20024
202-606-5000 TTY 800-833-3722
www.NationalService.gov