SOCIAL INNOVATION FUND

Budgeting for Rigorous Evaluation: Insights from the Social Innovation Fund

Budgeting for Rigorous Evaluation: Insights from the Social Innovation Fund

October 2013

Authors

Lily Zandniapour, Ph.D., Corporation for National and Community Service

Nicole Vicinanza, Ph.D., JBS International

Acknowledgements

The authors would like to thank Laissa Lai and Peter Lovegrove for their assistance with this paper. Ms. Lai assisted the authors by compiling and presenting the quantitative data on which this paper is based. Dr. Lovegrove contributed to the paper by conducting statistical analyses on the evaluation and program budget data and examining the relationship between cost measures and other explanatory variables. He also drafted the summary findings from the statistical analyses included in the Appendix.

Citation

Corporation for National and Community Service, Office of Research and Evaluation. (2013). *Budgeting for Rigorous Evaluation: Insights from the Social Innovation Fund*. (by Lily Zandniapour and Nicole Vicinanza). Washington, DC: Author.

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. Upon request, this material will be made available in alternative formats for people with disabilities.

Summary: This paper examines the challenges in allocating budgets sufficient to support rigorous evaluation of Social Innovation Fund projects. It concludes that ratios commonly used for evaluation – such as five to ten percent of program budgets – may not be adequate to meet the level of evidence and rigor required for a comprehensive evaluation. Experimental studies tend to be the most expensive; however, a variety of factors may affect these budgets, including the number of project sites and the targeted level of evidence for the evaluation. This paper includes reflections from grantees and provides a detailed snapshot of budgets for experimental, quasi-experimental, and non-experimental studies. The authors also examine budget trends for studies based on whether they seek preliminary, moderate, and strong evidence of success and the program and design factors that influence evaluation budget estimates.

Background and Overview

This paper provides information, insights, and tips on budgeting for evaluation based on the experiences of the 2010 and 2011 cohorts of Social Innovation Fund (SIF) grantees/intermediaries and subgrantees. The purpose of the paper is to provide useful information to future SIF intermediaries and subgrantees and the broader philanthropic and nonprofit communities to help them better budget for evaluation. SIF requires that intermediaries select subgrantees with experience conducting outcome evaluations or implementing programs with some supporting evaluation evidence. They are required to conduct third party evaluations that advance the evidence base for funded programs and increase the number of interventions with moderate and strong evidence of effectiveness. SIF evaluations include a range of study designs including experimental, quasi-experimental, and non-experimental designs. Many evaluations are a combination of implementation studies and impact evaluations that will provide causal evidence regarding program effectiveness within the timeframe of their SIF funding (usually five years). Evaluation is a key component of the SIF program, and grantees and subgrantees are encouraged to allocate sufficient resources to evaluation to ensure that the commissioned studies produce scientifically valid and rigorous evidence for the supported interventions. A rigorous evaluation is one that embeds sound evaluation principles and practices consistent with scientific research methods into each step of the process to ensure credible and useful results and minimize bias.

The challenge faced by many SIF grantees and sub-grantees, however, is how to estimate "sufficient" level of dedicated funds before hiring external evaluators and developing detailed evaluation budgets. Allocating too much money for evaluation will limit money available to serve program participants, while too little may mean that the evaluation component will not provide the level of evidence required by SIF.[1] Both underfunding and overfunding can become contentious points between the evaluator, the sub-grantee, the grantee, and the SIF. This is a challenge not exclusive to SIF; it is one shared by many grant-making and nonprofit organizations as they are increasingly asked to conduct evaluations of projects or programs to document and demonstrate their effectiveness. Providing useful information on the cost of evaluations and the development of evaluation budgets will address an information gap and a need in the larger social sector field.

When estimated budgets are realistic and adequate to the evaluation's purpose and approach, the evaluation experience can be very positive and informative. Evaluation budgets should be:

[1] For a full definition of the tiers or levels of evidence under SIF (preliminary, moderate and strong level of evidence), please refer to pp. 9-10 of the 2012 SIF Notice of Funding Availability (NOFA) at: Overview of Funding Opportunity (http://www.nationalservice.gov/pdf/12_0210_sif_nofa.pdf) or refer to the glossary at the end of this paper.

- Commensurate with stakeholder expectations and involvement;
- Appropriate for the research design and key research questions;
- Adequate for ensuring quality and rigor; and
- In line with the available program and organizational level resources.

There is limited guidance for programs on budgeting for evaluation. Evaluation budgets are generally considered proprietary as they include compensation information for personnel and/or organizational costs and fees. As such, they are not shared broadly. There are also contextual considerations that affect each evaluation budget and can make overly prescriptive budgetary parameters inappropriate. However, without any guidance, it is difficult for program directors and managers to make informed and sound decisions on the level of resources allocated for evaluation. Further, there is an information gap that makes it challenging for non-evaluators to develop reasonable evaluation budgets. This paper addresses that information gap and provides insights from the SIF experience, which involves a large number of rigorously planned evaluations across its large portfolio.

According to a leading evaluation resource with budgeting guidance: "Generally, an evaluation costs between 5 and 7 percent of a project's total budget."[2] The same resource notes that the specifics of the study can impact the funds required for evaluation and potentially lead to increased allocations for a number of line items in the budget.[3] This rule of thumb ratio has risen to around 10 percent as evaluations have increasingly involved more stakeholder engagement. The SIF experience, however, suggests that these estimates or allocations of a fixed percentage of a program budget for evaluation may not be sufficient, particularly for the rigorous evaluations funded by SIF that are to demonstrate program impact. During interviews with SIF grantees about the evaluation budget process, one grantee representative noted that, although they had advised their sub-grantees to allocate between 10 and 25 percent of their budget to evaluation, "percentages, I think in hindsight, were probably not the best way to do the budget but rather probably a benchmark …..regardless of your program size." Another noted, "I think that this idea of setting a target percentage or a target amount set aside for evaluation is just nearly impossible. I think even across [our] portfolio, there is a huge range of evaluation costs, which doesn't necessarily map to the size of the grants that those sub-grantees got."[4]

Looking across interventions supported by SIF and included in this study[5] (n=70), the average annual program budgets range from $100,000 to $5,460,618with an average of $1,104,649 and median of $593,309.[6]For these interventions, evaluation budgets range from $12,000 to $1,346,342 per year with an average and median costs of $216,838 and $81,471, respectively. Given different study designs and target levels of evidence, the average evaluation-to-program budget ratios vary widely from three percent to 83 percent with average and median of 19 percent and 15 percent, respectively. Thus, the average and

[2] *W.K. Kellogg Foundation Evaluation Handbook*. W.K. Kellogg Foundation, Battle Creek, MI, January 2004, p. 54. The document was developed in 1998 and updated in January 2004 and can be accessed at: W.K. Kellog Foundation Evaluation Handbook (http://www.wkkf.org/knowledge-center/resources/2010/W-K-Kellogg-Foundation-Evaluation-Handbook.aspx).

[3] Ibid. pp. 54-55.

[4] Quotes are drawn from unpublished interviews with SIF 2010 and 2011 Grantees conducted by Education Northwest during 2011.

[5] The date here refers to the date of the writing of the first draft of this paper in October of 2013.

[6] Tables 3, 4 and 5 at the end of this paper includes detailed information on evaluations covered in this paper.

median figures for these projects are much higher than the five to ten percent ratios for evaluation that are suggested in the literature. This data would indicate that studies that aim to establish causal impact, especially randomized control trials (RCTs), will likely require significantly higher percentages than likely expected for a quality evaluation.

To provide better information to support evaluation budget decisions, this paper draws on evaluation budget information developed by evaluators for the 2010 and 2011 SIF cohorts[7], as well as qualitative information from interviews with SIF intermediaries, sub-grantees, and evaluators about their evaluation budgeting and implementation experiences.

Types of Evaluations and Budget Impact

Evaluations range from preliminary, formative, or developmental studies that help programs through the early stages of design to implementation studies that consider various outcomes and impacts to determine the causality, merit, or return on investment of a program or intervention. Each type of evaluation requires different financial resources for external evaluation and the time spent by program staff and participants to support the evaluation. Most SIF evaluations are designed to provide causal evidence of program impacts and are expected to attain high levels of rigor and quality. Almost all SIF evaluations address outcomes or impacts as well as process and implementation. In some cases, the evaluation budget also covers a feasibility study or an evaluability assessment to ensure that conditions for conducting an impact evaluation are present prior to launch of a full impact study.

The SIF requires funded programs to come in with at least a preliminary level of evidence showing the intervention holds promise. The goals of the SIF evaluation program are to advance the evidence base for funded interventions and increase the number of interventions with a moderate or strong level of evidence over a five-year period. SIF grantees provide detailed evaluation plans, typically developed by an external evaluator contracted by the grantee/subgrantee, that include detailed annual and multi-year budgets. Given that some intermediaries or subgrantees have to adjust and often increase evaluation resources after the planning period, the program and evaluation budgets are updated and verified to ensure accuracy and account for subsequent adjustments. The information presented in this paper covers the first 70 evaluation plans that were submitted to CNCS and were either approved or were close to approval. The number of evaluation plans has continued to increase over time, with new cohorts of grantees and sub-grantees starting, and existing cohorts submitting additional plans as they conduct evaluations in a two-step process.[8] These plans detail evaluations ranging from implementation and feasibility studies (typically for the first program year only) and in some cases pre- and post-test assessments using Non-Experimental (NE) designs to multi-site randomized controlled trials (RCTs) lasting multiple years. The largest number of programs, however, selected a Quasi-Experimental design (QED) for the evaluation. The QEDs either use propensity score matching (PSM) or some other form of

[7] Figures presented in this paper are based on program and evaluation budgets provided by 2010 and 2011 cohort of SIF grantees that include a total of 16 awarded intermediaries each with between four and 46 subgrantees. Some intermediaries developed one evaluation plan across their portfolio of subgrantees, while others have one plan per subgrantee. SIF Evaluation Plans (SEPs) go through a process of review and approval by CNCS prior to implementation. Only data from SEPs that are approved or which have conditional approval are included in this study. All data are reported without reference to grantee or subgrantee name.

[8] It is expected that over 100 evaluations will be commissioned through the initiative within the next few years, providing future opportunities to revisit the budgets based on a larger set of data points.

matched comparison group design or use approaches such as Regression Discontinuity Design (RDD) and Interrupted Time Series (ITS) Designs.[9]The following two figures summarize information on the 70 evaluations included in this study. The first graph shows the breakdown of evaluations based on broad categories of design. The second graph shows the breakdown of studies in each design category.

Figure 1: SEP Designs of 2010 & 2011 Cohorts

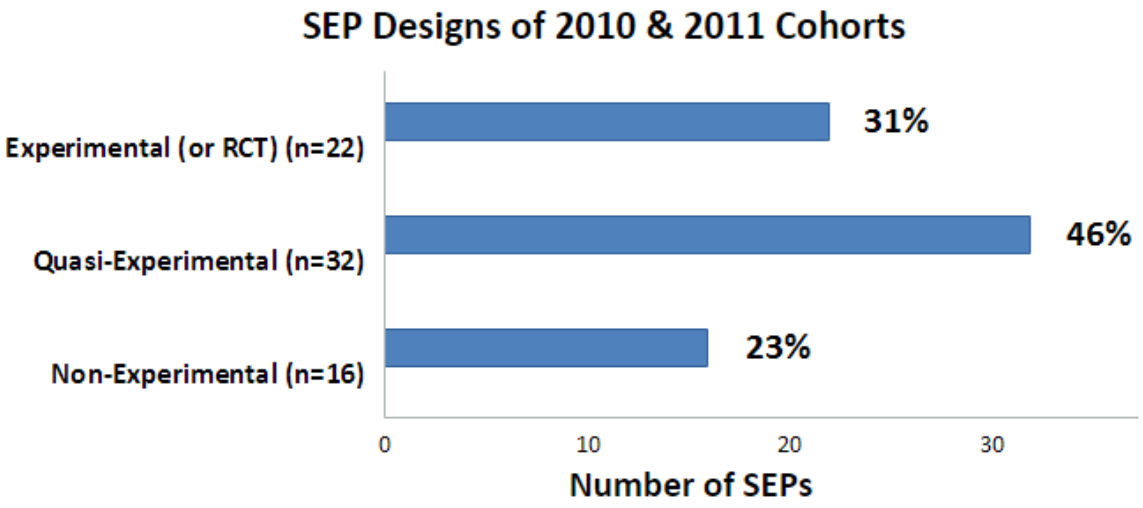
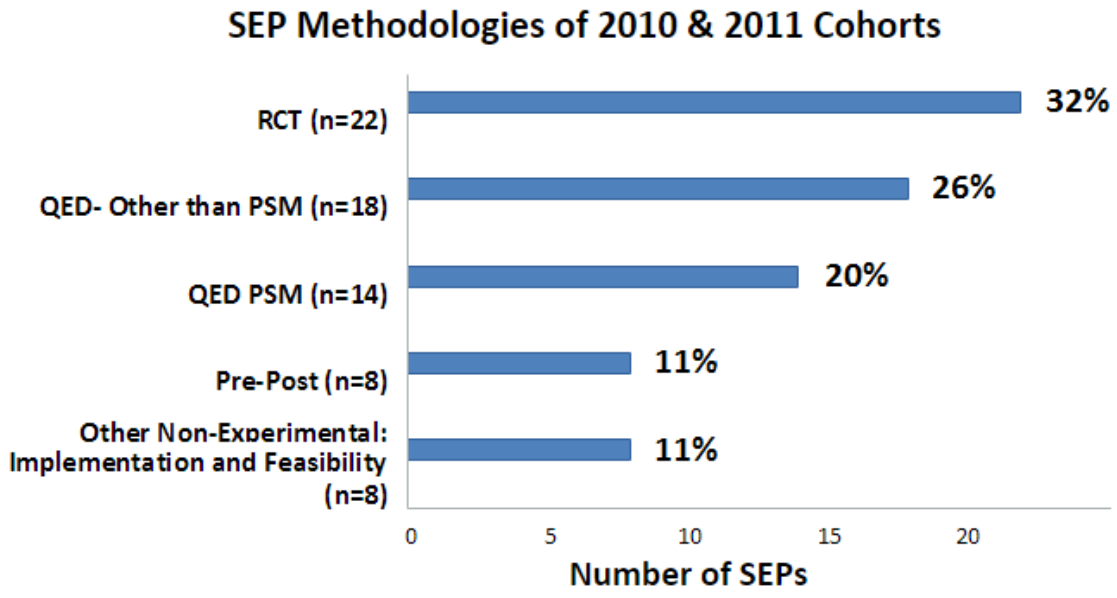


Figure 2: SEP Methodologies of 2010 & 2011 Cohorts



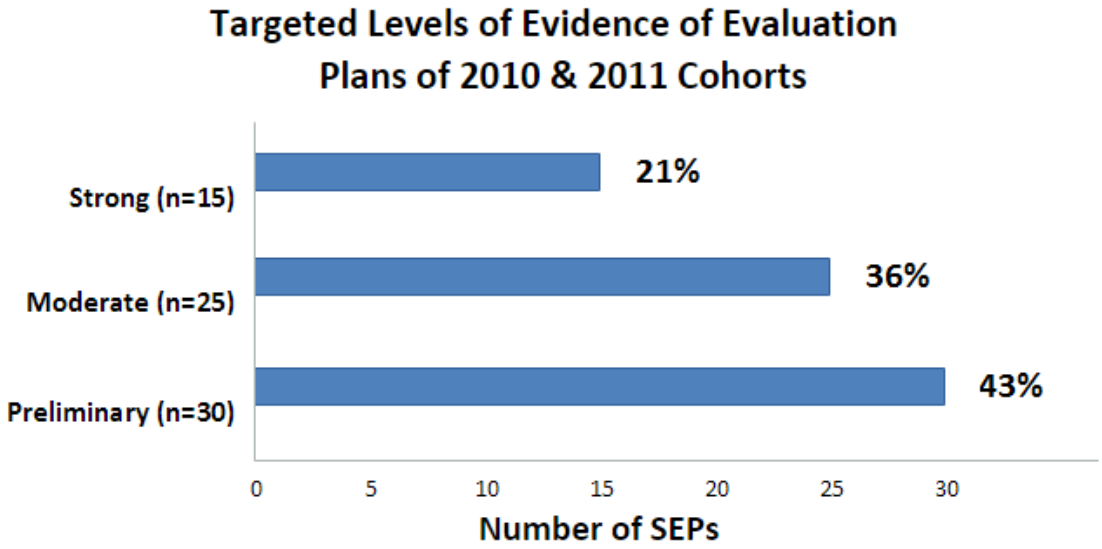
As mentioned earlier, many SIF evaluations include an implementation and an outcomes/impact component. In some cases, the study is a hybrid and uses more than one type of outcome or impact

[9] For more information about these study designs, please see the glossary of terms at the end of the paper.

design. For example, it may involve a quasi-experimental approach and an RCT in one study.[10]

A key benchmark for SIF-funded programs is the level of evidence expected in the evaluation plan. Plans rated as likely to achieve "preliminary" evidence either do not address program impacts or address them in ways that do not eliminate key threats to internal study validity. Plans rated as likely to achieve "moderate" levels of evidence address key threats to internal validity, while plans likely to achieve "strong" levels of evidence both eliminate threats to internal study validity and enhance external study validity. The next graph shows the evidence level for the 70 studies.

Figure 3: Targeted Levels of Evidence of Evaluation Plans of 2010 & 2011 Cohorts



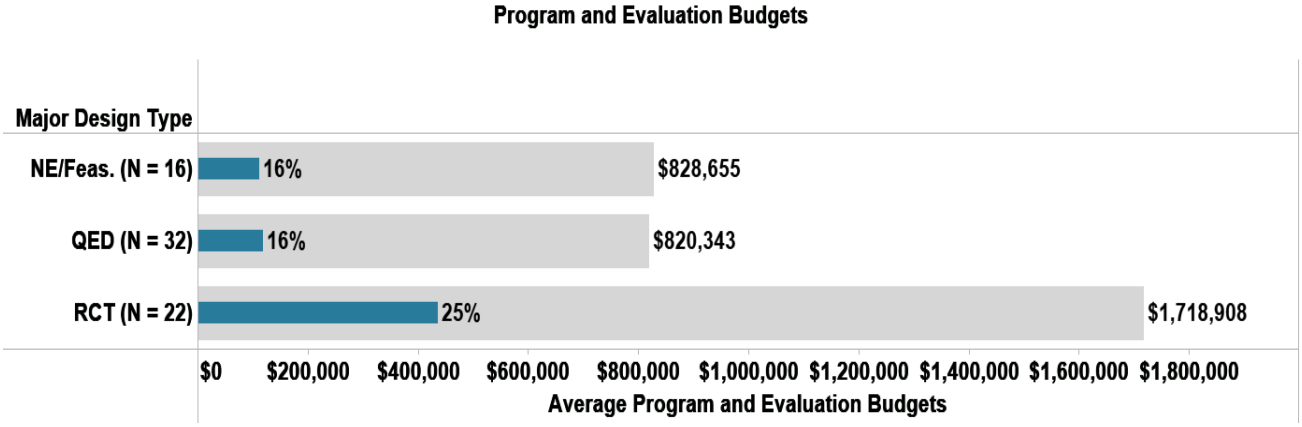
As noted earlier, the average evaluation-to-program budget ratios in SIF programs vary widely from three percent to 83 percent, with an average of 19 percent and a median of 15 percent. The average and median figures for these projects are much higher than the five to ten percent ratios suggested in the literature. While program and evaluation budgets differ based on study design and the target levels of evidence, it is clear that the evaluation trends within SIF likely require higher allocations for evaluation.

Study Designs

The type of study undertaken by the SIF project has major implications for the cost of evaluation, with RCTs requiring the highest percentage of allotted funds and non-experimental and QEDs having the lowest percentages (Figure 4). This figure also shows not only the percentage of program budget for evaluation, but also the range of annual costs for such designs.

[10] For consistency, studies were labeled in this paper based on the highest type of study design in a particular evaluation. Studies that used a combination of pre and post outcome design and a quasi-experimental design are categorized as QED and ones with an RCT and a QED are counted as RCT.

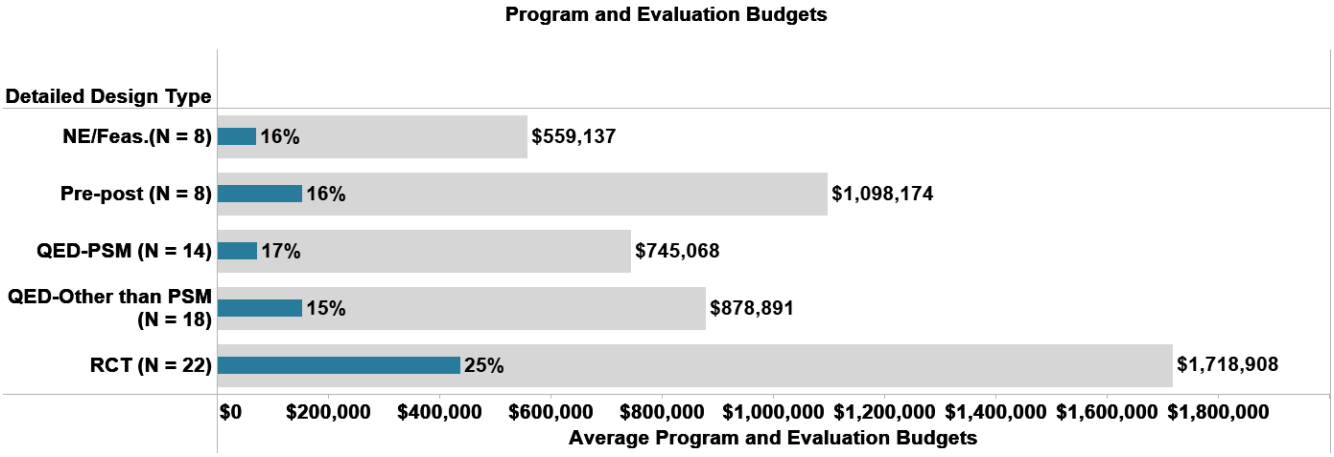
Figure 4: Annual Program Funding and Percent Allocated to Evaluation by Major Design Type



Non-experimental studies devote three to 45 percent to evaluation, while QEDs have evaluation amounts ranging from three to 56 percent. However, both types of studies have evaluation budgets with an average of 16 percent and a median of 13 percent. By comparison, designs using RCTs have the highest budgets in raw dollars and percent of program budget. RCT budgets range from nine to 83 percent of total program budgets, with average and median ratios of 25 percent and 22 percent, respectively. Looking at the dollar figures involved, the average cost of RCTs per year is almost four times that of non-experimental and QEDs, and the median cost of RCTs per year is over eight times higher than the median cost of non-experimental and quasi-experimental design studies.

Figure 5 shows that even among specific types of QED and non-experimental designs, RCTs still have substantially higher budgets. Non-experimental design studies are split into two groups: outcomes studies with a pre- and post-test approach and feasibility studies and other types of evaluations including implementation studies. The average annual cost of studies with a pre- and post-test approach ($153,014) is more than double the cost of other implementation and feasibility studies ($69,932). Evaluations using QEDs cover a number of different approaches, including propensity score matched designs (14 studies) and alternative matching approaches or designs such as Interrupted Time Series or Regression Discontinuity. Propensity-score-matched studies, on average, cost $71,898 per year, or 17 percent of program budgets, while other quasi-experimental studies cost $154,005, or 15 percent of program budgets.

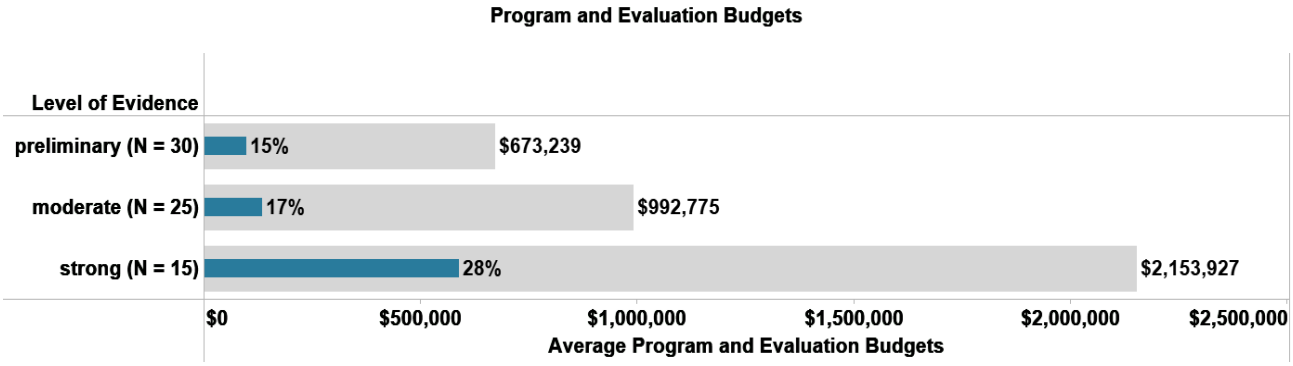
Figure 5: Annual Program Funding and Percent Allocated to Evaluation by Specific Design Type



Target Levels of Evidence

The level of evidence targeted by the study also has a major effect on evaluation costs. Overall, SIF evaluation studies become increasingly expensive as they move from a preliminary level of evidence to those targeting moderate and strong levels of evidence. As seen in Figure 6, evaluation costs as a proportion of average annual program budget and the overall evaluation costs are substantially higher for studies targeting strong level of evidence (28%, or $589,037) compared to those targeting moderate (17%, or $135,976) and preliminary levels of evidence (15%, or $98,123). Evaluations with strong levels of evidence are required to enhance external validity, which is typically done by conducting a large, multi-site study. Primary data collection for a larger number of study participants across multiple locales involving more stakeholders and the complexity of data analysis required are among the factors that drive costs. Therefore, it makes sense that evaluation costs for this level are higher. Studies that target strong level of evidence are six times more expensive as those targeting preliminary level of evidence and more than four times more expensive as those targeting moderate level of evidence.

Figure 6: Evaluation Costs by Levels of Evidence



To assess the validity of observed findings suggesting that RCTs and studies with strong level of evidence are more costly than other designs and approaches, a series of statistical analyses was

conducted on the data.[11] These analyses support the stated conclusions and ruled out cost differences due to other explanations such as a limited number of data points, characteristics of a few outlier studies or chance. (For more information on these statistical analyses and other findings regarding range of evaluation budgets across study types and level of evidence, see the Technical Appendix to this paper).

Revisiting Evaluation Budgets

In examining evaluation budget figures, it is important to note that SIF grantees estimated their evaluation budgets when they submitted their applications for funding. But, as noted by SIF Program Officers, some have adjusted their evaluation budgets over time as they learned more about evaluation expectations and goals and the costs involved in meeting those requirements. One grantee noted: "*I think we may have to rework our budget in years four and five. I think doing a moderate design would require more money for either the subgrantees and/or [our external evaluator].*"

Although some grantees had conducted similar evaluations in the past and were able to start with reasonable budget estimates, program officers noted that others had to revisit evaluation budgets after the award to increase funding. Based on these reports, it seems reasonable to assume that a substantial number of evaluation budgets are still likely on the lower bound of feasibility for the planned study designs. Further, some grantees and subgrantees that targeted higher levels of evidence could not develop plans within their budgets that achieved them. One grantee noted: *"On one [subgrantee evaluation plan], we dropped it from strong to moderate, because of what it would cost actually, to get to strong, that is not in the budget."*

Based on data gathered on SIF studies, including those presented here, the following key findings emerge:

- Evaluation costs and evaluation-to-program budget ratios vary based on the study design chosen and increase with designs that seek to establish causal impact.
- The rule-of-thumb ratios in use to date (i.e., 5% to 10%) lead to serious under budgeting of evaluations, especially for studies addressing both implementation and impact. The minimum percentage would appear to be 13 to 15 percent for non-experimental studies. Available data indicates that between 15 and 20 percent is more realistic for single site QEDs, and RCTs require 25 percent or more.
- Using a percentage of program budget may not be ideal for allocating evaluation funds. Evaluation and program costs should be considered in absolute dollar amounts as well as relative terms. For example, one likely cannot conduct an evaluation that targets a moderate level of evidence as defined by the SIF for less than $75,000 per year, unless the study is subsidized through in-kind contributions from the evaluator.
- The price of evaluation goes up as the level of desired evidence increases. Strong evidence is disproportionately more expensive. Preliminary and moderate studies are closer in cost, as the driving factor may be the number of sites in the evaluation.
- All design types have the potential to be expensive depending on the scope and number of sites in the study. For example, even implementation studies seeking preliminary evidence can be costly if programs operate across a large number of sites.

[11] Statistical analyses based on Bayesian multiple regressions support the stated conclusions. A more detailed and technical description of these analyses is included in the Appendix to this paper.

- It is not possible to conduct a rigorous evaluation on a shoestring budget. To conduct a robust evaluation for a high level of evidence, project leaders must be willing to budget accordingly.

One factor that may help grantees is to involve evaluators earlier in the budget process. This may be particularly helpful for projects with a strong level of evidence, as they may engage larger evaluation firms with detailed experience in budgeting. As one grantee noted, "*When we applied to the SIF our evaluation partner was part of our application, and they helped us craft the application around the…goals for the evaluation. And so we've had an easier time than I think a lot of the SIF grantees because we already had an evaluation partner in the process from the beginning.*" Given the cost variations for different types of studies, it is also important to consider the following issues when developing a budget estimate:

- Program factors, including the number of and distance between sites, the program services offered, and the type of population targeted;
- Design factors, such as key research questions, data collection strategies (including development and use of surveys), data sources, and amount of time to conduct data analysis;
- Management factors, including the need to build staff capacity for the desired evaluation; and
- Dissemination and use factors, such as the amount of time to document findings, prepare reports/deliverables, reflect on evaluation findings, and develop formal communications and dissemination plans to share evaluation results.

In addition, although these are typically not included in evaluation budgets, program planning and budgeting should take into account any time of program staff to help collect, enter, or clean data. Staff time will also be needed for project administration, monitoring, and quality control.

Tips for Evaluation Budget Planning

Given these factors, it is important that grantees address two key cost areas when planning for evaluation. The first is the cost of the evaluation itself, including the time, materials and other direct costs expended by the evaluation team on evaluation activities. These are typically listed as line items in a detailed evaluation budget. The second category is program costs for supporting the evaluation. While not typically included in an evaluation budget, these items include program staff and volunteer time spent in evaluation planning, oversight and supervision, data collection, entry and review, report development, program staff travel to support the evaluation, and other in-kind support. Failure to allocate sufficient resources in both areas can negatively impact the quality of the evaluation. One grantee noted that program costs for evaluation were an issue they addressed up front:

"*The evaluation itself will impose a burden on our program, and not an insignificant burden. They'll have to do significantly more work to make sure that the data is there for the evaluation, particularly for the comparison group. And the SIF …compensates them for their time in participating in the evaluation…The SIF creates just a context that will encourage people to provide higher quality data and to spend the time to ensure that we get all the data we need…. We've been clear with our grantees that the program is as much an evaluation program as it is a capacity building or, funding program for our grantees. And I think they've responded to that.*"

The following chart includes typical costs to include in a detailed budget, broken out by those costs incurred by the evaluation team and those incurred by the program funding the evaluation.

Costs to include in a detailed budget	**Evaluation Team**	**Program**
Staff time to conduct, or support[12]:		
Evaluation planning (including development of written evaluation, sampling, analysis, and reporting plans if needed)	x	X
Instrument selection, development, and any needed validation	x	X
Development/revision of IRB packages	x	X
Data collection	x	X
Data entry, cleaning, and coding	x	X
Data Analysis	x	
Reporting, including both funder required and evaluation specific reports	x	
Review and acceptance of reports		X
Travel required by the evaluation (e.g., to and from data collection and reporting activities)	x	
Interfacing for project and contract management	x	X
Development of evaluation capacity building/training activities	x	
Participation in training/capacity building	x	X
Travel[13]		
Airfare	x	X
Ground transportation	x	X
Lodging	x	X
Per diem/meals/incidental travel costs	x	X
Other Direct Costs[14]		
Communications—postage, telephone calls, etc.	x	X

[12] Evaluation staff and subcontractor salary and benefits and consultant time to conduct activities. Note that some contractors may provide separate line items for salary and benefits, while some may present a single, loaded rate, which includes salary, benefits, indirect rates, and fees.

[13] Travel expenses for staff and/or evaluators should be included as a line item in the budget. The travel costs vary from project to project with those across multiple sites around the country likely needing larger travel budgets compared to ones located in one site. Evaluator proximity can also affect travel costs. There may be travel costs associated with communication and dissemination plans. Data collection and capacity building components may also require travel by staff and/or consultants, etc. Ideally, travel should be estimated in association with specific evaluation tasks such as data collection or reporting.

[14] Other costs associated with the evaluation should be detailed in the budget. Note that the purchases of supplies and equipment may be limited by the funder to those that are specific to the evaluation and not re-useable on other efforts. Supplies used across multiple evaluations should be included in overhead costs.

Costs to include in a detailed budget	Evaluation Team	Program
Printing and copying—including both task-specific copying and general duplication.	x	X
Supplies and equipment that must be purchased or rented for the evaluation.	x	X
Overhead costs and fees (including indirect costs)****[15]	x	X

Conclusion

This paper is designed to provide readers with more data and guidance around budgeting for evaluation and to help program staff avoid the pitfalls of under-investing in evaluation. The specific evaluation needs, evaluation design, level of evidence targeted and other factors noted in this paper will affect evaluation costs. The average dollar figures for evaluations and percent of program budgets cited here may be helpful in making budget allocations. Other information provided on budget considerations may better prepare readers as they approach the budget development process.

In addition, below are other suggestions to lay the groundwork for developing evaluation budgets:

- Start by thinking about evaluation expectations. These can include requirements of funding agencies (i.e., evaluation requirements articulated by the Notice of Funds Availability in the case of SIF and CNCS or other funder requirements), as well as expectations of program staff and other internal or external stakeholders.
- Consider the level of engagement expected from stakeholder groups, including boards of directors, senior leadership, program and evaluation personnel and other stakeholders. How much time will different groups devote to the planning, implementation, and dissemination of evaluation findings? How much time and effort will be needed to ensure buy in and uptake?
- Consider how the evaluation approach will impact internal capacity building. For this effort, estimate the internal and external costs of such activities (trainings, technical assistance, coaching), the modes of delivery for these activities, and their intensity.
- Identify the key partners (evaluation firm(s), consultants, program partners, capacity builders, etc.) for this work to move forward.
- Review existing evidence from past evaluations and determine how to advance that evidence base. What approach would make sense to evaluate the program given the type of design and the targeted level of evidence?
- Develop, or request from an evaluation contractor, a detailed budget worksheet and estimated time/costs on a weekly, monthly, or annual basis (as appropriate) for all major categories relevant to evaluation needs. The timeframe reflected in the budget worksheet should cover the timeframe for the evaluation. If appropriate, the timeframe may need to be extended to cover relevant preparation and follow-up time as well.
- Get extensive input on the estimated budget to ensure completeness.

[15] These costs are typically allocated as a percentage of the overall budget and include expenses involved in operating a business including rent, computers, software licenses, utilities, and other business costs. For-profit firms also may include fees in their detailed budget categories.

- Plan for contingencies that can arise over the course of the evaluation. As one grantee noted: *"It's clear we can't always anticipate challenges - so expect not to be able to do that and plan accordingly with staff time and some contingency funds."*
- Be prepared to revise and adjust budgets as projects move from planning to implementation. Often, budgets must be revisited due to realities on the ground over the project timeline. One grantee provided an example of why its budget had to be adjusted: *"Recruitment challenges and small number of participating sites are extending study timeline and therefore, costs."*
- Make sure the evaluation process and findings are used to inform and improve this work.

About the Social Innovation Fund

The Social Innovation Fund is a federal program intended to foster innovation to transform lives and communities. A program of the Corporation for National and Community Service (CNCS) launched in 2010, it is one of six Obama Administration "tiered-evidence initiatives" embodying the principles of social innovation. As a program, it leverages federal funds through public-private collaborations by granting money to highly successful intermediary grant makers who in turn find, improve, and grow promising community solutions with evidence of successful outcomes in three core areas: youth development, economic opportunity, and healthy futures.

The SIF is characterized by the unique interplay of six key elements:

1. It relies on intermediary grant making institutions to implement the program – they take on the role of finding, selecting, monitoring, supporting, evaluating, and reporting on the nonprofit organizations implementing community-based interventions.
2. It is a tier-based evidence program that requires all funded programs/interventions to demonstrate at least preliminary evidence of effectiveness, or funding "what works."
3. It requires that all programs or interventions implement a rigorous evaluation that will build on their level of evidence.
4. It charges intermediaries with scaling evidence-based programs – increasing impact within their community or to communities across the country – and as such, grapples with a field-wide challenge of how best to successfully and efficiently do so.
5. It leverages public-private partnerships to effect large scale community impact in ways that neither a traditional federal grant investment nor a philanthropic grants investment could achieve on its own. This includes its unique leveraged funding model to support nonprofit programs.
6. The SIF is committed to improving the effectiveness of nonprofits, funders, and other federal agencies by capturing learning and best practices and promoting approaches that will generate the greatest impact for individuals and communities.

Glossary of Key Terms

SIF: The Social Innovation Fund (SIF) is a federal tiered-evidence initiative administered by the Corporation for National and Community Service. The program leverages federal funds through public-private collaborations by granting money to select intermediary grant makers who in turn find, improve, and grow promising community solutions that address pressing social challenges and needs.

CNCS: The Corporation for National and Community Service (CNCS) is an independent federal agency with the mission to improve lives, strengthen communities, and foster civic engagement through service and volunteering. The agency administers the SIF.

Focus Areas: Programmatic areas that address similar issues or challenges. The three focus areas that SIF supports are: youth development, healthy futures, and economic opportunity.

Evidence-based Initiative: In this context, a federal initiative that seeks to promote programs rooted in science and research. There currently are six federally funded evidence-based initiatives, of which SIF is one.

Intermediary: A non-profit, grant making organization that can apply for SIF funds. If chosen through CNCS' grant-making processes, this organization will select a number of subgrantees to participate in its SIF portfolio and to which it will disburse SIF funds. An intermediary is required to match SIF funds one-to-one.

Subgrantee: A non-profit organization that implements a program aimed at addressing a social or community challenge. These organizations are chosen to receive SIF funds through an intermediary organization, and join other such organizations as a part of the intermediary's SIF portfolio. Subgrantees are required to conduct a program evaluation that is rigorous and builds upon the existing body of evidence for the program's intervention. Subgrantees are required to match intermediary funds one-to-one.

SIF Cohort: A class of intermediaries chosen in a particular funding year. Currently, there are three SIF cohorts, one in each of the following years: 2010, 2011, and 2012. The term also encompasses each intermediary's subgrantees: 2010 cohort of 11 intermediaries, 154 subgrantees; 2011 cohort of five intermediaries, 48 subgrantees; and 2012 cohort of four intermediaries, that have selected a number of subgrantees, but are currently still in the process to select more.

Scaling Up: A term that means increasing the size and reach of a sub grantee's intervention or program.

Matching funds: Federal requirements of dollar match by a non-federal entity in order to be eligible to receive grant funding.

Intervention: A program's activity or model that addresses a social or community challenge. The intervention is what is evaluated in the SIF.

Body of Evidence: A collection of science-based studies or research that support the effectiveness of a sub grantee's program or intervention.

Level of Evidence: A particular location along a continuum of programmatic evidence, ranging from anecdotal information (participant stories) to rigorous causal studies. For the purposes of SIF, the continuum is broken up into three distinct segments: preliminary, moderate, and strong. A subgrantee is assessed after being selected by an intermediary, or upon "entry" to SIF, and again after completing the grant cycle, or upon "exit."

Subgrantee Evaluation Plan (SEP): This term signifies the evaluation plan developed by subgrantees.

Unified Subgrantee Evaluation Plan (UniSEP): A single subgrantee evaluation plan that an intermediary applies to the evaluation of multiple subgrantees' programs in cases where multiple subgrantees are implementing an intervention across multiple sites.

Entry Level of Evidence: A designation of preliminary, moderate, or strong based on CNCS' assessment of the initial body of evidence behind a program's intervention.

Target Level of Evidence: A designation of preliminary, moderate, or strong based on CNCS' assessment of the outcomes of the executed evaluation that was designated by the SEP upon exit from the grant program.

Preliminary Level of Evidence: Interventions with all other types of outcome studies (e.g., pre-post test studies, studies monitoring outcomes throughout an intervention) were designated as "Preliminary." Interventions that were based on reasonable hypotheses supported by research findings (e.g., a body of literature that supports the use of the general type of intervention, but not the specific program as conducted by the grantee/subgrantee) were also designated as having preliminary evidence.

Moderate Level of Evidence: Interventions were designated as having "Moderate" evidence if they had at least one well-designed and well-implemented experimental or quasi-experimental study or multiple examples of correlational research with statistical controls supporting the effectiveness of the program.

Strong Level of Evidence: Interventions were designated as having "strong" evidence if they had (1) more than one well-designed and well-implemented experimental study or well-designed and well-implemented quasi-experimental study that supports the effectiveness of the practice, strategy, or program; or (2) one large, well-designed and well-implemented randomized controlled multisite trial that supports the effectiveness of the practice, strategy, or program.

Internal Validity: The extent to which a study can support causal conclusions by reducing systematic error or biases.

External Validity: The extent to which a study's results can be generalized to locations, contexts, or populations beyond those actually in the study itself.

Experimental Design: Experimental design studies using random control trials or RCTs assign program participants to two distinct groups (at random): the treatment group, which receives program services, and the control group, which does not. The control group is called the "counterfactual," representing the condition in which the program or intervention is absent. Random assignment ensures that the treatment and control groups are initially similar and do not differ on background characteristics or other factors. Random assignment, thus, creates an evaluation design where any observed differences between the two groups after the program intervention takes place can be attributed to the intervention with a high degree of confidence.

Random Assignment: A process that uses randomly generated numbers or other approaches to assign study units to groups in ways that are unaffected by the characteristics of the study units. With random assignment, any differences between the groups at pre-test can be attributed only to chance. The use, or lack of use, of this process differentiates experimental designs from non-experimental designs.

Quasi-Experimental Design: A design that forms a counterfactual group by means other than random assignment. This approach is used for conducting impact evaluations where observed changes in the treatment group are compared with a comparison group (as a counterfactual representing an absence of intervention) to assess and estimate the impact of the program on participants. However, groups formed in these designs typically differ for reasons other than chance, and these differences may influence the impact estimate. There are different types of approaches used in quasi-experimental designs such as

those using Propensity Score Matching (PSM), Regression Discontinuity, Interrupted Time Series (ITS) and others.

Propensity Score Matching: A statistical matching approach that is sometimes employed in quasi-experimental design studies for the purposes of developing a comparison group. This approach is based on a predicted probability of group membership (e.g., intervention vs. control) using measured characteristics of study units as predictors. The predicted probabilities are typically obtained from logistic regression.

Regression Discontinuity Design: This is a specific quasi-experimental design approach that is used for evaluating causal effects of interventions. Under this approach, assignment to a treatment is determined at least partly by the value of an observed covariate lying on either side of a fixed threshold. The intervention and control group are formed using a well-defined cutoff score. The group below the cutoff score receives the intervention and the group above does not, or vice versa. For example, if students are selected for a program based on test scores, those just above the score and just below the score are expected to be very similar except for participation in the program, and can be compared with each other to determine the program's impact.

Interrupted Time Series: This is a specific quasi-experimental design approach that is used for evaluating causal effects of interventions. Under this approach, multiple observations are obtained prior to the intervention to establish a baseline. Multiple observations are also obtained after the intervention. Effects are demonstrated when the observations after the intervention deviate from expectations derived from baseline projections.

Non-Experimental Design: The term is a catch-all category that refers to a range of research and evaluation studies that do not fall under the experimental or quasi-experimental research designs. They include process and outcomes evaluations, case studies, cost effectiveness or cost benefit analysis, feasibility studies, rapid assessments, situational and contribution analysis, developmental evaluation, strategic learning, systems change studies, and others.

Appendix

Details of the Multiple Regression Analysis

Confidence in findings and conclusions about the differences between the budgets for different types of studies and for different levels of evidence might be limited if: a) the differences between the study budgets was no bigger than could be expected due to chance; b) estimates of differences were affected by the small number of studies; c) were overly-determined by the characteristics of a few outlying studies; and d) did not rule out as many alternative explanations as possible. In order to assess the validity of the conclusions, a series of statistical analyses were conducted. The statistical analyses were based on Bayesian multiple regression where small sample size does not necessarily limit the extent of findings' plausibility. Additionally, the distribution of budget numbers was adjusted to draw in outliers.[16] Finally, the study characteristics, such as the impact evaluation's sample size and the program's focus area, were controlled for so that the effect of a program's study design could be isolated from other factors.

More specifically, a multiple regression was calculated using each program's budget amount as the dependent variable. The natural log of the budget amounts was taken prior to the regression, to pull in outliers and make the variable more normally-distributed. All program factors were entered into the regression simultaneously, and so the estimates of a single factor's relation to a program's budget is a net of (or, controls for) the effect of other factors in the regression. Not all data were available for every program, and there was a relatively small number of evaluations included in the analysis (N=70). To address these two issues, the regression estimation used Full Information Maximum Likelihood (FIML) procedures to include programs with some missing data, and Bayesian estimation, which does not make assumptions about sample size and estimate accuracy. Using this procedure, any correlation between independent variables is modeled, so that non-multicollinearity assumptions of the regression are not likely to have been violated. Results from the analysis conducted in SPSS and Mplus are presented, for comparison purposes.

The tables below present the unstandardized (B) and standardized (Beta) estimates, with standardization on the dependent variables in the regression based on their distribution's mean and standard deviation. Thus, the standardized beta estimates are effect sizes, and their magnitude can be directly compared across regression equations because their unit has been converted to standard deviations of the dependent variable. The B estimates are interpreted as the size of the unit change in the log of outcome measure associated with a one unit increase in the independent variable, net of other variables in the model. For example, a B of 0.5 when 'preliminary level of evidence' was the independent variable and log of total budget was the dependent variable, would indicate that, net of other factors, a study targeting preliminary level of evidence was associated with a 0.5 increase in the log of total dollars spent. The interpretation of the Beta estimate is the change in *standard deviations* of the dependent variable associated with a change in X. Statistical significance of the estimates in the model is indicated using p-values. The point of significance used is 0.05.

[16] Logarithmic transformation of budget figures was used for this purpose.

Table 1: SPSS Regression Analysis

	% Total Budget Used for Evaluation: B	% Total Budget Used for Evaluation: Beta	% Total Budget Used for Evaluation: Sig	Total Evaluation Budget: B	Total Evaluation Budget: Beta	Total Evaluation Budget: Sig	Total Budget: B	Total Budget: Beta	Total Budget: Sig
Sample Size	.072	.100		.116	.074		.044	.033	
Focus Area (Youth Outcomes is Reference)									
Economic Opportunities	-.157	-.079		1.122	.261	**	1.279	.350	**
Healthy Futures	.399	.201		-.370	-.086		-.769	-.210	*
Level of Evidence (Moderate is Reference)									
Preliminary Plus	.152	.083		-.411	-.104		-.563	-.167	
Preliminary	-.292	-.174		-1.255	-.345	*	-.963	-.312	**
Strong	.839	.473	**	1.765	.459	*	.925	.283	**

	% Total Budget Used for Evaluation: B	% Total Budget Used for Evaluation: Beta	% Total Budget Used for Evaluation: Sig	Total Evaluation Budget: B	Total Evaluation Budget: Beta	Total Evaluation Budget: Sig	Total Budget: B	Total Budget: Beta	Total Budget: Sig
Sample Size	-.014	-.020		-.065	-.042		-.051	-.039	
Focus Area (Youth Outcomes is Reference)									
Economic Opportunities	-.084	-.042		1.194	.277	**	1.278	.349	**
Healthy Futures	.329	.166		-.805	-.187	+	-1.134	-.310	**
Design Type (RCT is Reference)									
Feasibility	-.756	-.343	*	-2.501	-.523	**	-1.744	-.429	**
Pre-Post	-.609	-.258	+	-.918	-.179		-.308	-.071	
QED	-.690	-.421	**	-1.802	-.507	**	-1.112	-.368	**
QED - PSM	-.744	-.432	**	-1.467	-.393	**	-.723	-.228	+

**p<.01 * p< .05; + p< .10

Table 2: Regression Analysis using Bayes

USING BAYES, FOR WHICH SMALL SAMPLE SIZE IS NOT AN ASSUMPTION ISSUE, AND WHICH INCLUDES CASES WITH MISSING DATA (N=70)

	% Total Budget Used for Evaluation: B	% Total Budget Used for Evaluation: Beta	% Total Budget Used for Evaluation: Sig	Total Evaluation Budget: B	Total Evaluation Budget: Beta	Total Evaluation Budget: Sig	Total Budget: B	Total Budget: Beta	Total Budget: Sig
Sample Size	.052	.070		.123	.078		.056	.042	
Focus Area (Youth Outcomes is Reference)									
Economic Opportunities	-0.045	-.020		1.149	0.249	*	1.242	.317	*
Healthy Futures	.121	.057		-.811	-.190	*	-.907	-.249	*
Level of Evidence (Moderate is Reference)									
Preliminary Plus	.132	.063		-.361	-.081		-.479	-.127	
Preliminary	-0.206	-.126		-.837	-.231	*	-.600	-.195	+
Strong	.575	.310	*	1.701	.432	*	1.097	.328	*

	% Total Budget Used for Evaluation: B	% Total Budget Used for Evaluation: Beta	% Total Budget Used for Evaluation: Sig	Total Evaluation Budget: B	Total Evaluation Budget: Beta	Total Evaluation Budget: Sig	Total Budget: B	Total Budget: Beta	Total Budget: Sig
Sample Size	-.013	-.016		-0.06	-.037		-0.058	-.043	
Focus Area (Youth Outcomes is Reference)									
Economic Opportunities	0.014	.007		1.253	.268	*	1.28	.323	*
Healthy Futures	.102	.054		-1.031	-.237	*	-1.139	-.309	*
Design Type (RCT is reference)									
Feasibility	-.697	-.290	*	-2.205	-.427	*	-1.609	-.367	*
Pre-Post	-.666	-.285	*	-1.311	-.255	*	-0.653	-.150	
QED	-.569	-.329	*	-1.765	-.470	*	-1.199	-.377	*
QED - PSM	-.561	-.285	*	-1.454	-.355	*	-0.923	-.266	*

* p< .05; + .10

Study Type

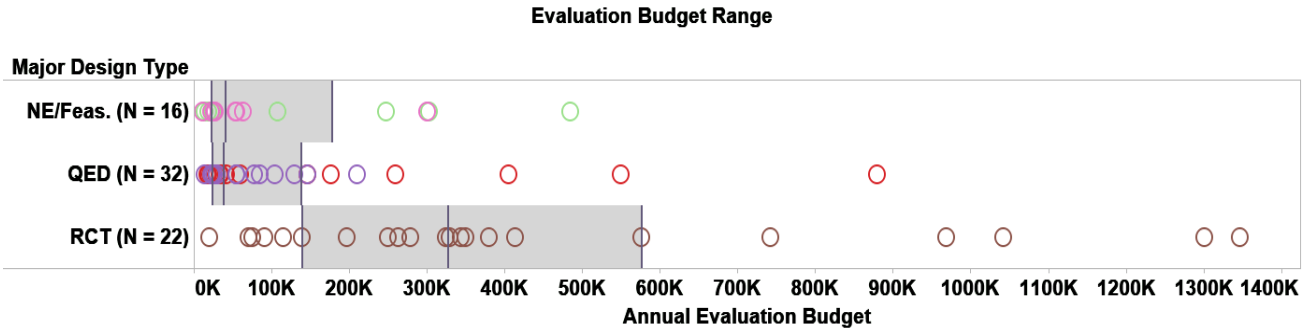
Results of analyses also showed that a program's total evaluation budget was higher –statistically significantly higher – when a study was an RCT compared to other study types. A program's total overall budget was significantly higher for RCTs compared to Feasibility, QED, and propensity score matched QED studies. The proportion of a program's total budget spent on evaluation was significantly higher when a study was an RCT, compared to any other study design type. These differences were separate from the effect that a study's sample size and program focus area had on a program's budgeting. This means that RCTs used in any of the SIF focus areas such as youth development, economic opportunity, and healthy futures were substantially more costly compared to other designs. The same was true for evaluations that targeted a strong level of evidence. Moreover, the size of the sample used in the studies did not change these conclusions.

Thus, differences in budgeted amount can be considered statistically meaningful, and confidence in assertions made are not limited due to the small number of studies included in the analysis, or by the small number of outliers in the array of programs considered. The analyses largely ruled out the possibility that the relation between study design and budgetary differences are actually due to the sample size or focus area of the studies. Finally, because there is substantial overlap between the type of study design and the level of evidence, with programs targeting strong level of evidence using RCTs most often, it is difficult to statistically examine level of evidence and design type simultaneously. However, analyses conducted showed that studies targeting a "strong" level of evidence have both larger evaluation budgets, and use a larger portion of the program budget when both sample size and program focus area are held constant.

Other Results

The figures below show the range of budget costs by major and specific design types. The shaded areas showing the middle quartiles of the range indicate a clear overlap between non-experimental and quasi-experimental budgets, while the RCT budgets distribution overlaps very little with the other two types.

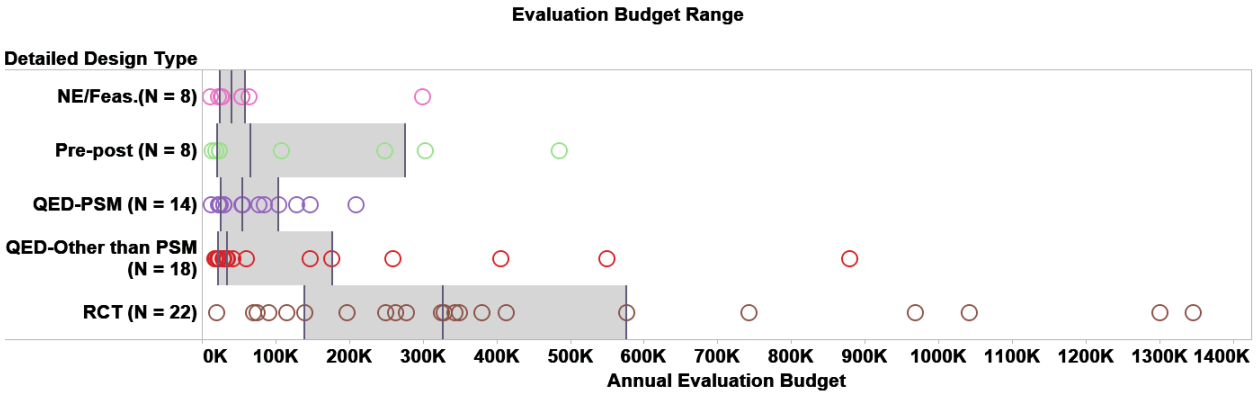
Figure 7: Evaluation Budget Range by Major Design Type



The average and median evaluation budgets for non-experimental design studies were $111,473 and $40,700 per year, with a few studies pulling up the average for these types of designs. The QED studies were clustered much more closely together at the lower end. The average and median cost of these studies was$118,083 and $38,434 and the above figure shows that costs for these studies were affected by a number of more expensive evaluations. The experimental design studies have a wider spread, from a minimum of $20,000 to a maximum of $1,346,342 per year.

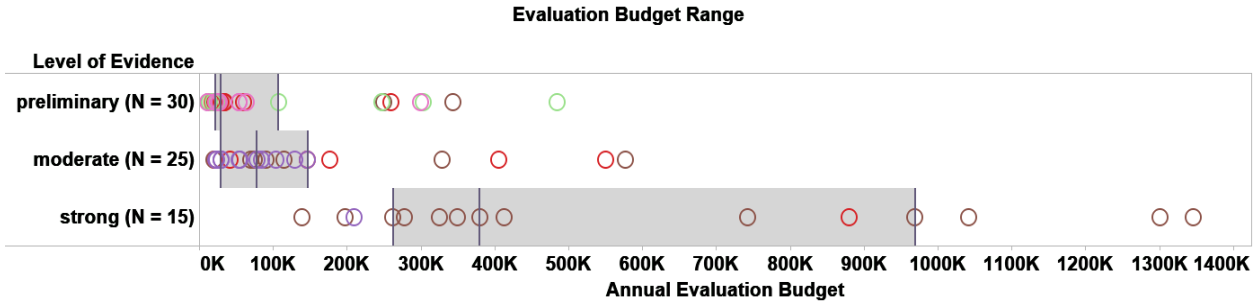
A more detailed breakout of design types shows that RCT costs overlap slightly with both pre- and post-test designs and QEDs that did not use propensity score matching. A number of high-cost QED studies without Propensity Score Matching (PSM) pushed up the average cost of studies in this category relative to QEDs that use PSM.

Figure 8: Evaluation Budget Range by Specific Design Type



The Figure 9 illustrates that there was a distinct break between the budgets for programs targeting strong evidence and those targeting preliminary or moderate levels of evidence.

Figure 9: Evaluation Budget Range by Levels of Evidence



Although the average budget for evaluations with preliminary level of evidence was $98,123, the median was only $29,250. A small number of very costly preliminary studies raised the average for this level of evidence.

Considering the maximum cost of evaluation per year under each type of evaluation design suggests that all design types have the potential to be expensive. When the available budgets are disaggregated by level of evidence and design type, however, more variability can be detected.[17] For example, some high quality, large scale rigorous non-experimental approaches designed to address issues other than causal

[17] In order to keep the grantees/subgrantees' identities confidential, breakdowns by level of evidence and design type are not presented here. However, it should be noted that studies that target preliminary level of evidence include non-experimental design studies as well as quasi-experimental design studies and RCTs. Studies that target moderate and strong levels of evidence include QEDs of different types as well as RCTs.

impacts can be as costly as designs targeting moderate levels of evidence. Finally, feasibility studies were pulled out from across different design categories in order to gauge what they cost. Average cost of feasibility studies across the portfolio (n=12) was $113,788 per year with an evaluation to budget ratio of 18 percent. In cases where the feasibility study was in preparation to conduct QEDs or RCTs that target a strong level of evidence, these studies tended to get very expensive. In addition, a couple of feasibility studies that were combined with pre and post designs were also very costly.

Detailed SIF Evaluation Budgets

Tables 3, 4 and 5 below provides detailed budget information from the 70 SIF studies underway with the minimum, maximum, average and median levels of annual SIF program funding, evaluation funding, and ratio of annual evaluation budget to program budget. This information is also grouped by study type and levels of evidence targeted. The data clearly demonstrate the higher average cost of RCT studies and those seeking to target a strong level of evidence. These costs are also reflected in the higher evaluation to program budget ratios for RCTs and those requiring strong evidence.

Table 3: SIF Average Annual Program Budget by Design Type and Target Levels of Evidence

Evaluation Study Design	Average	Median	Minimum	Maximum	N
Non-Experimental (NE)	$828,655	$420,000	$100,000	$2,821,953	16
Implementation/ Feasibility	$559,137	$420,000	$100,000	$2,000,000	*8*
Pre-post	$1,098,174	$469,286	$100,000	$2,821,953	*8*
Quasi-Experimental Design (QED)	$820,343	$362,008	$100,000	$4,748,313	32
QED-Other than PSM	$878,891	$402,008	$100,000	$3,500,000	*18*
QED-PSM	$745,068	$324,888	$100,000	$4,748,313	*14*
Experimental Design (RCT)	$1,718,908	$1,350,000	$100,000	$5,460,618	22
Overall	$1,104,649	$593,309	$100,000	$5,460,618	70
Target Level of Evidence					
Preliminary	$673,239	$326,896	$100,000	$2,821,953	30
Moderate	$992,775	$513,000	$100,000	$4,748,313	25
Strong	$2,153,927	$2,000,000	$1,020,751	$5,460,618	15

Table 4: SIF Average Annual Evaluation Budget by Design Type and Target Levels of Evidence

Evaluation Study Design	Average	Median	Minimum	Maximum	N
Non-Experimental (NE)	$111,473	$40,700	$12,000	$484,790	16
Implementation/ Feasibility	$69,932	$40,700	$12,000	$300,233	*8*
Pre-post	$153,014	$65,871	$14,167	$484,790	*8*
Quasi-Experimental Design (QED)	$118,083	$38,434	$13,333	$879,667	32
QED-Other than PSM	$154,005	$34,135	$17,625	$879,667	*18*
QED-PSM	$71,898	$54,700	$13,333	$209,763	*14*
Experimental Design (RCT)	$437,110	$327,251	$20,000	$1,346,342	22
Overall	$216,838	$81,471	$12,000	$1,346,342	70
Target Level of Evidence					
Preliminary	$98,123	$29,250	$12,000	$484,790	30
Moderate	$135,976	$77,943	$20,000	$576,667	25
Strong	$589,037	$379,800	$139,189	$1,346,342	15

Table 5: SIF Average Annual Evaluation Budget to Program Budget Ratio by Design Type and Target Levels of Evidence

Evaluation Study Design	Average	Median	Minimum	Maximum	N
Non-Experimental (NE)	16%	13%	3%	45%	16
Implementation/ Feasibility	16%	14%	3%	45%	*8*
Pre-post	16%	11%	4%	43%	*8*
Quasi-Experimental Design (QED)	16%	13%	3%	56%	32
QED-Other than PSM	15%	14%	4%	44%	*18*
QED-PSM	17%	13%	3%	56%	*14*
Experimental Design (RCT)	25%	22%	9%	83%	22
Overall	19%	15%	3%	83%	70
Target Level of Evidence					
Preliminary	15%	14%	3%	45%	30
Moderate	17%	15%	3%	56%	25
Strong	28%	24%	11%	83%	15