I. Introduction: Program Background and Problem Definition, Overview of Prior Research, Overview of Study, and Connection of this Study to Future Research

To contextualize the evaluation of the program, it is important to provide a strong description of the program and its origins. The relationship of the program to the problem it is designed to address is also important for understanding the overall evaluation design.

The Program Background and Problem Definition section describes the problem or issue that the program addresses. Use research related to the program, and the problem or issue it addresses, to provide context. The description does not have to be exhaustive, but it should provide sufficient background for understanding why and how the program came to be developed.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Program Background and Problem Definition

- ☑ The policy issue or problem the program is designed to address and why it should be addressed is succinctly described.
- ☑ The issue or problem is framed based on a review of research.
- ☑ The program model is described briefly in the introduction, and includes key information, such as who participants will be, the intervention level, and key outcomes.

Example: Introduction: Program Background and Problem Definition, Overview of Prior Research, Overview of Study, and Connection of this Study to Future Research
Excerpted from the SEP by the SEED program, a subgrantee of the Edna McConnell Clark Foundation.

Sweeping changes in the U.S. economy and labor market over the past three decades have dramatically reduced the availability of well-paying jobs for workers without post-secondary education. And yet, one-fourth of high school freshmen nationwide do not graduate in four years, and many who do complete school are not ready to perform college level work. These trends are particularly pronounced in urban areas, and among students from low-income and underserved families.

The policy problem that the program addresses is presented in the first two paragraphs.

With this in mind, policy makers, practitioners and researchers have developed and promoted a variety of approaches to improving students' high school options. One approach operates outside of the existing public education system by creating new

This paragraph provides background from previous research to situate the problem that the SEED program tackles.

institutions, such as charter schools, that offer an alternative to and thus compete with public high schools (Gleason, et al., 2010). A second approach operates within the existing public education system by attempting comprehensive reform of failing high schools based on reform models such as Career Academies (Kemple, 2008), Talent Development Schools (Kemple, Herlihy, & Smith, 2005), Project GRAD (Snipes, Holton, & Doolittle, 2006), and others. A key element of many of these reform models is the creation of small learning communities within existing schools. A third approach, which also works inside the existing public education system, is to close failing high schools (which in urban areas are often quite large) and replace them with new public high schools that are small in size and open to all students.

The SEED Foundation is a national nonprofit that partners with urban communities to provide innovative educational opportunities that prepare underserved students for success in college and beyond. SEED believes in college attainment and completion as a solution to urban poverty. The SEED Foundation is the only organization in the U.S. that has successfully started and sustained urban, public, and college-prep boarding schools. The SEED School of Washington, D.C. (SEED DC), which opened in 1998, and The SEED School of Maryland, which opened in 2008, currently serve 582 students and are growing to serve 730 students in grades 6-12. In addition, SEED plans to open at least two new schools with support from the Social Innovation Fund (SIF), likely in 2012 and 2013.

The subsequent paragraphs introduce the SEED program and provide information about the program theory and program components.

SEED schools attempt to take the best of the options detailed above and combine them together into an intensive and holistic intervention. SEED's boarding school model is predicated on the assumption that, for certain disadvantaged students who face overwhelming barriers to success at home and in the community, the typical school reforms and enhancements (for example, after school programs, extended school hours) will not be sufficient (as outlined by SEED in the figure below). Rather, SEED believes that, for these students, achieving success in high school and beyond requires a fully-integrated academic and boarding program that also provides scheduled study time, constant access to positive role models, and life skills training.

In 2004, Perry Bacon of TIME magazine profiled students in the SEED School of Washington, D.C. and described their daily life as follows:

"The dorms are divided into "houses" of 10 to 14 students, which are named after universities, reflecting the school's emphasis on college preparation. Two students share each small room, which contains little more than twin beds, two desks and, for upperclassmen, desktop computers. The schedule is purposefully intense. Each morning, after waking up at 5:45 a.m., the kids make their bed, get dressed in their uniform of khaki pants and white polo or Oxford shirt, then line up single file to go to the cafeteria for breakfast. Classes begin at 8 a.m. and last until 4 p.m. The late-afternoon hours are filled with extracurricular activities that range from choir to flag football. After dinner, the students go back to their dorms for an hour-long study hall before a half-hour of "quiet time," then go to bed. There are few behavior problems. "They don't have a lot of time to get into that stuff," says Roz Fuller, the associate boarding director."

SEED has been the subject of several qualitative studies and one impact study, conducted by Vilsa Curto and Roland Fryer, Jr. of Harvard University.[1] Curto and Fryer's study built on the admissions lottery for SEED DC to compare state standardized test scores for students who "won" the lottery and were offered admission to SEED for the 2006 and 2007 school years, with students who "lost" the lotteries in those years. The analysis, based on a small sample and a short follow-up period, found that SEED increased reading scores by 0.198 standard deviations and math scores by 0.230 standard deviations, per year of attendance. While both boys and girls experienced statistically significant positive effects of SEED enrollment, results were much stronger for girls than for boys.[2]

The research activities described below will build on the Curto and Fryer analysis by following a larger number of students for a longer period, obtaining data on a broader range of outcomes from school records and a student survey, and conducting an implementation study to more fully describe the SEED intervention. The design discussed below is limited by the size of the SEED schools, our ability to follow only one cohort through high school to four-year graduation, and the relatively recent establishment of the Maryland school. Thus, while the analysis will include all students who have gone through the SEED DC lotteries since 2006, and follow them for as long as possible within the SIF timeframe, the sample sizes for some parts of the analysis (those that focus on later high school outcomes) are relatively

[1] Curto and Fryer, 2011.

[2] According to Curto and Fryer, the 2SLS estimates for females (including controls for baseline scores and demographic characteristics) are 0.382 in reading (-0.138 for males) and 0.265 in math (0.037 for males). The difference between males and females is significant for reading, but they cannot reject the null hypothesis that the effects are the same for math.

small. If possible, it will be important to continue the analysis beyond the three-year SIF period in order to obtain data on longer-term outcomes for a larger number of students.

Solution Spectrum for Improving Educational Outcomes

Enrichment/ Support Programs ↔ **Year-Round Intervention**

Type:	**Strengthened Day School**	**After-School Tutoring**	**Extended-Hours School**	**Boarding School**	**Year-Round Immersion**
Target Population Children:	• High-achieving children needing more challenging curriculum	• Academically motivated children needing specific, supporting aid	• Motivated children needing more extensive support and aid	• Academic underperformers who, if given extensive support around the clock, can go from at-risk to college-bound	• Children with serious behavioral and/or emotional problems; resistant to education
Target Population Environment:	• Stable and supporting home and background	• Supportive home	• Supportive individuals	• May lack any supportive individuals • Could benefit from a residential setting	• May lack any supportive individuals • Need change in residential setting

SEED Foundation

II. Program Theory and Logic Model

The SEP should include an overview of your program's theory and a logic model that guides the evaluation and complies with all of the guidance in this document. A description of program theory will provide the reader with a better understanding of how your program is expected to achieve its targeted outcomes and impacts.

A description of program theory, coupled with an informative logic model, frames the evaluation design. Understanding how a program is designed to work in theory is an important precursor upon which all subsequent evaluation activities fundamentally rest. A program logic model usually includes both a graphic display and a narrative description of the resources/inputs, the program activities that constitute the intervention, and desired participant outcomes/results.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Logic Model

- ☑ Both a narrative and a graphic display that follows the chain of reasoning are included.
- ☑ The logic model concepts, including all outcomes to be measured, are clearly defined.
- ☑ How the resources and activities lead to the outcomes is described.
- ☑ Only aspects directly related to the theory of change are included.
- ☑ Existing literature to describe the theoretical or research basis that the model draws upon is included.
- ☑ Existing literature to support the connections between the activities and outcomes is used.

Example: Program Theory and Logic Model

Excerpted from the SEP by Gateway to College, a subgrantee of the Edna McConnell Clark Foundation.

As discussed [in the introduction], the goal of the Gateway to College program is to reconnect students who have dropped out of high school or are at risk of dropping out of high school by putting them on a path to earn their high school diploma while also earning credits towards an Associate's degree or certificate. This section describes the various players, activities, and key outputs that feed into attaining the student-level outcomes being measured in this evaluation. This section also describes and provides support for the assumptions behind the logic model (Figure 2.1) as available through past research and a review of the literature.

Inputs

As demonstrated in Figure 2.1, the Gateway to College logic model has three main inputs, which represent the organizational infrastructure required to successfully implement the Gateway to

The individual component parts of the logic model are each described in detail in these paragraphs.

College model: (1) the Gateway to College National Network (GtCNN); (2) community colleges; and (3) K-12 school districts.

The first input is the Gateway to College National Network, which is based in Portland, Oregon and is charged with overseeing the implementation of the Gateway to College program, including providing training and support for instructors and staff, as well as developing and sustaining partnerships between K-12 school districts and community colleges. The second input is community colleges, which serve as the institutional hosts for all Gateway to College classes and are also where the Gateway to College staff are located. Participating students have full access to college courses, facilities, and support services. The community college partnership also provides flexible class times for non-traditional students' schedules. Currently, there are 29 colleges in 16 states in the Gateway to College network. The third input is the K-12 school districts, which provide funding for tuition and books, are a central source of student data, and provide referrals to the Gateway to College program. Together, the Gateway to College National Network, community colleges, and K-12 school districts provide the infrastructure for the implementation of the Gateway to College program and related activities.

Activities, Outputs, & Outcomes

The five core program activities of Gateway to College as outlined in the logic model are a holistic and interdependent network of services that include both student-level activities and support for Gateway to College staff. The student-level activities and services include: (1) the Foundation Experience; (2) transitioning to general college classes; and (3) support from Resource Specialists. The activities associated with program staff include: (4) Implementation of student instruction and support based on the Gateway to College "Principles of Teaching & Learning"; and (5) Support from the Gateway to College National Network through ongoing training, technical assistance, and professional development.

All students begin with the Foundation experience where a learning community of 20-25 students takes developmental reading, writing, math and college preparatory courses taught primarily by Gateway to College faculty. The Foundation Experience is intended to create a shared experience among students, as well as a network of peer support. This network of peer support is designed to strengthen over time and is expected to carry over to the second activity outlined in the theory of change--the transition to general college classes. The theory of change posits that the learning community and network of peer support created by these two activities is a key factor that helps to facilitate improved retention and academic outcomes (e.g. more credits earned, increased number of college courses completed.). Tinto (1997) [3], who conducted a study on the impact of learning communities in a community college setting, argues that "learning communities promote persistence by facilitating the creation of supportive peer groups among students, encouraging shared learning, and giving students the opportunity to actively participate in knowledge selection" (Bailey, 2005).[4]

In addition, a recent report by MDRC on the impact of learning communities on community college students found that those who participated in a learning community attempted and passed their developmental math classes at higher rates than those students who were not a part of the learning community (Weissman, 2011)[5]. While these results diminished in subsequent semesters, students in the learning communities still reported higher levels of engagement and greater satisfaction with their college experience, and were more likely to take and pass an English assessment test required for graduation or transfer more than a year later.[6] In addition, this study also found that linked classes can have an impact on students' achievement during the program semester – a feature that is inherent in the Gateway to College program design.

Existing literature and research are cited throughout the section to illustrate the basis for the model as well as the connections between component parts.

Moreover, as activities in the logic model, the Foundation Experience and the transition to general college classes also lead to two key outputs of the Gateway to College program: the dual-enrollment component and the partnership between K-12 and postsecondary institutions. While dual-enrollment programs have been in existence for many years, they were once reserved for high achieving students, and have only recently become increasingly available for moderate to lower achieving students such as those targeted by Gateway to College (Bailey, 2003)[7]. Research by the American Association of State College and Universities (2002)[8] suggests that dual enrollment may reduce high school dropout rates, increase student aspirations, and decrease the amount of remediation needed by incoming college students. Past research

[3] Tinto, V. (1997). Classrooms and communities: Exploring the educational character of student persistence. *Journal of Higher Education, 68*(6), 599-623.

[4] Bailey, T., Alfonso, M. (2005). "Paths to Persistence: An Analysis of Research on Program Effectiveness at Community Colleges." *Community College Research Center, Teachers College, Columbia University.*

[5] Weissman, E. (2011). "Learning Communities for Students in Developmental Math. Impact Studies at Kingsborough and Houston Community Colleges." MDRC.

[6] Ibid.

[7] Bailey, T., & Karp, M. M. (2003). *Promoting college access and success: A review of dual credit and other high school/college transition programs.* Washington, DC: U.S. Department of Education.

[8] American Association of State Colleges and Universities. (2002). The open door: Assessing the promise and problems of dual enrollment. *AASCU State Policy Briefing, 1*(1), 1–10.

also suggests that one reason dual enrollment programs that serve at-risk students have the potential to reduce the high school dropout rate is because it provides students the opportunity to be academically challenged and partake in more engaging coursework (Lords, 2000)[9] – an opportunity that may not have been present in some traditional K-12 settings, and one that is embedded in Gateway to College instruction and support.

In addition to increasing high school completion rates, this program theory of change hypothesizes that Gateway to College also has an impact on longer-term outcomes such as postsecondary access and success. While the impact on these outcomes will not be tested during the three-year timeframe of this evaluation, the literature on programs with dual enrollment components leading to postsecondary gains is promising and should be noted, particularly if the timeframe of this evaluation can be extended and additional follow-up is conducted. In a study by Karp (2007)[10] using data obtained from the State of Florida and the City University of New York (CUNY), it was found that dual enrollment was positively related to enrollment in college, and also increased the likelihood of enrolling in a four-year institution.

Activities 3, 4, and 5, which are: (3) support provided by Resource Specialists; (4) the use of the "Principles of Teaching & Learning" as the guiding framework for how to deliver student instruction; (5) and ongoing training, technical assistance, and professional development support provided by the Gateway to College National Network, together represent a host of wrap around support services that make up two key outputs: (1) Innovative Teaching & Learning; and (2) Intentional Collaboration.

The Innovative Teaching & Learning output encourages Gateway to College instructors and staff to implement innovative pedagogical techniques guided by the Gateway to College "Principles of Teaching & Learning." Drawing from literature in K-12 education, numerous studies have revealed the tremendous impact teachers/instructors have on student achievement. For example, in a study by Marzano (2003)[11], students of teachers who were characterized as "most effective" posted gains of 53 percentage points over the course of one academic year, as compared to 14 percentage points for students taught by teachers

[9] Lords, E. (2000). "New efforts at community colleges focus on underachieving teens." *The Chronicle of Higher Education.* June 30, 2000, p. A45.

[10] Karp, Melinda Mechur, Juan Carlos Calcagno, Katherine L. Hughes, Dong Wook Jeong & Thomas R. Bailey. 2007. *The Postsecondary Achievement of Participants in Dual Enrollment: An Analysis of Student Outcomes in Two States.* St. Paul, MN: National Research Center for Career and Technical Education, University of Minnesota.

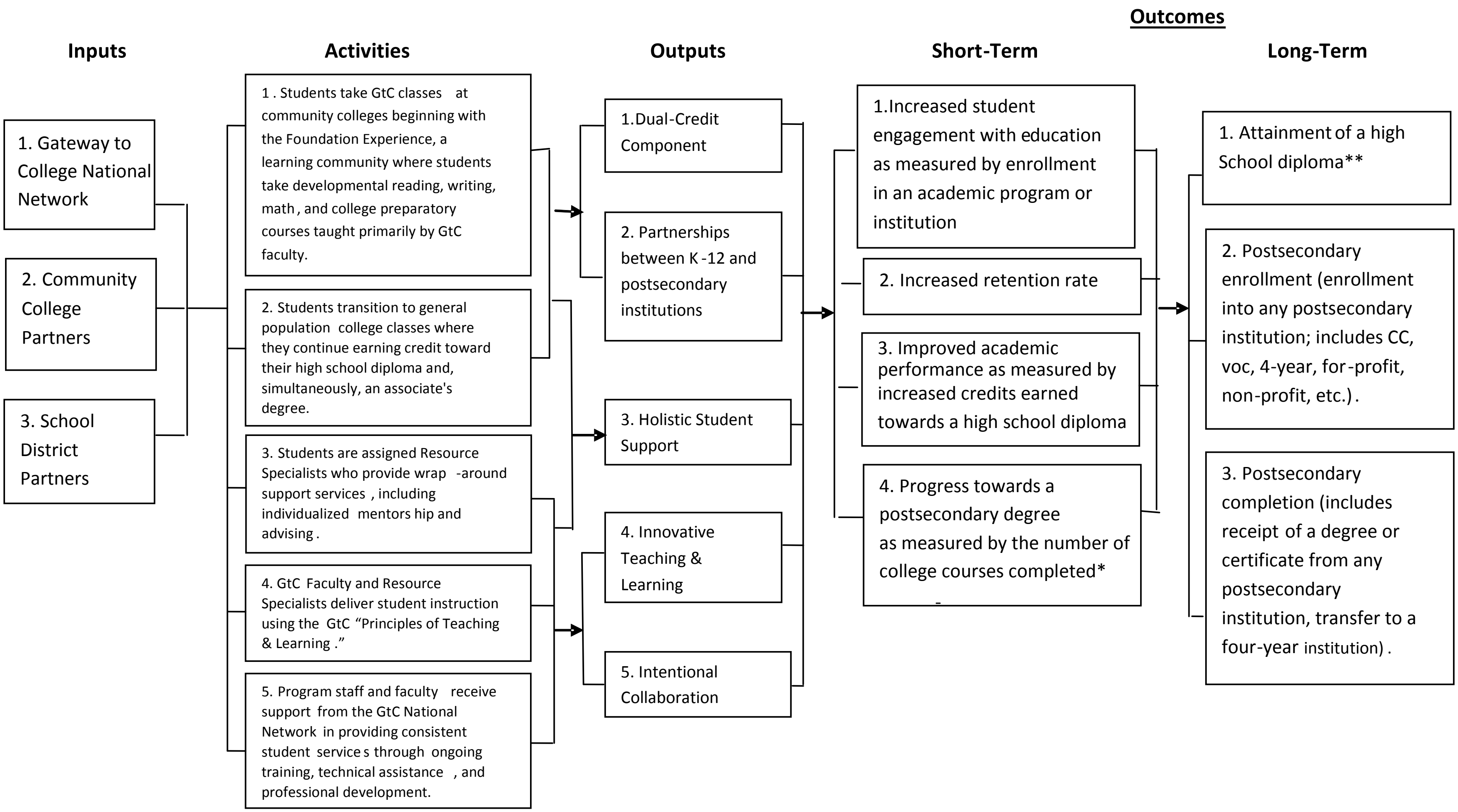
[11] Marzano, R. J. (2003). What works in schools: Translating research into action. Alexandria, VA: Association for Supervision and Curriculum Development.

that were “least effective.” While the context in the Gateway to College program is different from that of K-12 school settings, the implications of having a strong cadre of instructors is clear -- particularly for the at-risk student population that Gateway to College serves. As such, the Gateway to College theory of change posits that the high level of training and ongoing support that the Gateway to College National Network provides its staff, as well as the network of professional collaboration across the network plays a critical role in impacting the academic achievement of students.

As discussed above, the short-term outcomes being measured in this evaluation are promising intermediate indicators of the longer-term outcomes such as attainment of a high school diploma and postsecondary access and success. Taken together, we expect that the holistic and interdependent network of services coupled with high quality instruction, dual-enrollment framework, and strong culture of collaboration and support (for both students and instructors) provides a promising framework for positively impacting the student outcomes being measured in this evaluation.

The logic model graphic and narrative section focus on the essential components of the program and the processes related to it.

Figure 2.1: Gateway to College Logic Model



Outcomes

Inputs

1. Gateway to College National Network

2. Community College Partners

3. School District Partners

Activities

1 . Students take GtC classes at community colleges beginning with the Foundation Experience, a learning community where students take developmental reading, writing, math , and college preparatory courses taught primarily by GtC faculty.

2. Students transition to general population college classes where they continue earning credit toward their high school diploma and, simultaneously, an associate's degree.

3. Students are assigned Resource Specialists who provide wrap -around support services , including individualized mentors hip and advising .

4. GtC Faculty and Resource Specialists deliver student instruction using the GtC "Principles of Teaching & Learning ."

5. Program staff and faculty receive support from the GtC National Network in providing consistent student service s through ongoing training, technical assistance , and professional development.

Outputs

1.Dual-Credit Component

2. Partnerships between K -12 and postsecondary institutions

3. Holistic Student Support

4. Innovative Teaching & Learning

5. Intentional Collaboration

Short-Term

1.Increased student engagement with education as measured by enrollment in an academic program or institution

2. Increased retention rate

3. Improved academic performance as measured by increased credits earned towards a high school diploma

4. Progress towards a postsecondary degree as measured by the number of college courses completed*

Long-Term

1. Attainment of a high School diploma**

2. Postsecondary enrollment (enrollment into any postsecondary institution; includes CC, voc, 4-year, for-profit, non-profit, etc.) .

3. Postsecondary completion (includes receipt of a degree or certificate from any postsecondary institution, transfer to a four-year institution) .

Assumptions
1. A learning community among students and Gateway to College staff and instructors is created and maintained as a result of all the program activities.

External Factors
1. K-12 and postsecondary partnerships are maintained.
2. Funding and policies at K-12 institutions and at the state level are conducive to continuing program implementation.

*Due to the unfeasibility and costs associated with obtaining student-level postsecondary progress data (e.g. through credits earned) directly from each postsecondary institution that treatment and control students are enrolled in, we may incorporate additional proxy measures of postsecondary progress towards a degree in addition to the number of college courses completed.

**While we expect to be able to fully measure impacts on the attainment of a high school diploma in the long-term with an extended evaluation timeframe, it is possible that we may also see impacts on high school diploma receipt in the short-term.

III. Research Questions and Contribution of the Study

The SEP should include both the questions that the evaluation hopes to answer and a description of how answering these questions will contribute to a greater understanding of programs, outcomes, and/or policies. In this section, questions related to both the impact, and if appropriate, implementation evaluations should be included.

Impact evaluations pose questions about the outcome of the program for beneficiaries/participants and on the impact of program services or participation more generally. Impact research questions may be confirmatory or exploratory in nature. Confirmatory questions represent the main impact questions for the primary outcome that the study can address with a known level of statistical precision (based on statistical power). Exploratory questions are those that are posed during the design phase, and implied by, or stated in the logic model, but cannot be answered with adequate statistical power.

Implementation evaluations pose questions related to the process of developing, running, or expanding a program, and potentially also about participants' experiences of program participation. These questions should focus on the process by which the program operates, rather than on the outcomes among beneficiaries.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Impact Questions

- ☑ Program impact questions that the study will address are clearly stated.
- ☑ The questions are consistent with the logic model's stated outcomes.
- ☑ If there are multiple questions related to the same outcome, confirmatory and exploratory questions are clearly defined.
- ☑ The confirmatory research question(s) represent(s) the main impact question(s) for the primary outcome that the study can address with a known level of statistical precision (based on statistical power).
- ☑ Questions address programmatic activity in the program participant group and, when feasible, the comparison or control group.
- ☑ Questions are phrased such that groups can be compared to understand net effect of intervention and differences.
- ☑ Impact questions do not include implementation questions.

Example: Research Questions
Excerpted from the SEP by the Latin America Youth Center, a subgrantee of Venture Philanthropy Partners.

The external evaluation of the Promotores Pathway will seek to answer four primary research questions.

Program Impact Questions
The first question will focus on program impact and is the **confirmatory research question**:

1) Is the Promotores Pathway leading to improved outcomes for participants in the areas of academics, employment, and healthy behaviors when compared to LAYC participants not matched with a Promotor?

The second question focuses on subsets of participants and outcomes and is the **exploratory research question**:

2) Are there differences in the outcomes achieved among subgroups of Promotores participants? For example, are Latino participants performing differently than African American participants, or are males performing differently than females?

The research questions are divided up between impact and implementation evaluations. Additionally, the confirmatory and exploratory questions are appropriately labeled.

Program Implementation Questions
The final two research questions deal with program implementation and potential replication in other locations:
3) What are the key components of Promotores Pathway and how does it operate „on the ground"?

4) Is there a core group of components that make up the Promotores Pathway that could be replicated in other multi-service direct-service nonprofit organizations?

Contribution of the Study

Using the overview of prior research as context, the SEP should clearly state what contribution the proposed study will make to the overall knowledge about the program's interventions and outcomes. The SEP should clearly state what level of evidence, according to the SIF framework, the proposed evaluation plans to attain (preliminary, moderate, or strong), and why this level is appropriate for the program. Although SIF-funded subgrantees should generally target moderate or strong levels of evidence, in some cases a design targeting preliminary evidence may be more appropriate. If an evaluation targeting preliminary evidence is proposed, the section should provide a justification for this, and needs to describe how this evaluation will lead to evaluations targeting moderate or strong levels of evidence within the timeframe of the program's involvement with SIF. Include specifics regarding the timeframe for moving to a higher level of evidence.

Because this section was not required of earlier cohorts of SEPs, the following is an example created using the research questions above, but was not in the Latin America Youth Center's SEP.

SEP Review Checklist: Contribution of the Study

- ☑ The contribution to understanding that the evaluation will make is clearly stated.
- ☑ The level of evidence the program is targeting is described.
- ☑ How the proposed evaluation meets the criteria for this level of evidence is included.
- ☑ If an evaluation targeting a preliminary level of evidence is proposed, the section details why this is appropriate, and when and how the program will move to a higher level of evidence.

Example: Contribution of the Study

By using a randomized control trial design, the evaluation of the Latin America Youth Center (LAYC) will expand the level of evidence related to mentoring programs for high risk youth. The LAYC program evaluation will also provide evidence related to paid mentors (promotores) working with older, high risk youth in longer term relationships. The structure of the LAYC program is based on previous research that showed that these particular participant and program characteristics have not been adequately addressed either from the programmatic or research standpoint. This evaluation will therefore help LAYC in assessing program efficacy and generate moderate to strong evidence. The use of a randomized control trial research designs minimizes threats to internal validity, as it controls for characteristics of participants; the use of multiple sites should maximize external validity through the inclusion of youth from different places.

IV A Combination of Designs and/or Analyses:

In some situations, using a combination of several study designs may be useful to capture the evidence of effectiveness of the program. In the example below, the SEP outlines a randomized control trial design that will be implemented in conjunction with an interrupted time series design. Both design types are outlined briefly here.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked. The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

Impact Evaluation: Randomized Between-Groups Design

The strongest evaluation design available is random assignment of program participants (or groups of participants, program sites, etc.) to either a program participation group or a control group that is not exposed to the program (often referred to as the treatment or intervention). If individuals are randomly assigned to the program and control groups, the groups are statistically equivalent on measured and unmeasured characteristics—including unmeasured characteristics that evaluators may not have considered when designing the evaluation. Random assignment allows evaluators to infer that changes in those measured characteristics are due to the intervention, regardless of the characteristics of any of the individuals that are easily recorded (such as race or gender) or less easily recorded (such as motivation or beliefs).

Impact Evaluation: Interrupted Time Series Design

Some evaluations examine a single group of individuals before and after participation in a program, with no attempt at controlling who is part of the group. This form of evaluation, referred to as an interrupted time series design with a single group (such as a school or classroom), attempts to capture any change that occurred to the individuals after program participation by examining the general trend found in multiple measures of an outcome over time.

SEP Review Checklist: Combination Design

☑ Clear details of all design components are provided.
☑ A rational for using a combined approach that identifies how that approach addresses threats to internal and external validity is discussed.

SEP Review Checklist: Randomized Between-Groups Design

☑ Unit of random assignment is clearly identified (and aligned with the unit of analysis).
☐ Procedures to conduct the random assignment, including who implemented the random assignments, how the procedures were implemented, and procedures used to verify that 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.
☐ Any concerns that proposed strategies or approaches will lead to nonequivalent groups are discussed.

Example: Impact Evaluation, Interrupted Time Series

Excerpted from the SEP by the Communities in Schools program, a subgrantee of the Edna McConnell Clark Foundation.

This section provides detail on our proposed evaluation design for estimating the impact of CIS. The CIS model, as described earlier, operates on two levels: Level 2 intensive services targeted for individual high-need students (e.g., who typically have poor academic achievement, poor attendance, or other signs of risk for school failure or dropping out), and Level 1 preventive services for the whole school. Given the multilevel nature of the intervention strategies, our proposed evaluation design uses two complementary studies to address the effects of CIS at each of its levels. The first study is an individual-level randomized control trial (RCT) in middle schools and high schools, designed to investigate the impact of Level 2 services for individual students, and the second uses a school-level comparative interrupted time-series (CITS) approach to investigate impacts of the full model for the whole school. The RCT design seeks to evaluate the student-level impact of the most intensive CIS service provision for the students with greatest need for support, while the CITS design will explore the broader impact of CIS services experienced across a whole school at the elementary, middle and high school levels. Details of each of these approaches are described throughout this plan.

The introductory paragraph describes how each design will address specific research questions and aspects of the program (i.e., different levels of analysis).

In an effort to provide robust estimates of program impact, the evaluation seeks to measure outcomes both within and across schools, pooling results across schools to improve the power of the research to detect impacts by increasing the sample sizes for statistical analyses. When possible, analyses will also pool data across grade levels within each school, also improving the power of our analyses by increasing sample sizes. The inclusion of multiple schools and grades also improves the generalizability of the overall impact findings. This approach of pooling together findings across schools is not uncommon, and has been used, for example, when estimating the impact of the *Reading First* program.[12]

SEP Review Checklist: Interrupted Time Series Design

- ☑ Number of measurement points before and after the intervention is described.
- □ The number of measures during each measurement phase is shown to be sufficient to establish a trend and rule out rival explanations.
- ☑ The timing of measures pre/post interruption is shown to be appropriate to the intervention.
- ☑ Comparison cases are clearly described.

[12] U.S. Department of Education, National Center Educational Evaluation and Regional Assistance (2008)

Student-Level Impact Evaluation of CIS Level 2 Services:
The Random Assignment Study (Level 2 RCT)

The randomized control trial is considered the "gold standard" for rigor in evaluating program effectiveness. The key distinguishing feature of this design is that students, after being assessed for eligibility and recruited for participation in the program, but before the intervention to be studied begins, are randomly assigned to receive the services or "treatment" being investigated. In this case, a subset of students in each grade level at each CIS school will be selected randomly from a larger pool of eligible students to receive the Level 2 CIS intervention services. The other eligible students, those not selected, are assigned to a "control" or "business as usual" condition within the school and do not receive the Level 2 services. The random assignment process creates two comparable groups of students. Thus, it can be inferred that any differences in post-random assignment outcomes between students within grades and/or within schools are attributable to the impact of the CIS Level 2 services.[13]

The description of the RCT is limited in this section, but is outlined elsewhere in the SEP. Ideally, more information about the way random assignment will be done, any blocking to be undertaken, and any concerns about differences between the groups would be included here as well.

School-Level Impact Evaluation: Comparative Interrupted Time-Series Design (CITS Study)

The impact of CIS on a student and/or school outcome equals the *difference* between what the outcome was after CIS was under way and what it would have been without CIS. One can estimate this difference by comparing the change in outcomes over time for schools that adopted CIS with the corresponding change for similar comparison schools that did not adopt it (the "counterfactual"). Thus, the impact estimate represents the observed improvement of the CIS program schools relative to the observed improvement of their comparison schools.[14] Ideally the time-series design used to produce impact estimates should have data on consistently measured outcomes for multiple pre-intervention baseline years, multiple post-intervention follow-up years, multiple program schools, and multiple comparison schools.[15] Exhibit 4.a is a representation of this CITS design that uses five observations (*O*) prior to an intervention and three after it has been implemented.

The number of measurement points is described here, although it would be helpful to also have an indication that these are a sufficient number of points for the analysis.

The footnotes indicate that the timing is appropriate and also describe the comparison cases.

[13] The overall effect of CIS program on student outcomes will be based on all schools pooled together. The impact of the CIS program, by design, will be evaluated by estimating the average impact on students within each school and then pooling these results across schools.

[14] Shadish, Cook & Campbell (2002); Pedhazur & Schmelkin (1991)

[15] Multiple baseline years help to provide a reliable benchmark and trend of pre-intervention outcome. Multiple follow-up years help to provide the elapsed time needed for a reform to be implemented and thus to begin to take effect. Multiple program schools help to provide a reliable measure of change over time in the presence of the program. This reliability stems from (1) the ability of multi-school averages to reduce random year-to-year fluctuations in student outcomes and (2) their ability to "dampen the shocks" that can occur at a single school due to idiosyncratic local events, such as a change in principal. For the same reasons, multiple comparison schools can help to provide a reliable basis for estimating the change over time in student outcomes that would have occurred without the program.

Exhibit 4.a: Schematic of Comparative Interrupted Time Series Design.

$$\begin{array}{ccccccccc} O_1 & O_2 & O_3 & O_4 & O_5 & \text{X} & O_6 & O_7 & O_8 \\ \hline O_1 & O_2 & O_3 & O_4 & O_5 & & O_6 & O_7 & O_8 \end{array}$$

The key to this design is to know when the intervention occurred. The logic is straightforward: if the intervention has had an impact, the causal hypothesis is that the observations after the intervention will have a different slope or level from those before the intervention (i.e., the series should reflect an “interruption” in the prior pattern or trend at the time the intervention was delivered or implemented). The comparison group allows us to address the threat to our ability to assert a causal relationship between CIS and changes in outcomes after its implementation that some other concurrent event may have caused changes in the patterns of observed outcomes. If some other concurrent event (e.g., within a district or state) is a plausible cause of the outcomes, then we should also see similar changes in the pattern of outcomes for the comparison group schools. If not, then CIS is more likely to have been the cause. Thus, the CITS is a powerful quasi-experimental design alternative when randomization is not feasible since it combines the traditional interrupted time series analysis and a comparative schools analysis, each building on strengths of the other.[16]

[16] Bloom (2003). Corrin & Cook (1998) also discuss how different design elements (e.g., measurement of outcomes over time and matched comparison groups) can be combined to form stronger evaluation designs.

IV A Pre-Experimental Designs

Sometimes it is not feasible to conduct an experimental or quasi-experimental evaluation of a program for a variety of reasons. Instead, an initial evaluation of a program's effectiveness is more appropriate and serves as a precursor to a more rigorous evaluation. Pre-experimental design evaluations are characterized by their lack of a control or statistically matched comparison group, and often incorporate more limited data collection. Included in this design evaluation category are studies with a single group that uses a post-test, or pre- and post-test and studies with two-groups that are not randomly assigned, statistically matched, or controlled, but that use post-test or pre- and post-test comparisons. These methods of evaluation only have a limited ability to curtail problems with internal and external validity because they cannot adequately account for either measured or unmeasured characteristics of participants.

The example excerpted below describes both a pre-experimental design that would likely yield preliminary evidence, but also includes information on how the evaluation will be further developed later during the period of SIF involvement to develop a more robust evaluation capable of yielding moderate to strong evidence of effectiveness.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked. The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Pre-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 any counterfactual groups 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.

Example: Pre-Experimental Design
Excerpted from the SEP by AIDS United.

The evaluation for Access to Care is designed to have three complementary components: monitoring of national evaluation measures; case studies; and cost analysis. The outcome evaluation draws from each of these three components.

Monitoring of national evaluation measures: This portion of the outcome evaluation uses a pre/post test design of program participants without a comparison group. Data are collected longitudinally at baseline, six months (Time 1), twelve months

This section outlines the pre-experimental study design.

(Time 2), and eighteen months (Time 3) on all study participants. Baseline is defined as the date the client enrolls in the project and is most often operationalized as the date that informed consent is signed. Baseline data is gathered either at enrollment or shortly after enrollment. The windows for Time 1-3 are four months wide and are skewed to the right. Time 1 is defined as months 3, 4, 5, and 6; Time 2 is months 9, 10, 11, and 12; Time 3 is months 15, 16, 17, and 18. The reason that the windows are so heavily skewed toward lower months is that some programs are only engaging clients for three or four months, and thus requested that Time 1 include months 3 and 4 to better ensure participation in follow-up data collection. If multiple measures are taken during a follow-up window, we have asked sites to take the measurement closest to six months. Sites report the mean number of days from baseline that outcome measures were taken to monitor that measurements are representative of six, twelve, and eighteen months.

Case Study (network analysis): The case study network analysis uses a retrospective pre/post design. An on-line survey will be used to obtain retrospective and current network data on relationships between organizations. The network analysis will calculate density and average node degree. These data will be collected by asking "Does your organization collaborate with [insert name of organization] in the implementation of A2C?" and "In the six months before A2C, did your organization work with any of the following organizations to link PLWHA into HIV care and treatment?" Network analysis data will be collected approximately one year into program implementation.

Cost Analysis: We plan to conduct a cost and threshold analysis to assess: cost per client and cost per contact of delivering the program, economic threshold for the cost per HIV infection averted compared to current standard of care, economic threshold for cost per QALY averted. The threshold analysis of transmissions averted will allow us to calculate if the intervention is cost-saving and the threshold analysis of QALYs will allow us to assess if the interventions are cost-effective.

To achieve these aims we will employ standard methods of cost, threshold and cost-effectiveness analyses as recommended by the U.S. Panel on Cost-effectiveness in Health and Medicine, and as adapted to HIV/AIDS programs by Holtgrave (Holtgrave, 1998). The cost analysis will employ a U.S. Panel recommended micro-costing approach that has also been adopted by the U.S. Centers for Disease Control and Prevention. The threshold analysis will take the results of the cost analysis, and determine how many HIV infections from clients living with HIV must be averted to HIV seronegative partners in order to claim that the A2C program is cost-saving. The threshold analysis will also determine how much improvement in the quality of life of A2C clients much be realized in order to claim that the program services are cost-effective (even if not cost-saving) at a well-utilized standard of $100,000 per QALY saved. At the conclusion of the project, we will combine the cost information and the outcome data to conduct a cost-effectiveness analysis so as to determine the actual cost-per-quality-adjusted-life-year saved by the A2C project services. Uncertainty in any input parameters will be examined via sensitivity analysis so that robustness of results to changes in parameter values can be gauged.

Limitations of Initial Evaluation

This evaluation has many limitations that are worthy of mention. We are using a pre-post design to assess trends in participants' health. Because we do not have a control or comparison group, we will not be able to test causal hypotheses. Each site has a unique program, recruitment methodology, and data collection methodology, and thus combining national evaluation construct data across site may not be appropriate. In addition, due to the vast differences between sites, we are not able to make meaningful comparisons across sites. For the case studies, only one individual will be analyzing and coding the data and therefore we will not able to assess reliability. For the case studies, we will not interview all individuals who participate in the A2C program, and therefore some key information and details will most likely be omitted.

This section outlines the barriers to achieving moderate or strong evidence in the initial evaluation period.

However, selection of the participants for interviews will attempt to recruit key informants and critical staff. The decision to recruit two key staff members from each participating organization is based on best practices from the literature (Kwait, Valente, & Celentano, 2001). The cost analyses will make some assumptions, such as lifetime HIV costs of approximately $355,000 and $100K as the willingness to pay per QALY. However, these cost assumptions are based on the literature and represents standard amounts used for cost analyses. In addition, HIV infections averted is modeled and not tested biologically.

Plan to Achieve Moderate Evidence

This section outlines plans for continued evaluation during years two through five of SIF involvement, which would likely yield moderate or strong evidence.

We propose a comparative analysis of Access to Care and U.S. national data on linkage and retention in care. To better understand the unmet HIV care and treatment needs for individuals living with HIV, scientists have recently constructed cascades of the spectrum of engagement in care. The spectrum of engagement in care includes being HIV infected, diagnosed HIV positive, linked to HIV care, retained in HIV care, on ART, and having an undetectable viral load. The cascades estimate the number of PLWHA in each category and allow researchers to estimate the percentage of PLWHA who are linked to care and the percentage of PLWHA who have an undetectable viral load. For example, Gardner recently estimated that 19 percent of PLWHA have an undetectable viral load and the CDC estimated that 28 percent of PLWHA have a suppressed viral load (CDC, 2011; Gardner, McLees, Steiner, Del Rio, & Burman, 2011). The national evaluation will include a comparison of the SIF cascade to the CDC's national-level cascade. Thus, the CDC's national data on HIV linkage to care would serve as our 'comparison group'. Both cascades would include the following: Achieving a moderate level of evidence: evaluation activities for year four and five

During year two and three of the project, JHU will work with a selected site (or sites) to design an evaluation that meets the requirements of a moderate or strong level of evidence. We anticipate working closely with AU and a site to develop a detailed evaluation proposal that would outline the research questions to be answered, the evaluation design (including sampling, recruitment, and retention), an analysis strategy, and limitations. We anticipate that data collection would begin in year four.

While it is not possible at this time to speculate on any of the details of this evaluation plan, below we discuss some of the types of designs we think would be feasible. In suggesting these designs, we are assuming that we are working with a site that has access to existing medical records.

1. *Randomized comparison group design*. Randomization could occur at the individual level or at the sub-site level. If we randomized at the individual level, our sampling frame could be a list of patients who had been out of care for a year (failing to have two visits two months apart in the past 12 months). These patients could be randomized to an intervention condition (e.g. outreach followed by six months of peer navigation) or a standard of care condition (e.g. a call and referral by a case manager). If we

randomized at the sub-site level, the approximately ten to twelve partner organizations for a site could be randomized to either an intervention condition or a standard of care condition. Data on linkage to care and retention in care, CD4 and VL could be collected at baseline, six months, and twelve months for the intervention and the control group. This design would allow us to compare differences between the intervention and the control group at baseline, six months, and twelve months.

This is a fairly robust design as it includes a randomized control group which should limit biases in how participants are enrolled in the program. Because the control group would have some interaction with study staff in order for data collection on CD4 and VL to be possible, we would expect our results to be biased toward null findings. Another alternative to this design would be to only track the outcomes of linkage to care and retention in care. This could be done passively for participants in the control group, and therefore would be less burdensome.

2. *Interrupted time series design without a comparison group.* Clients who were identified as being out of care based on their previous medical records would be enrolled in the study and followed for six months. During this time, they would receive the usual standard of care and data would be collected on participants by an outreach worker on visit history, CD4, and viral load. After six months, participants would be exposed to the intervention condition (e.g. six months of peer navigation). The study would compare visit history and health outcomes prior to exposure to the program to visit history and health outcomes after exposure to the program.

   This design has appeal because all out of care participants are exposed to the program, there is simply a lag in time from their enrollment to the start of the intervention. However, this lag time could be problematic. It becomes increasingly difficult to locate out of care clients as time passes. There are additional limitations of this design that are worthy of mention. Because data are not collected on the intervention group and the comparison group at the same time, factors other than exposure to the program that occurred at the same time as exposure to the program could impact the outcome variables measured.

Given our previous experience working with sites on implementing the national evaluation, it is more likely that sites have the capacity to conduct an interrupted time series design. However, we anticipate exploring both options equally.

IV A Impact Evaluation: Non-Randomized Group Designs – Groups Formed by Matching (Propensity Score Matching)

When it is not feasible to randomly assign potential program participants (or groups or sites) to either a treatment group or a control group, a quasi-experimental research design that uses matching of participants and non-participants is an effective alternative. A comparison group can be formed by matching study participants, or clusters of study participants, on a set of pre-intervention measures of the program outcome and/or pre-intervention measures that are highly correlated with the program outcome. The main idea is to have two groups of individuals (or sites) that are as similar as possible on as many characteristics as possible.

Propensity scoring methods, or statistical assessments of similarity among participants, are preferred when groups are matched with multiple pre-intervention measures. However, the type of matching algorithm used to implement the propensity scoring should be carefully selected based on simulation studies, previous research that demonstrates the validity of the algorithms, and the goals of the evaluation.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Between-Groups Design-Formed by Matching

- ☑ Unit of matching is clearly identified (and aligned with the unit of analysis).
- ☑ Procedures to carry out the matching to form a comparison group are described.
- □ A precedent in the literature for including the variables used in the matching is included.
- ☑ Methods used to form the proposed comparison group are described such that the validity of the matching is explained.
- ☑ Reasons why the comparison group might differ from the treatment group and threaten internal validity, and the ways in which the proposed methods adjust for those differences, are discussed.

Example: Impact Evaluation using Propensity Score Matching

Excerpted from the SEP by the National Fund for Workforce Solutions program, a subgrantee of Jobs for the Future

The unit of matching is clearly identified.

The quasi-experimental evaluation study will provide rigorous estimates of the impact of NFWS/SIF programs on participant employment outcomes. **Using a quasi-experimental approach, IMPAQ will estimate program impacts by comparing the outcomes of program participants (treatment group) to the**

outcomes of non-participants who are observationally equivalent to program participants (comparison group). Implementing this approach involves the following steps:

The steps outline the matching procedures, and greater detail is provided below.

1. Collect state administrative data from states in which the evaluation sites are located.
2. Merge participant-level and ES administrative data files, appending demographic and program participation information from the UI and WIA data, as available.
3. Apply matching methods using state administrative data to construct appropriate comparison groups, comprised of non-participants with the same characteristics and resided in the same labor market area as program participants. Comparison groups will be constructed separately for unemployed and incumbent workers.
4. Construct common outcome measures for the treatment and the comparison group members based on state administrative data.
5. Produce rigorous estimates of program impacts through outcomes comparisons between the treatment and the comparison group.

Construct Comparison Groups Using Matching Methods.

One key component of this evaluation is to construct matched comparison groups for unemployed and incumbent workers who are program participants in the NFWS/SIF evaluation sites and who entered the program from April 2011 through February 2012. Our matching approach will ensure that program participants are compared to a sample of non-participants with the same observed characteristics and work history, and who reside within the same labor market area.

Matching methods have emerged as a reliable approach for producing rigorous evaluations of workforce programs, particularly when a random assignment design is not feasible. Matching methods rely on the *conditional independence assumption* (CIA): the outcome (the outcome of the individual not participating in the program) is independent of program participation, controlling for observed characteristics. The implication is that non-participants who are observationally comparable to participants can be used as a comparison group for

This section would be a great place to add in references from the literature concerning the precedent for this type of matching.

the evaluation. Matching methods provide credible impact estimates when: 1) the data include large samples of non-participants and 2) matching is performed based on rich information on participant and non-participant characteristics, employment activities over the two years prior to program entry, and local labor market.

The treatment group in this study is program participants in selected NFWS/SIF sites. To construct matched comparison groups, IMPAQ will rely on ES data, which provide rich information on the characteristics and work history of all individuals who sought state employment services. ES is particularly appropriate to identify comparison groups for a number of reasons. First, ES data includes large samples and their numbers will exceed those of evaluation site participants. Second, the ES population includes all unemployed workers seeking employment assistance and incumbent workers seeking training/education services. A matched comparison sample based on the ES data will be nearly observationally identical to program participants; matched comparison groups will be constructed separately for unemployed and for incumbent workers.

In this study, IMPAQ will use the *Propensity Score Matching (PSM)* method. PSM techniques facilitate construction of a sample of nonparticipants with characteristics that closely correspond to those of program participants. We will implement PSM using the following steps:

Step 1: Merge data – Participant data provided by evaluation sites will be merged with ES and Wage Record data from the state in which sites operate based on participant SSN. The merged data will include all available characteristics and outcomes of participants and non-participants. UI receipt and WIA participation from the UI and WIA data will also be merged with these data, as available.

Step 2: Produce propensity score – We will use a logit model to estimate the likelihood of program participation based on available control variables: 1) socioeconomic characteristics at program entry, work history, and prior services/training participation (ES data), 2) prior wages and employment (Wage Records), 3) prior UI receipt (UI data), and 4) prior participation in WIA training (WIA data). We will then use the model results to produce a propensity score for each participant and non-participant in the data; the propensity score is equal to the predicted probability of program participation based on observed individual characteristics.

Step 3: Use propensity score to match participants with non-participants – Pair-wise matching (one comparison case is matched with each participant) was once the accepted method, but recent work shows that radius matching, where one treatment case is matched to multiple comparison cases, provides the most efficient estimates. We will use radius matching to match participant cases to one or more comparison cases that have identical or nearly identical propensity scores.

Step 4: Test comparison sample and modify specification if necessary – **Once matching is achieved, it is necessary to test if participants and comparison individuals share similar characteristics. These tests involve comparisons of variable means and standard deviations between the treatment and the comparison group. If treatment-comparison differences in characteristics are detected, IMPAQ will modify the logit model specification (e.g., include polynomials to capture nonlinearities or multiplicative terms to capture variable interactions) to eliminate such differences and ensure that a successful matching is achieved**

This section indicates how internal validity concerns will be addressed in the matching procedure.

IV A Impact Evaluation: Randomized Between-Groups Design

The strongest evaluation design available is random assignment of program participants (or groups of participants, program sites, etc.) to either a program participation group or a control group that is not exposed to the program (often referred to as the treatment or intervention). If individuals are randomly assigned to the program and control groups, the groups are statistically equivalent on measured and unmeasured characteristics—including unmeasured characteristics that evaluators may not have considered when designing the evaluation. Random assignment allows evaluators to infer that changes in those measured are due to the intervention, regardless of the characteristics of any of the individuals that are easily recorded (such as race or gender) or less easily recorded (such as motivation or beliefs).

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Randomized Between-Groups Design

- ☑ Unit of random assignment is clearly identified (and aligned with the unit of analysis).
- ☑ Procedures to conduct the random assignment, including who implemented the random assignments, how the procedures were implemented, and procedures used to verify that 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.
- ☑ Any concerns that proposed strategies or approaches will lead to nonequivalent groups are discussed.

Example: Impact Evaluation: Randomized Between-Groups Design

Excerpted from the SEP for the Latin America Youth Center program, a subgrantee of Venture Philanthropy Partners

An impact study [of the Latin America Youth Center program] will use a random assignment design to compare participants with nonparticipants in terms of academic and employment outcomes, and other areas of life central to the successful transition to adulthood. The study will last 40 months. The Promotores Pathway Model (PPM) impact study will:

1) Provide a detailed understanding of how the PPM works and assess programmatic achievements and challenges;

2) Help explain how and why the PPM does or does not achieve desired impacts on participants; and

3) Improve practice and advance the field’s understanding of effective strategies for reconnecting seriously disconnected youth.

For the research design of the impact study, LAYC and P/PV will utilize a random assignment experimental design, commonly referred to as a randomized control trial (RCT). As P/PV writes in their original application for the external evaluation, ‘In a random assignment evaluation, individuals eligible for a program are randomly assigned to a treatment group that is allowed to participate in the program and a control group that is not. Thus, for all intents and purposes the two groups can be seen as identical before participation in the Promotores Pathway.’ Using this design, any differences in outcomes observed between the two groups can be presumed to be caused by participation (or lack of) in the Promotores Pathway.

These first several paragraphs identify the unit of randomization and also explicitly relate the randomization procedures that will be used. The details are further explained below. The program group and control conditions are also thoroughly laid out.

The intake procedure for the Promotores Pathway creates an ideal situation for randomization of potential participants. Each youth that is referred to the Promotores Pathway is required to go through an intake procedure where demographic information is collected. In addition, each youth will complete LAYC’s Risk Screening Tool, which contains 25 yes/no questions asking about participant behaviors, current educational and employment status, and risky behaviors. Each of the questions and answer categories in the Risk Screening Tool has weights – allowing LAYC to compile a total risk score for each youth. All youth that receive a risk score that is equal to, or exceeds, 3 is at the cutoff point deemed eligible for the Promotores Pathway.

Once a youth receives a qualifying risk score, they are then required to complete the consent process for the impact evaluation. This includes consenting to the randomization process, agreeing to take all baseline and follow-up surveys, and understanding that they are part of a research study. Those youth that do not consent to the study are placed on a Promotores wait list, and are referred to other LAYC services that they may be eligible for.

If a youth provides his/her consent, then they are entered into the randomization process (commonly referred to as the Promotores lottery). The Director of the Promotores Pathway submits a group of three to five youth two times a week (Wednesday and Friday) to P/PV for randomization. P/PV then randomizes the youth into the control or treatment groups, and the results of the randomization are then provided to the Director.

Treatment group youth are then matched with a Promotor. Control group youth are placed on a Promotores Pathway waiting list, and will remain on the waiting list until the end of the evaluation. No youth on the waiting list will receive services through the Promotores Pathway until the evaluation is completed.

If a youth is randomized to the control group, the Promotor will inform the youth of the result, and at the same time make referrals to other LAYC programs when appropriate. These referrals are based on the results of the Risk Screening Tool. For example, if the youth self-reports a potential mental health problem, the Promotor will refer the youth to LAYC‘s mental health treatment services. However, the Promotor simply makes this referral by providing contact information and does not assist the youth in fulfilling the referral in any way. This is in contrast to how the Promotores work with treatment group youth – where a Promotor will (initially) set up meetings as part of the referral process, and transport (and accompany) youth to the referral programs.

[…]

P/PV describes the random assignment procedures in the following way:

‘To further ensure the integrity and consistency of random assignment, and to facilitate monitoring of sample buildup, we recommend that sample members be randomly assigned by a third party. The study team has enlisted the services of Ewald & Wasserman Research Consultants, LLC (E&W), a survey research firm with experience in managing random assignment and tracking and conducting survey interviews with hard-to-reach populations.

LAYC staff will gather required information (including consent, contact information and required data fields from the PPM intake data) and submit that information to E&W once they are sure an individual is eligible and a strong candidate for PPM. E&W staff will review materials for completeness and notify the staff regarding any missing information, illegible notations and other omissions that might compromise the integrity of that respondent‘s intake. E&W will carefully monitor the intake data for both duplicate records due to multiple entries and other criteria that might implicate a reason to exclude a potential participant from the sample.

After intake forms have been reviewed and any missing entries obtained, E&W will assign the sample member to one of the two groups, treatment or control. The assignment algorithm will follow a 2:1 ratio. The random assignment algorithm will use blocking to ensure stable assignment ratios for each site

This paragraph explains the blocking procedure that will be used.

(the PPM is offered at three different LAYC locations – DC, Silver Spring, and Langley Park). The algorithm will also verify that an applicant has not already been assigned (for example, in the case that a control group member reapplies) and will result in a string of no more than two control or two treatment group assignments in a row.

Following assignment, E&W will fax or email each sample member's treatment or control group designation back to LAYC and the research team within 24 hours. At that point, those designated as treatment group members can begin to receive services, and those designated as control members will be notified of their status and informed of alternative options.'

LAYC's Director of the Promotores Pathway (and only the Director) submits a group consisting of 3-5 youth for randomization two times a week. By submitting youth for randomization in groups, potential threats to the integrity of the randomization are minimized.

The randomization process has worked quite well in reality. LAYC submits potential youth for randomization twice a week to E&W, and the results are provided back to LAYC via e-mail usually within an hour of submission (the maximum length of time between submission and randomization results has been 24 hours). E&W, P/PV, and LAYC all keep master lists of the randomization results for all submitted participants.

IV B Implementation Evaluation

Implementation evaluation is an assessment of how well a program does what it sets out to do. Rather than focusing on the *outcomes,* however, implementation evaluations focus on the *process* by which a program provides services or otherwise accomplishes its mission. Implementation studies center on discerning how closely the actual running of the program matches the theory that generated both the program in general, and the particular components that participants experience.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist:
Implementation Evaluation

- ☑ Specific plans for measuring fidelity of program implementation (i.e., how well the program was actually implemented) in the program group are presented.
- ☑ Plans for measuring the level of program services the program group actually received, including the criteria for assessing whether an adequate amount and quality of the program was delivered to participants, are described.
- □ Plans for assessing whether the control or comparison group received program services, including the criteria for assessing the extent to which there was diffusion of the program to the control or comparison group, are provided.

Example: Implementation Evaluation Design

Excerpted from the SEP by the Children's Institute, a subgrantee of the Edna McConnell Clark Foundation.

Program managers at Children's Institute Inc. (CII) have expressed a strong interest in including an implementation study in our evaluation to complement the fidelity assessments. We also believe that an implementation study will be important in order to provide an assessment of the overall quality of CII's program and to better understand how the EBTs fit into the overall structure of service delivery.

These paragraphs outline specific plans for measuring program implementation fidelity.

In addition to the fidelity assessment questions that will be addressed (see [previous section]), our plans include at least two rounds of site visits to answer the implementation-related research questions outlined in [another section]. During these visits, we will conduct semi-structured interviews with program managers and staff, observe program activities, and meet with other local service providers and key stakeholders. We will also augment the findings from the site visits with what was learned in conducting the fidelity assessments since implementation is a key focus of those assessments. In addition, we will work with CII to conduct analysis of CII's MIS to better understand how youth flow through program services and to determine what services in addition to the EBTs are being provided to youth.

Our current plan is to visit four campuses during the first visit (Otis Booth, Watts, Torrance, and Central LA) since these campuses offer the three EBT's that will be assessed. During the second round, we will revisit these same four campuses, and add the Long Beach campus, which only offers TF-CBT.

Interviews will be conducted in group settings as much as possible (e.g. a group of clinicians that provide FFT will be interviewed together) to maximize evaluation staff time on site and minimize disruption to program operations.

Implementation Fidelity

Consistent with the primary implementation questions outlined in [a previous section], measures of implementation fidelity of the EBTs will be assessed utilizing semi-structured interviews. Interviews with senior staff, including the president and program directors, will inform how the organization's leadership developed and implemented the systems CII uses to promote fidelity. Interviews with program managers will shed light on the overall client referral, enrollment, and assessment process (primary implementation questions 1 and 2) and the staff recruitment and training process (primary implementation question 5). Supervisors and staff (e.g. clinicians that provide EBTs and their supervisors) will be interviewed to better understand how client services are documented and tracked (primary implementation question 3), how EBTs fit within the overall service delivery structure (primary implementation question 4), how CII addresses decisions around which EBT to utilize or not utilize, and how difficulties during the course of treatment are resolved (primary implementation questions 5, 6 and 7). Since semi-structured interviews will be conducted across all of CII's sites, the team will be able to assess whether there is variation across the five sites (primary implementation question 8).

These sentences detail the data collection methods that will be used in the implementation evaluation.

If the evaluation design used a comparison group, a similar discussion of data collection to understand any diffusion of the program could be included here.

In addition, participant demographics and participation data will be collected from the CII MIS to assess the level at which the program was successful in providing services to the intended target population.

Organizational Implementation Assessment

The research team and CII are interested in better understanding CII's overall program model, as well as how the EBTs fit within the organization, especially as it relates to CII's continuum of care and its trauma-lens. Moreover, CII is interested in the value added of their other services, such as youth development and family support services. In order to address these secondary implementation questions, the semi-structured interviews described above will also address the following. Interviews with program managers and other stakeholders in the community (e.g. staff from the Los Angeles Department of Social Services, a key referring agency) will help the evaluators better understand issues related to the local context and infrastructure (secondary implementation question 9) and what other services are available in the community (secondary implementation question 10). Interviews with program managers, supervisors, and staff will also help the evaluators better understand CII's overall organization structure, service delivery, and how the trauma-lens factors into services (secondary implementation question 11) as well as the overall strengths and challenges of the CII model (secondary implementation question 12). The protocols for the semi-structure interviews will be developed approximately six to nine months prior to the first round of site visits.

IV C 1 a Sample Plan and Power Calculations

The sampling plan provides a complete description of who will participate in the study and how they will be selected. A well-thought out sampling plan can help prevent problems with internal and external study validity, so the sampling plan and processes to be used should be clearly described. A power analysis is a calculation that estimates, given a specific sample size, how likely there will be significant results in the study. There are different techniques available for calculating a power analysis, and exactly which power analysis formula will be used will depend on the details of the study, including the amount and types of information collected (the independent variables), and the desired level of explanation the study hopes to provide (the size of the expected R-squared).

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Sampling Plan and Power Analysis

- ☑ The size and composition of the sample is described and is consistent with the "Budget" section.
- ☑ Sampling plan is designed to select participants that are representative of the population from which they were selected.
- ☑ The target population from which the sample was selected and the sample selection procedures, including any geographic regions targeted, are described.

Example: Sample Plan

Excerpted from the SEP by the Gateway to College program, a subgrantee of the Edna McConnell Clark Foundation.

The target population will be drawn from nine Gateway to College program sites from across the country and will be selected based on the program's general eligibility requirements for participation. The program's eligibility requirements include students that are/have:

- Between the ages of 16-20 (and able to complete high school diploma by age 21);
- Between 5-17 credits away from high school completion;
- Current or former high school students;
- Reading at the 8th grade level; and
- A low GPA or history of absenteeism.

This section provides information concerning sample selection and representativeness.

Minimum Detectable Effects (MDEs)

The Minimum Detectable Effect (MDE) is the smallest true impact that an experiment has a "good" chance of detecting (Bloom, 1995).[17] The smaller the MDE, the more likely a study is to be able to detect impacts of a small magnitude.

SEP Review Checklist: Sampling Plan and Power Analysis (cont'd)

- ☑ Statistical power is estimated and consistent with the study design.
- ☑ The statistical power analysis used to arrive at the sample size is described, and includes the minimum detectable effect size (MDES) that has an 80 percent chance of being statistically significant at a specific alpha level.
- ☑ Outcome(s) and assumptions used in the statistical power calculations are described.
- ☑ When there are plans to conduct analyses of subgroups, additional statistical power analyses are presented to estimate those MDES.

The MDE for binary outcomes is calculated using the following formula[18]:

$$(1)\ MDE = 2.80 * \sqrt{\frac{\pi(1-\pi)(1-R^2)}{T(1-T)n}}$$

Where:

$2.80 =$ The appropriate multiplier for 80 percent power and a five percent significance level
$\pi =$ The expected control group success rate
$R^2 =$ The explanatory power of the impact regression
$T =$ The proportion of the study sample that is randomly assigned to the treatment group
$n =$ The study sample size

As equation 1 shows, the main factors that influence the MDE are:

(1) The control group's success rate (Π),
(2) The study sample size (n),
(3) The proportion of the sample randomly assigned to the program and control groups (T)

This section clearly illustrates the power and MDES calculations, and additionally describes how the 2:1 treatment to control design affects both statistics.

In MDE calculations 80 percent power and a five percent significance level are assumed, as is customary. Conservatively, it assumed that R^2 is 0, meaning that baseline covariates do not help to explain any of the variation in outcomes.[19]

[17] A "good chance" is usually defined as having an 80 percent chance of detecting an impact estimate that is statistically significant at the 5 percent level.

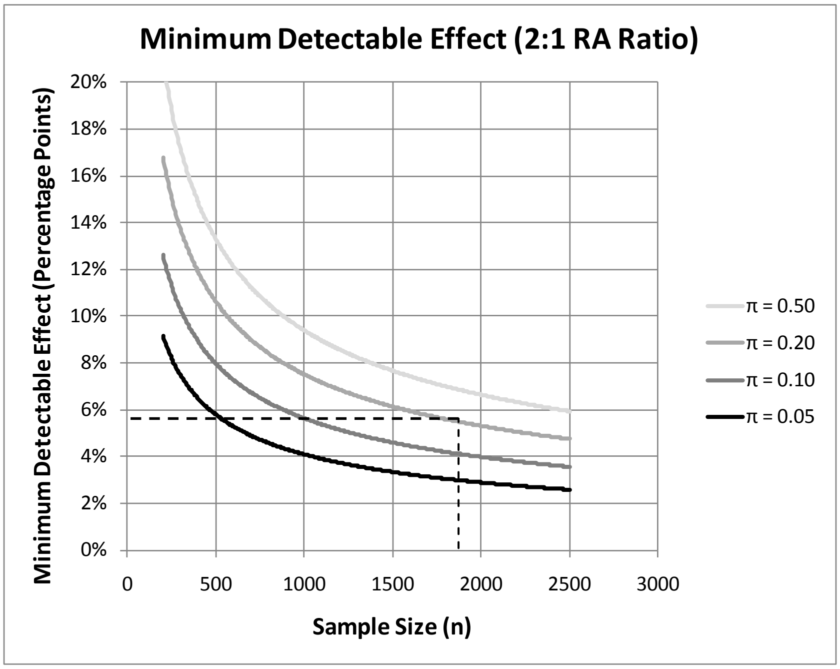
[18] Taking a conservative approach to our MDE calculations, this formula does not account for blocking, which could potentially result in a slight improvement in the precision of our impact estimates (i.e., lowering the MDE). The impact regression will include a series of indicator variables that represent each of the nine blocks. Precision gains can occur if these block indicators explain some of the variation in the outcomes of interest. The inclusion of indicator variables also results in a reduction in degrees of freedom, which is sometimes associated with a loss of precision. However, when sample sizes are as large as those that we have proposed, the affect on degrees of freedom is essentially irrelevant.

[19] We recognize that assuming r^2 is 0 is conservative. However, in our experience with postsecondary research, it has been difficult to predict outcomes based on baseline information. In some cases, we may find a correlation of .30 (r^2 = .10), which would result in a 5% reduction in the

Sample Size Under the 2:1 Treatment to Control Random Assignment Ratio

Figure 1 displays the MDE under a 2:1 random assignment ratio. As the total sample size decreases, and the expected control group success rate increases, the MDE increases. In other words, a smaller sample results in there being less sensitivity to detect small impacts. Similarly, a higher control group success rate increases the benchmark that an intervention has to surpass, resulting in there being less sensitivity to detect small impacts.

Figure 1. MDE when T=.66 (a 2:1 RA Ratio)



This section includes additional information on how the MDES would change depending on desired parameters and data collection activities.

Our power analysis finds that assuming a control group success rate of 20 percent and a target sample size of approximately 1,800 students, with 1,200 in the treatment group and 600 in the control group (see Table 4.2 for the sample breakdown by group) using the 2:1 ratio is sufficient to yield a minimum detectable effect of around 5.5 percentage points.20 The dotted line on Figure 1

MDE, but the correlations generally don't reach that level. Moreover, for binary outcomes, which are particularly hard to predict, we typically achieve an r^2 that's between .01 and .05, making the effect on the MDE negligible (between 1% and 3%).

[20] Setting the sample target size conservatively at 1800 allows room to reduce the sample size if issues arise. For example, if Gateway to College is only able to randomly assign 1350 youth, the MDE would be around 6.5 percentage points. Moreover, if the control group success rate was underestimated and was actually closer to 50 percent, the MDE for a sample size of 1350 would be just under 8.1 percentage points. Assuming an

illustrates this outcome. The third curved line from the bottom shows the MDEs associated with a control group success rate of 20 percent. At a sample size of 1800, the MDE is around 5.5 percentage points. Assuming an 80 percent response rate on the student survey, and using the same assumptions outlined above, the MDE for the survey is just over six percentage points.

With a total of nine sites, we expect that each site will contribute at least two cohorts of students. While each program site's capacity may vary and some sites may be able to contribute more students to the study sample than others, we expect that each program site will contribute approximately 100 students each year across two cohorts. A breakdown of the target study sample by cycle and site is provided in Table 4.3.

Table 4.2: Study Sample by Treatment and Control

	Full Study Sample (# of students)
Treatment	1,200
Control	600
Total	**1,800**

Table 4.3: Target Number of Students in Study Sample by Cycle & Site

	Cycle 1 Fall 2011	**Cycle 2 Spring 2012**	**Cycle 3 Fall 2012**	**Cycle 4 Spring 2013**	**Total**
Site 1	75	75	75	75	300
Site 2	75	75	75	75	300
Site 3	75	75	75	75	300
Site 4			75	75	150
Site 5			75	75	150
Site 6			75	75	150
Site 7			75	75	150
Site 8			75	75	150
Site 9			75	75	150
Total	**225**	**225**	**675**	**675**	**1800**

80 percent response rate on the student survey, under this scenario, the MDE would still be less than 10 percentage points. MDEs for subgroup analysis will be addressed when decisions are made about specific subgroups.

IV C 1 b Sample Retention

To ensure that the evaluation is as strong as possible, it is important to try to maximize the number of participants (or groups or sites) who take part in the study. This means not only recruiting participants, but also making sure that as many people as possible are part of data collection efforts during the study, as well as in any follow-up period. This is true for program participants as well as for any control or comparison group members.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Sample Retention

- ☑ Strategies to recruit and retain study participants are described.
- ☑ Alternative strategies that can be implemented if the initial strategy is unsuccessful are outlined.
- ☑ Plan describes strategies and incentives to recruit and retain study participants (e.g. remuneration for all participants). Justification is given for amount of remuneration.
- ☑ A management plan to monitor, track, and troubleshoot issues that arise in retaining the study sample during program implementation is described.

Example: Sample Retention

Excerpted from the SEP by the BELL program, a subgrantee of the Edna McConnell Clark Foundation.

The study design recognizes the need to avoid loss of the sample due to missing data. Much of the impact study will rely on existing administrative records from BELL (the application, which provides baseline characteristics) or participating school districts (which provide test score information at baseline and follow up from existing state or local standardized tests plus other student records on student characteristics, attendance, promotion, etc**). It is possible that some students served in a summer will not be tested in the following spring because they have left a participating school district. We will access records data on students who switch schools within a district.**

This section describes how participants will be tracked even if they are no longer available at the original site of data collection.

The crucial point at which sample attrition through missing data could occur in the random assignment study is the special testing of achievement the study will undertake, typically after the summer program when state or local standardized tests are not available. Our preferred approach will be to conduct this special testing at the start of school after the BELL summer program serves the program group. We recognize that some participating school districts may not want to introduce additional testing at the start of the school year and we may have to field the test in late summer, prior to the start of the school year. As long as the testing is done at the same time for the program and control groups, shifting from the start of school to late summer will not affect the internal validity of the impact estimate. **We will attempt to test all sample members attending any school in the district at that time. Testing students at the start of school will be logistically easier, but we are prepared to conduct summer testing, as was required in our first site to start, Springfield, MA.**

These sentences indicate both the preferred manner of collecting data and an additional alternative that can be undertaken as needed.

MDRC is working with Survey Research Management, with which it has partnered many times in similar studies to produce a high sample response rate on this testing. In prior studies, we have achieved response rates of 90 percent or above in such testing. This project poses special challenges, since some school districts in the project will want end-of-summer testing to occur prior to the start of school to avoid any disruption of instruction. Absence special procedures and effort to produce high response rates, this can create problems, as the prior evaluation of BELL experienced. Our plan to produce high response rates includes the following strategies:

- Hiring of local staff to make contact with sample members early in the summer to remind them of the later testing, explain how their participation is vital to learning about the effectiveness of the BELL program, provide initial information about when and how it will be conducted, and identify and begin to address issues that create barriers to participation in the testing.

This section includes information about strategies for recruiting and retaining participants, including remuneration. These points also address efforts to maintain participation among study participants.

- Help in arranging attendance at the testing can include customized information about transportation options to the testing locations, and alternative times for testing that can be convenient for parents and student with difference schedules.

- These early contacts will include information about compensation they will receive for the effort involved in getting their child to the testing. Since the testing will involve one-hour tests in both reading and math, it will need to occur on two days. Thus, we anticipate compensation for the effort involved in participating in the testing will be gift cards for a local store of up to one hundred dollars per family.

- Follow-up contact with families who miss the initially scheduled dates urging them to get their child to make-up sessions.

We expect that program group members will be much easier to locate and test as the summer ends because most will still be attending the BELL program. We intend to capitalize on this to test members of the program group still attending the BELL program prior to the end of their program participation and in a site where their program operates. We will work with BELL staff to attempt to equalize as much as possible the testing conditions for these program group members, other sample members who are in the program group but no longer participating, and students in the control group. The entire sample (program and control groups) will be tested by SRM staff. We believe that it will be possible to make testing conditions reasonably comparable across the members of the sample and the substantial savings in evaluation resources made possible by piggy-backing the evaluation testing on the end of the program will allow us to devote the resources needed to assure a high testing rate and similar timing of the testing for the control group and the non-attending program group members, which will be vital for the success of the evaluation.

Based on past experience in similar studies of programs operating outside the regular school calendar (especially our prior study of afterschool programs), we believe that these strategies will avoid differential attrition of the sample or differential timing of data collection.

IV.C.2. Measures and Measure Validity, Reliability, and History of Use

Ensuring that the measures selected are reliable, valid, and appropriate to use for the study is a key way to reduce threats to internal validity caused by the study itself. The SEP should provide a clear indication of how each measure aligns with the outcomes in the logic model. As such, each outcome should be clearly defined as a confirmatory measure or an exploratory measure.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Measures

- ☑ How each outcome is aligned with the logic model presented earlier in the SEP is described.
- ☑ If there are multiple outcomes, as indicated by the logic model and the research questions, the SEP differentiates between confirmatory and exploratory outcomes consistent with the logic model and research questions.
- ☑ How each variable from the logic model will be measured is detailed.
- ☑ For each measure, whether the measure will be developed or has already been developed (i.e., a commercially available, off-the-shelf measure) is explained.
- ☑ If the outcome measure differs across sites or groups/subgroups of sample members, how the different measures will be combined is described.

Example: Measures & Measure Validity, Reliability, and History of Use

Excerpted from the SEP by the BELL program, a subgrantee of the Edna McConnell Clark Foundation.

The outcome measures for the impact analysis will closely track the logic model presented earlier in this plan and are shown explicitly in Exhibit 3. Within the initial three-year stage of the evaluation, we will focus on the early outcomes of improvements in basic skills over the summer (with an end-of-summer test) and in the school year following program participation (with a spring test). To the extent feasible, we will utilize existing state or local tests of reading and math already being fielded in participating schools. This lessens the testing burden on students and schools and provides a measure most relevant for practitioners and policy makers. If we use existing tests, test scores from different tests across sites will be standardized (z-scored) by district to make it possible to pool impacts from different local assessments.[21]

This section and the accompanying table present the connection between the measures and the logic model and details about the measures. The final sentence also provides evidence for the proposed standardization of measures.

In most cases, for the random assignment design, we anticipate we will need to field a special achievement test for the evaluation for the end-of-summer achievement measure. Existing tests will

[21] For a discussion of the analytics and advantages and disadvantages of using state tests in evaluations see May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using State Tests in Education Experiments: A Discussion of the Issues.* Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education (NCEE 2009-013).

typically be fielded in the spring of each academic year for students in grades three and up in reading, math, and other academic subjects. For example, in the first district to begin the evaluation, Springfield, MA, the Massachusetts Comprehensive Assessment System testing will provide spring scores on English language arts, math, and science/technology for students in grades three and up. In this district, the GRADE tests of math and reading tests will be fielded to sample members near the end of the BELL program to provide an end-of-summer measure of reading and math achievements. The GRADE test is a widely used assessment with a long track record of use and scores can be provided as scaled scores, normal curve equivalents, and percentiles.

If we are able to conduct a regression discontinuity analysis of BELL's impacts in Detroit, we will rely on the fall administration of the Michigan state test for an end of summer measure of reading and math achievement. For longer follow up, we will either rely on a spring test fielded by Detroit Public Schools (if an appropriate one exists) or use the fall administration of the Michigan test in the subsequent year.

We will also field a survey of students in our random assignment impact analysis to collect information on students' attitudes toward school and learning and academic self-concept, and increased involvement in the community. We are currently exploring the advantages of various scales for use in our student surveys. Past studies of summer programs have used the Perceptions of Ability Scale Scores (PASS) survey and this is under consideration for this study, but it has not proven sensitive enough to pick up small changes in attitudes. We are also examining other survey measures for use in the survey. At this point, we have not budgeted for a parent survey given the substantial incremental costs of such a survey. If we are able to conduct a parent survey, we will also be able to estimate impacts on social skills at home and the engagement of parents. Even without a survey, we will explore these topics through interviews with summer program staff.

SEP Review Checklist: Measure Validity, Reliability and History of Use

- ☑ Information regarding each measure's reliability and validity (e.g., Cronbach's alpha), and validation against outcomes as well as historical use, if available, is provided.
- ☑ If reliability and validity of measures have yet to be determined, an analysis plan for doing so is presented.

Throughout this section, the use of specific tests as measures is described along with their history of use. Additional information on validity and reliability of these tests, especially as related to the way they will be used in this evaluation, would be helpful.

Exhibit 3: Measures and Data Sources for the Impact Evaluation

Outcome	Source	Timing
A. Student-Level Impact Evaluation		
Academic Outcomes		
Math and reading test scores at the end of the summer program	Typically test administered by evaluators	Late summer 2011 and late summer 2012
Math and reading state test scores at the end of the school year	Typically state/district test score from school records	Spring 2012 and spring 2013
Grade promotion	School records	Spring 2012 and spring 2013
Behavioral and Non-Academic Outcomes		
Attendance	School records	Spring 2012 and spring 2013
Attitudes and self-esteem	Student survey	Late summer 2011 and 2012 and perhaps Spring 2012 and 2013
Baseline Characteristics		
Demographic information	School records	Summer 2011 and summer 2012
State or local test scores (math and reading)	School records	Summer 2011 and summer 2012

IV D Missing Data

It is to be expected that some participants (or control or comparison group members) may drop out of the study, the program, or otherwise become impossible to contact. This creates "missing data," or holes in the collected data, that need to be dealt with in order for statistically sound analysis to take place. When researchers and evaluators refer to missing data, they generally are referring to information that was not collected, but could have been.

Missing data can be a problem when trying to understand the effects of a program. For example, if only half of all participants complete an entire program, but all of them show positive change, it remains unclear if the impact is due to the program or if it is due to the characteristics of the people who completed the program. If nothing is known about the people who did not complete the program, it would be difficult to say with certainty that any change found among participants was due to program participation.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Sample Retention and Missing Data

☑ How overall and differential attrition will be calculated and assessed is detailed.

☑ An outline of specific procedures planned for assessing and adjusting for potential biases, due to non-consent and data non-response, is included.

☑ Any intentions to use multiple imputation for missing data is discussed. This imputation should match the analysis to be conducted (i.e., the imputation model should use multi-level procedures if the analysis is multi-level; if the statistical analysis uses maximum likelihood or Bayes to incorporate missing data patterns, then this should also be noted).

☑ A brief description of how the plan is designed to minimize missing data with particular focus on minimizing differential attrition is provided.

Example: Missing Data

Excerpted from the SEP by Child Trends, a subgrantee of the Edna McConnell Clark Foundation.

In order to minimize missing and inaccurate data in the family finding web-based database, Child Trends provides ongoing training and technical support for data entry, and conducts regular audits of the data to ensure the completeness and accuracy of the information program staff are entering. In addition, we developed a system to extract the data from the case management system and conduct regular data exports so analyses can be conducted using SAS programming. Child Trends currently provides reports to program staff to monitor program implementation. These reports include numbers of children served, number and types of kin discovered, number of interactions that family finding staff have with kin, number of family meetings held, and reasons for family finding case closure.

For the in-person interviews, the majority of children are relatively easy to locate at the 12 month time period and more difficult at the 24 month time period as some have emancipated from foster care and no longer locatable through the child welfare agency resources. This issue is of lesser importance for youth in the treatment group given that family finding workers will have discovered and engaged a large number of relatives and other adult supports, making tracking of runaway or otherwise disconnected youth far more easier and our estimated response rates take this into consideration.

The first two paragraphs provide a general plan for dealing with missing data for the two types of data collection activities that will be undertaken in the evaluation.

As noted previously, for the full sample, data on the key outcome for the impact analysis (as well as for other child welfare outcomes) will come from North Carolina state administrative data. One potential source for missing data could be North Carolina's inability to find records for specific children for whom we request data. This could occur, for example, if a case identifier has been recorded incorrectly in the Family Finding case management database. To address this problem, we have carried out "pre-tests" involving requesting preliminary data from the North Carolina in order to ascertain how many records we are able to obtain. Results indicate that we are able to obtain administrative data records for all children who have undergone random assignment. Another potential issue could be that, even though we are able to obtain data records, relevant information may be missing from records. During the coming months, we will be examining the extent of missing data.

This paragraph addresses potential sources of missing data and how the evaluation team will attempt to mitigate them.

Similarly, with the in-person interviews, respondents may decline to answer some questions. In the case of missing data on covariates, we will carry out multiple imputation. While listwise deletion generally yields less biased parameter estimates than single imputation, multiple imputation is an improvement over both methods by producing less biased results than single imputation and by maximizing the use of available data (Allison, 2009). In multiply imputing the data, we will include a set of auxiliary variables in our multiple imputation model in order to increase the plausibility of the assumption necessary for this approach, that data are missing at random (Allison, 2009). Additionally, we will include variables used as regressands in our multiple imputation model in order to increase efficiency and potentially reduce bias (Allison, 2009). However, if records are missing on outcome variables, they will be omitted from analyses of impacts using the multiply imputed data (i.e., listwise deletion).

This paragraph explains the imputation methods that will be used to deal with missing data.

IV E Multiple Outcome Measures

The SEP should describe the ways in which the evaluation design will take into account the use of, and potential problems related to that use of, multiple outcome measures. When multiple outcome measures are specified, evaluators need to be careful of errors that may develop in comparing them. If there are multiple related confirmatory questions, or a single confirmatory question evaluated using multiple outcomes, the SEP should detail the adjustments to be made to analyses to reduce the likelihood of a Type-I error (incorrectly rejecting a hypothesis that is in fact confirmed) occurring. Techniques that can be used to adjust for such a possibility are the ordering of multiple outcomes (primary, secondary, etc.) and accordingly adjusting the p-values for each outcome, or the use of a statistical procedure such as a Bonferroni adjustment.

SEP Review Checklist: Multiple Outcome Measures

☑ If the proposal has multiple related confirmatory research questions, or a single confirmatory question evaluated using multiple outcomes, adjustments made to reduce the chance of a Type-I error are described.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

Example: Multiple Outcomes Measures

Extracted from the SEP developed by the Local Initiatives Support Corporation for the Financial Opportunities Centers' program

We plan to test multiple measures of program impacts within each domain of outcomes (employment, credit ratings, net income, net worth). The more statistical tests one conducts, the greater the probability of finding a statistically significant impact estimate purely by chance. **While we have denoted one confirmatory research question within each domain, if we do find statistically significant positive impacts for the exploratory measures, we will consider the effects of multiple hypothesis testing and present adjusted significance levels using the Benjamini-Hochberg family-wise adjustment.** This involves comparing each estimated p-value with an adjusted p-value criterion based on the formula, pi = i * (α/M) where α is the target level of statistical significance, M is the total number of p-values estimated within the domain of outcomes, and i is the rank of the

This sentence explicitly identifies the technique that will be used to address the multiple outcome measures problem.

p-value, with i = 1 through m. Estimated p-values that are less than the adjusted p-values are judged to be statistically significant.

IV D Statistical Analysis of Impacts

To ensure the strongest possible evidence results from the evaluation, the correct statistical analysis techniques must be employed. The statistical technique chosen will depend on the types of research questions and outcomes or impacts specified in the research design. The technique will also depend on the type(s) and quantity of data collected.

The gray boxes to the right contain the checklist items from the 2011 SEP Guidance for this SEP topic. If the example addresses the checklist item, then the item is checked. (Because the examples are drawn from real SEPs developed using an earlier guidance, and because not every checklist item applies to every SEP, all boxes may not be checked.) The white callout boxes below indicate the sections in the example where the checklist items are included, and offer suggestions for where missing items could be inserted.

SEP Review Checklist: Statistical Analysis of Impacts

☑ If a between-groups design is planned, an Intent-to-Treat (ITT) analysis is described, which compares outcomes between those assigned program services and those not assigned.

☑ 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.

☑ The statistical model used to estimate the program effect is fully specified and all variables in the model (and their coefficients) are defined.

☑ Assumptions of the model are listed.

Example: Statistical Analysis of Impacts

Excerpted from the SEP by the National Fund for Workforce Solutions, Jobs for the Future.

Descriptive Analyses of Evaluation Sample. The matching approach, as described above, will enable us to produce matched comparison groups for unemployed and incumbent workers who participate in training. Prior to conducting any impact analyses, IMPAQ will develop descriptive analyses for treatment and matched comparison group members based on information available in the state UI administrative data. These analyses will provide an overview of the baseline characteristics of treatment and matched comparison group members, including:

- Socioeconomic characteristics from ES data (e.g., gender, race, ethnicity, education);
- Employment history from ES data (e.g., industry of prior employment, occupation of prior employment, tenure with prior employer, prior self-employment experience);
- Wage history from Wage Records (e.g., wages in each of 8 quarters prior to program entry, wage growth prior to program entry, number of quarters with positive wages prior to program entry);

The unit of analysis and unit of assignment are specified.

- Prior participation in training programs from WIA data (e.g., participated in WIA training, WIA services received); and
- Receipt of UI benefits from UI data (e.g., prior receipt of benefits, number of IMPAQeks on UI, benefit amounts collected).

These descriptive analyses, which will be produced separately for unemployed and incumbent workers, are necessary to provide an overall characterization of NFWS/SIF participants and their matched comparison cases. Additionally, these analyses will provide evidence on whether the matching approach was effective in identifying comparison cases that are observationally similar to the treatment cases. To show that matching was done effectively, IMPAQ will perform the following tests:

- *Treatment-comparison group comparisons in observed characteristics* – We will produce means comparisons of each available characteristic in the ES, WIA, and UI data. Using t-tests, IMPAQ will assess if there are any statistically significant differences in characteristics between treatment and matched comparison group members. If matching was done effectively, IMPAQ will not detect any statistically significant differences in characteristics between treatment and comparison cases.

The steps in the analysis, and the way in which the evaluation team will examine both the characteristics of the group and the outcomes are clearly related.

- *Estimate likelihood of treatment group assignment* – We will estimate a linear regression model for each state, using the treatment and matched comparison group members. The dependent variable in this model is the likelihood of being a program participant and independent variables include all available characteristics in the ES, WIA, and UI data. If matching was done effectively, the estimated parameters will not be statistically significant.

Additionally, IMPAQ will produce descriptive statistics of treatment and matched comparison group outcomes in the 15-month period following program entry. To analyze labor market outcomes, IMPAQ will use the Wage Record data to construct the following outcomes in terms of quarters:

- *Likelihood of Employment* – Equals 1 if worker had positive wages in a given quarter, 0 otherwise. This measure will identify if treatment and control group members were employed in each of the 5 quarters following program entry.

- *Likelihood of Retaining Employment* – Two measures of employment retention will be constructed: 1) equals 1 if worker had positive wages in specified consecutive quarters, 0 otherwise, and 2) equals 1 if worker had positive wages in specified consecutive quarters from the same employer, 0 otherwise. These measures will identify whether workers were able to obtain and retain employment following program entry.

- *Quarterly Earnings* – Equals the quarterly wage amounts earned in each of the five quarters after program entry.

- *Earnings Growth* – Equals the change in quarterly earnings in each of the five quarters after program entry. This outcome measures the wage growth experienced by workers following program entry.

- *Industry of Employment* – Equals 1 if the worker found employment in the industry focus of the program, 0 otherwise. This measure will identify whether individuals were employed in the industry focus of the program.

In addition, IMPAQ will rely on WIA and UI data, as available, to construct the following outcomes:

- *Likelihood of UI Receipt* – Equals 1 if worker collected UI benefits following program entry, 0 otherwise.

- *Likelihood of Receiving WIA Training* – Equals 1 if worker received WIA training following program entry, 0 otherwise.

- *Likelihood of Receiving Educational/Training Credential* – Equals 1 if worker received an educational/training credential as a result of WIA or other training, 0 otherwise.

This section outlines the model, the variables to be used in it, and the assumption upon which the model is based. It also indicates that a *treatment-on-treated* approach will be used.

Descriptive analyses of these measures will enable us to observe patterns in the labor market and other outcomes for treatment and control group members in the 15-month period following program entry.

Impact Analyses. To estimate the impact of NFWS/SIF programs on participant outcomes, IMPAQ will examine outcome differences between the treatment group (program participants) and the comparison group. These impacts are formally termed the *impacts of the treatment on the treated.* To estimate program impacts with increased statistical efficiency, IMPAQ will use multivariate regression models, which control for available socioeconomic, employment, and other characteristics. The impact analysis regression models can be expressed by the following equation:

$$Y = \alpha \cdot T + X \cdot \beta + EMP \cdot \gamma + OTHER \cdot \delta + SITE \cdot \varepsilon + u$$

This model will be estimated on the combined sample of treatment and matched comparison groups. The dependent variable in this model (Y) is the participant outcome of interest (e.g., likelihood of employment, likelihood of retaining employment, quarterly earnings). Control variables include the following:

- *T*, which equals 1 if the individual was in the treatment group and 0 otherwise;

- *X*, which includes all available individual socioeconomic characteristics (e.g., gender, race, age, education) and a constant term;

- *EMP*, which includes variables capturing individual employment and wage history (e.g., industry and occupation of prior employment, tenure with prior employer, wages in each of the 8 quarters prior to program entry);

- *OTHER*, which includes variables capturing prior participation in WIA training and receipt of UI benefits, as available;

- *SITE*, which includes NFWS/SIF site characteristics (e.g., site fixed effects, industry focus of training, number/types of services received); and

- *u*, which is a zero mean disturbance term.

The parameter of interest in this model is α, the regression-adjusted treatment effect of the NFWS/SIF program on the outcome of interest. This parameter represents a rigorous estimate of the impact of receiving NFWS/SIF program services. The model will be estimated separately for unemployed and incumbent workers, for each available outcome of interest. To account for differences in the outcomes distribution across sites, IMPAQ will produce robust standard errors clustered by site – this will ensure that standard errors of estimated parameters are accurately estimated. Once IMPAQ estimate these models, standard errors will be used to produce t-tests to determine if the estimated program impact is statistically significant.

This section indicates alignment between the research questions and the statistical analysis plan.

We will also identify if there were differential impacts by key individual characteristics (e.g., gender, age, education). To do so, IMPAQ will include interactions between these characteristics and the treatment indicator. The parameters of these interactions will capture differential program impacts; t-tests will be used to determine their statistical significance. Similar analyses will be performed to assess differential impacts based on site characteristics, including: 1) the industry focus of the training, 2) types of services received, and 3) number of services received.