

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/275353531>

Evaluating program effectiveness: Key concepts and how to use coarsened exact matching

Technical Report · April 2015

DOI: 10.13140/RG.2.1.2609.6167

CITATION

1

READS

1,430

1 author:



[Rebecca Firestone](#)

36 PUBLICATIONS 769 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:

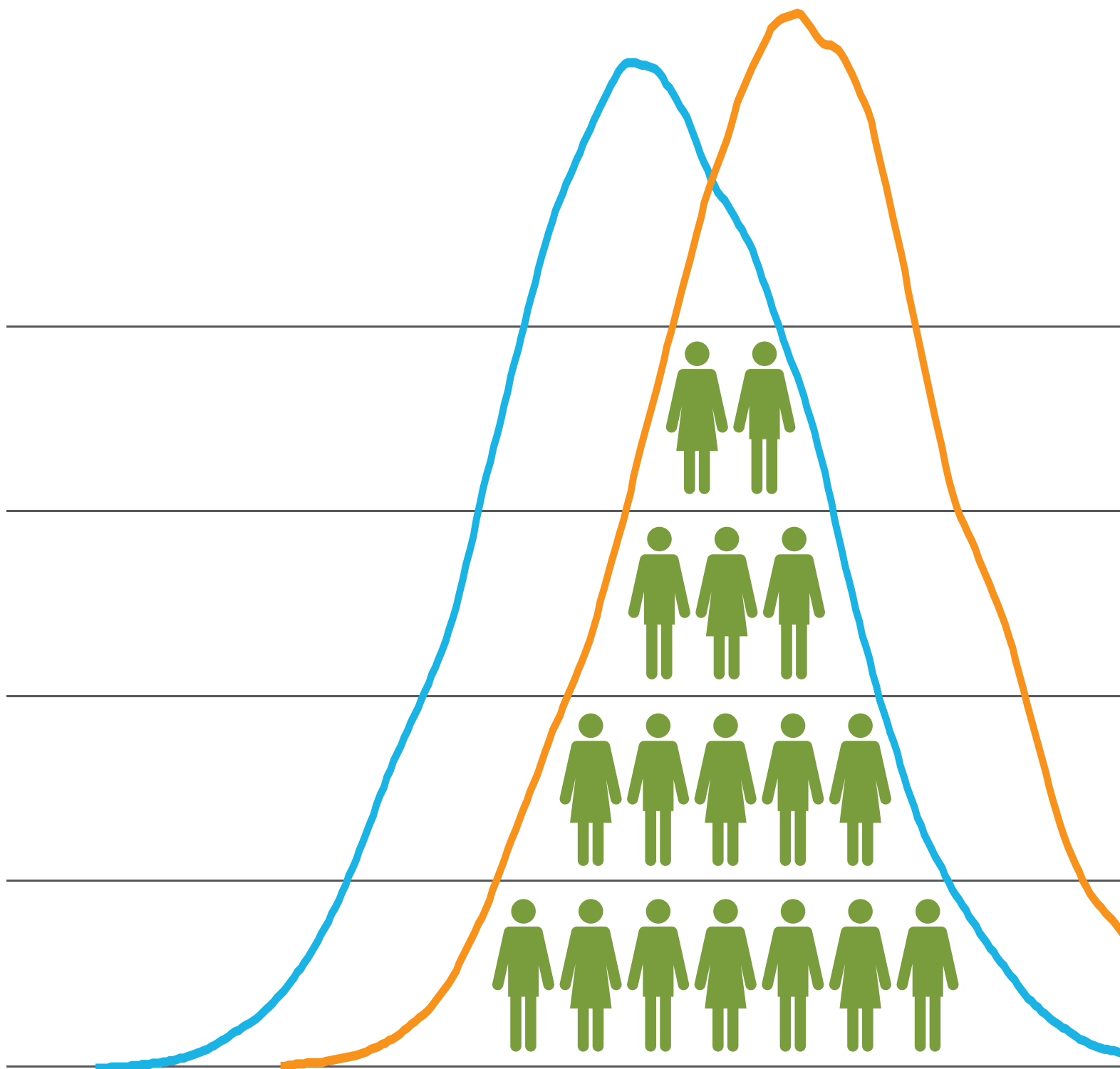


Equity Methods [View project](#)



USAID CAP-3D project [View project](#)

EVALUATING PROGRAM EFFECTIVENESS: KEY CONCEPTS AND HOW TO USE COARSENEDED EXACT MATCHING



FRONT MATTER

This document may be freely reviewed, quoted, reproduced or translated, in part or in full, provided the source is acknowledged.

RECOMMENDED CITATION

Firestone R. (2015). Evaluating Program Effectiveness: Key Concepts and How to Use Coarsened Exact Matching. Washington, DC: PSI.

ORDERING INFORMATION

This publication is available for electronic download at: <http://www.psi.org>

PSI shares its technical guides with all interested individuals or organizations. Please note that technical guides are updated periodically based on the latest available epidemiological, demographic, intervention effectiveness and utilization data. As a result, numbers used in this document should be considered as illustrative only.

For more information about this technical guide contact:

Rebecca Firestone
1120 19th St., NW, Suite 600
Washington, DC 20036
rfirestone@psi.org

ACKNOWLEDGEMENTS

This technical guide was developed by Rebecca Firestone with research assistance from Cassandra Rowe and Dana Sievers. An earlier version of the guide was created by Virgile Capo-Chichi. We would like to thank Gary King and Matthew Blackwell of Harvard University for on-going support in use of coarsened exact matching as an evaluation methodology. Special thanks to Jennifer Wheeler for contributions to earlier drafts of this document, to Nicole Bellows for editorial support, and to Amy Gregowski for technical guidance. Noah Taruberekera and Rebecca Simmons reviewed earlier drafts of the guide. We would like to acknowledge the PSI network affiliate research teams who contributed evaluation findings and model questionnaires to this guide.



© Population Services International, 2015
POPULATION SERVICES INTERNATIONAL (PSI)
1120 19th Street, NW, Suite 600
Washington, DC 20036

TABLE OF CONTENTS

04 | INTRODUCTION

05 | SECTION 1. CONCEPTS UNDERLYING EVALUATION
RESEARCH USING COARSENEDED EXACT MATCHING

13 | SECTION 2. PLANNING PHASE

21 | SECTION 3. ANALYSIS PHASE

43 | SECTION 4. DATA INTERPRETATION AND PRESENTATION

49 | CONCLUSIONS

51 | APPENDICES

51 | GLOSSARY OF TERMS

52 | INSTALLATION INSTRUCTIONS

53 | SPSS SYNTAX FOR USING CEM

54 | EXPOSURE QUESTIONNAIRES

60 | MODEL TABLES

65 | REFERENCES

INTRODUCTION

This guide provides practical and technical guidance on designing evaluations to assess whether PSI's social marketing programs are effective. This guide is intended for PSI's researchers and internal evaluators to guide decision-making on how to design an evaluation and analyze data using the technique of coarsened exact matching (CEM).

There are many ways to conduct an impact evaluation, including experimental designs with random assignment, quasi-experimental designs with non-random controls, and non-experimental statistical approaches (World Bank Group 2011). One quasi-experimental approach is the use of statistical matching, where researchers can compare program participants to non-participants, controlling for factors believed to influence outcomes of interest. PSI uses impact evaluation to determine if their programs are working. One approach to impact evaluation is the use of CEM, which is a statistical technique used in an evaluation to determine program effectiveness.

PSI researchers can conduct impact evaluation more efficiently with CEM than when conducting standard TRaC analyses. By using CEM, researchers can focus on assessing a PSI program's effectiveness through clear comparisons between who received the program and who did not. Further, an evaluation with CEM can be focused on higher levels of the [logframe](#) (purpose or objective levels) as opposed to a broad range of outputs and intermediate factors unless they have been identified as a key causal link to the primary outcomes of interest. Perhaps most critically, a CEM analysis only requires one round of data collection, whereas TRaC analyses require at least two rounds. As such, using CEM is a more cost-effective approach to evaluation.

This guide has the following objectives:

- 1) Define impact evaluation and coarsened exact matching as a method of conducting an impact evaluation,
- 2) Provide guidance on study design that will provide useful results when analyzed with CEM,
- 3) Assist in decision-making during data analysis to implement CEM effectively, and
- 4) Explain how to interpret and report results from a study that used CEM.

HOW THIS DOCUMENT IS ORGANIZED

The document is organized into four sections. The first section defines some core concepts underlying CEM in order to describe the types of questions that CEM can answer and when the method is useful. The next section walks through important decisions to be made during the study design phase when statistical matching and specifically CEM will be used during data analysis. The third section describes the data analysis steps to properly apply CEM and the fourth section provides guidance on how to report and disseminate the results.

This document should be used in conjunction with the [study planning toolkit](#) to ensure that the evaluation planned meets program needs. Any study with primary data collection involving human subjects should be conducted in accordance with PSI's research [standard operating procedures](#).

SECTION 1. CONCEPTS UNDERLYING EVALUATION RESEARCH USING COARSENEDED EXACT MATCHING

A. WHAT IS IMPACT EVALUATION?

There are several different evaluation approaches one can take depending on the objectives of the evaluation and expectations of those who commissioned the evaluation. **Impact evaluations** are used to address questions of program effectiveness by linking outcomes to a specific program, policy, or intervention (Gertler et al. 2011). For example, an impact evaluation might assess whether a communications campaign leads to increased bednet use, whether drop-in centers are effective methods for HIV prevention, or whether a voucher scheme can increase the use of health services.

Put simply, impact evaluations try to answer questions and assemble evidence about whether or not programs work. An impact evaluation is therefore designed to quantitatively test a hypothesis that a program actually caused a particular outcome. This hypothesis should be supported by a specific theory of change as to why the program is expected to influence these outcomes (Center for Theory of Change 2013). Other evaluation methods, including qualitative approaches, may be needed to account for why the relationship exists and whether the program theory of change is correct.

COMPONENTS FOR AN IMPACT EVALUATION

To answer a question about program effectiveness, an impact evaluation needs to have: (1) data on relevant outcomes of interest, (2) data on clear definitions of exposure, and (3) a study design that rules out other explanations for a quantitative finding of effect.

1. Outcomes of interest

An impact evaluation can be designed to assess changes in a variety of outcomes. The PSI **logframe** shows that PSI impact evaluations are designed to look at changes in goal-level indicators, such as measures of health status, health behaviors, or intermediate outcomes when behavior changes are difficult to observe. **Box 1** shows examples of PSI outcomes that are assessed in impact evaluations. Impact evaluations are generally not designed to estimate how many people have been reached by a program or how satisfied program beneficiaries may be with services. These may be important evaluation questions, but can be addressed by other types of evaluation.

BOX 1 | EXAMPLE IMPACT EVALUATION OUTCOME VARIABLES

Did the PSI program reduce HIV in the target population?

- Health status outcomes – reduced STI incidence
- Behavioral outcomes – increased consistent condom use, increased post-exposure prophylaxis
- Intermediate outcomes – improved condom self-efficacy, improved social support to know HIV status

Did the PSI program reduce morbidity/mortality due to malaria?

- Health status outcomes – reduced parasite prevalence
- Behavioral outcomes – increased pregnant women/children sleeping under bednets
- Intermediate outcomes – improved knowledge on malaria transmission and treatment

2. Exposure

An impact evaluation also needs to have a defined program or intervention to evaluate and a clear understanding of who has received or been influenced by this intervention. Several evaluation research techniques have emerged from medical research, particularly the methods for conducting randomized controlled trials, where the effectiveness of a medical treatment is tested. Many evaluations in public health and other fields will refer to a program being evaluated as a **treatment** or indicate that the people receiving program benefits are in a **treatment group**.

The term **program exposure** comes from communications literature to acknowledge that communications program often reach a large audience, but individuals targeted are likely to have a fairly passive engagement with the program (Morris et al. 2009; Piotrow et al. 1997). For example, they see a billboard, hear a radio message, or an IPC worker talks to them. Many PSI programs, operating at geographic scale and using mass media to influence an audience, present a particular challenge to evaluators, because it is often difficult to determine who has been exposed to the program.

The program evaluation literature also refers to **program participation** to talk about program beneficiaries engaged more actively with a program. Traditionally, program evaluation literature looks at evaluations where program participation is extremely clear: an individual must sign up or be signed up to enroll in a program and receive an intervention (Warlick 1981). In that situation, program records can be used to determine participation, and the task of the evaluator is to determine what particular factors related to the voluntary nature of participation need to be accounted for in design and analysis to make an appropriate **inference** about the effects of the program on outcomes.

For the purposes of this guide, the term program exposure is used to discuss how the people PSI serves may engage with a program. This term makes the least amount of assumptions about what PSI programs may be trying to achieve and how they engage with beneficiaries.

BOX 2 | EXPOSURE VS. PARTICIPATION

Program exposure is passive. An individual exposed to a PSI program has been reached by the program through the program's initiatives.

- Saw a mass media communication
- Read a flier
- Receive an home visit from an IPC worker

Program participation is active. A program participant makes an active decision to use PSI program services or resources.

- Went to a drop-in center
- Received services from a social franchise

3. Appropriate study design to rule out competing explanations for an effect

Attributing changes in outcomes to a program requires a study design that assists in ruling out other explanations that could account for those outcomes. It is primarily through study design that an evaluator has the flexibility and control to ensure that the evaluation is able to account for other factors, besides the program, that might influence the evaluation's outcomes of interest. Different types of experimental and quasi-experimental study designs are appropriate for impact evaluation; a review of them is provided in **Box 3**.

One key feature of an impact evaluation is the need for a designated counterfactual. The **counterfactual** is what would have happened to the target population in the absence of the program (Gertler et al. 2011). Impact evaluations make an explicit comparison between outcomes observed among people exposed and outcomes observed among people not exposed to the program in order to assess the difference made by the program.

BOX 3 | EXPERIMENTAL VS. QUASI-EXPERIMENTAL DESIGNS

In an experiment, a group of individuals (or other units of analysis) are brought together, with all of them equally willing to be exposed to a program. These individuals are then randomly assigned to receive the program (treatment group) or not receive the program (control group). The strength of an experiment, also called a randomized evaluation, is that chance, otherwise called sampling error, determines who gets the group and who does not. The counterfactual is clearly defined when someone is randomly assigned to not receive the program, and in this condition, no factors other than the treatment itself can influence the study's outcomes of interest. Random assignment between treatment and control groups directly addresses risks of selection bias, because chance, rather than individual preferences or program implementation decisions, determines who receives the program (Shadish, Cook & Campbell 2002).

Experimental designs are powerful tools for making causal inferences and considered a gold standard for designing impact evaluations. However, some potential challenges can arise when conducting a random experiment:

- It may not be ethical to withhold a program or treatment from potential beneficiaries
- The scale of the program, if it is national or widespread may make it difficult to establish a control group
- The people assigned to not receive the program may end up being able to access program benefits
- It may be difficult to track study participants over time

Quasi-experiments are a family of study designs that lack random assignment but otherwise have similar purposes and design features of an experiment (Shadish, Cook & Campbell 2002). A range of quasi-experiments are possible, with the common feature being that the program should occur prior in time to expected outcomes and that other explanations for the relationship have been made implausible due to the evaluation's design and analysis. Potential quasi-experiments include:

- Statistical matching (including CEM)
- Difference-in-difference
- Regression-discontinuity
- Interrupted time series

A counterfactual is commonly created by establishing a **control group** that does not receive the program, but a variety of study designs can create a counterfactual (Shadish, Cook & Campbell 2002). In CEM, the counterfactual is created by matching exposed and unexposed individuals on key characteristics so that they are equivalent groups appropriate for comparison.

Because an impact evaluation is trying to make a case that the program and only the program is responsible for the outcomes, impact evaluations need to ensure that all other explanations for a change in outcomes can be accounted for in order to ensure that the results are valid, i.e. that they are an accurate and truthful reflection of the program-outcome relationship in the population under study. Impact evaluations therefore need to address a range of potential threats to the internal **validity** of expected results (Shadish, Cook & Campbell 2002). There are two main sources of threats to validity: sampling error and bias.

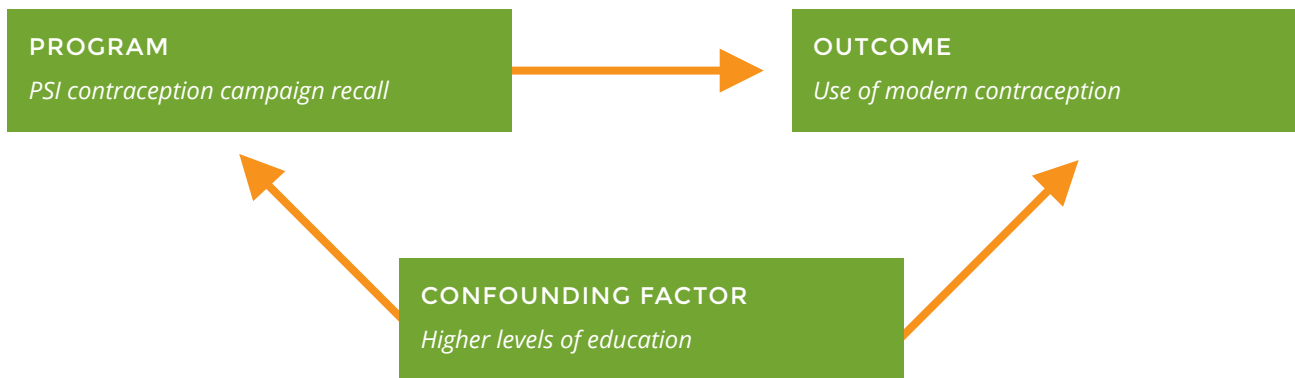
Sampling error is based on the concept that the sample population may be different from the general population. This error is based on probability theory, which states that there is always a possibility that findings from a research study may result from random variation, since it's not always possible to take a complete census of an entire target population. The best way to reduce sampling error is to ensure the study has a large enough sample size to answer the primary research question. See Step 3 in Section 2 below on sampling guidance for evaluations using statistical matching.

A **bias** is a systematic error in the design, conduct, or analysis of a study that leads to mistakes in estimates of **results** (Schlesselman 1982). When we are trying to say that program exposure caused a change in outcomes to occur, it is important to ensure that our results are not biased. A variety of different sources of bias can threaten the validity of research results and a task of all research studies is to identify and prevent these risks (Shadish, Cook and Campbell 2002). Impact evaluations need to be particularly concerned with **selection bias** – which occurs when the way that subjects are selected to participate in a study distorts the estimate of effect (World Bank Group 2011).

Many factors can influence an individual's likelihood to be exposed to a PSI program. Program participation is non-random and based on individual's voluntary willingness to participate in the program, in addition to programmatic decisions on where and how to engage with the target population. The factors that influence who actually receive the program, rather than the program itself, could be what is in reality leading people to adopt the outcome of interest. For example, young adults who recall being exposed to a condom campaign may be those who are already more motivated to use condoms. Observed outcomes in this group might indicate that the program has an effect for motivated participants, but may not reflect the program's effects in the complete target population (World Bank Group 2011). The TRaC studies that PSI commonly conducts, which use repeated cross-sectional surveys, are not designed to easily address selection bias.

When risks of selection bias are measured, they can be directly addressed in the data analysis. The term **omitted variable bias** is used when an unmeasured variable creates misleading results. This bias may still be captured in the error term of the regression model used to estimate the program's effects.

When conducting an impact evaluation, it is important to carefully research and understand the causal relationship between the program and outcome. If a third variable, called a **confounder**, is related to both the exposure and the outcome in quasi-experimental designs, one can erroneously attribute changes in outcomes to program exposure. A confounder explains variations in both the exposure and the outcome but is not caused by the exposure or outcome (Yanovitsky, Zanutto & Hornik 2005). Figure 1 below provides a visual of **confounding**, with arrows showing direction of effects. Variables that are confounders occur prior in time to program exposure.

FIGURE 1. HOW CONFOUNDING WORKS

For example, education may act as a confounding factor when estimating the relationship between a PSI-operated media campaign and use of modern contraception. Women with higher levels of education generally are more likely to use modern contraception. Women with higher levels of education may also be more likely to have seen or recalled seeing PSI's communications on family planning. Further, an individual woman is most likely to have completed her education before she ever saw PSI's communications. If education is not accounted for when estimating the relationship between the communications campaign and use of modern contraception, this estimate may not be valid.

It is important for program implementers and researchers to think about what factors could confound the relationship between the program being evaluated and the outcomes that the program seeks to achieve. These confounding factors can be addressed through the evaluation's design or through data analysis techniques, if data on these factors has been collected.

B. WHAT IS COARSENEDED EXACT MATCHING?

One research design for impact evaluation is statistical matching, which is used to compare exposed and unexposed subjects to evaluate an intervention. The purpose of matching is to take each exposed subject and pair it with an unexposed subject that has the same characteristics so that comparisons between the exposed and unexposed are not biased. Because statistical matching establishes a clear counterfactual, this design is considered a quasi-experiment, even though the analysis may be based on data collected from a cross-sectional study.

CEM is a specific method of statistical matching used in program evaluation to match persons who are exposed and unexposed to a program, based on a set of specific factors identified in the evaluation design (Iacus, King & Porro 2011). By identifying people exposed and not exposed in the data and making sure that these two groups are as similar as possible to each other on all factors that could influence exposure, we can then assume that statistically significant differences between the two groups are because of the program, rather than other factors (Blackwell et al. 2009). Statistical analysis helps address the risk of selection bias, because the evaluation specifically aims to measure the factors that could influence who participates in the program.

CEM matches data by categorizing each of the covariates identified to be included in the matching procedure¹. A **covariate** is an analysis variable that can affect the relationship between the dependent variable and the independent variables of interest, typically the outcome variables. Some covariates tend to have distinct categories, such as religion or district of residence. Other covariates are more nuanced or continuous and need to be grouped into broader categories for matching. In CEM, this categorization process is called **coarsening**, where some details in the variables are lost in order to increase the likelihood of matching exposed and unexposed cases. **Box 4** gives examples of coarsening variables for CEM.

BOX 4 | EXAMPLES OF COARSENEDED VARIABLES

Age = 48.54 years	<i>coarsened to</i>	47-49 years
Education = 5 years	<i>coarsened to</i>	primary only
Hours of radio/week = 17	<i>coarsened to</i>	15-20 hours/week
Asset index score = 1.456	<i>coarsened to</i>	2nd wealth quintile

Once covariates are coarsened appropriately, based on the variability and size of the sample, each individual in the exposure group is matched with one or more individuals in the control group who have the same specified covariate values. Matching individuals in this way produces treated and control groups that are similar with respect to the covariates. CEM assigns each case into one of a specified set of strata in which members are exactly matched on a set of coarsened, i.e. categorized, variables. Matched members are then assigned a weight specific to that stratum and representative of the proportion of all members present in the stratum.

The aim of the matching procedure is to reduce **imbalance**, the extent to which exposed and unexposed groups are different from each other on values of the variables used in matching. If exposed and unexposed groups are balanced on their covariates, individuals in each group should not be measurably different from each other on important factors that influence probability of exposure. Therefore, if a difference is found between exposed and unexposed groups on the outcome of interest, the only likely reason for this difference should be the program that the exposed group was part of.

Coarsening facilitates identification of matches but in the subsequent analysis, only the uncoarsened values of the covariates of the matched units are maintained. The data analyst then proceeds to test for program effects in a smaller sample of just the cases that have been matched in order to estimate the relationship between program exposure and the outcome of interest. See section 3 on analysis for more detail on how to implement this analysis.

CEM can be implemented with almost any type of quantitative data in order to develop statistically equivalent groups. The method discussed in this toolkit is for program evaluations requiring population-based survey data.

¹Note that CEM was created as a method for researchers who were not involved in primary data collection and thus did not have control over initial study design or how data were collected.

C. ADVANTAGES, APPLICATIONS, AND LIMITATIONS OF CEM

Previously, PSI research supported an evaluation analysis approach within TRaC, in which baseline levels of outcomes were compared to follow-up levels among individuals both exposed and unexposed to PSI programming.

CEM has several advantages over this analysis:

1. CEM accounts for effects of who self-selects into receiving the program, a major risk of bias in the previous analysis,
2. A clear counterfactual is defined, and the definition is developed during study design, rather than during data analysis, and
3. Operationally, CEM is easier and more cost effective to conduct because it only requires one round of data collection and therefore does not require a baseline survey.

While CEM is a useful tool for measuring program effectiveness, it is not a one-size-fits-all evaluation method. Matching studies are subject to limitations due to omitted variable bias. CEM has no way of accounting for factors not measured, unlike some other evaluation designs. The study and its inferences are only as good as the variables that are measured, and we cannot account for unmeasured **endogeneity**, when the outcome variable is associated with the error term (Wooldridge 2009). Thus, the imperative is on the research team in conjunction with programmers to develop extremely clear measures of exposure and completely identify potential matching variables.

CEM should be used when the method fits well within program implementation plans and when results generated from the analysis are of strategic importance to internal and external stakeholders to the extent that it's worthwhile to do the evaluation. As with any form of program evaluation, a CEM study requires time and effort on the part of program, marketing and research teams. CEM is a good evaluation approach when there is an internal or external audience that wants to understand whether differences in outcomes can be attributed to a specific program, however, if there is no demand for results on program effectiveness from either a PSI decision maker or an external stakeholder, then a CEM analysis is unnecessary.

PSI's application of CEM is based on one round of data collection. As such, it can detect differences between groups, but not changes over time. **Temporality** – ensuring that program participation occurs prior in time to the desired outcome – is a key aspect of making valid causal inferences. If it's not plausible to imagine that the population has had a chance to change (whether knowledge, attitudes, norms, or behaviors are of interest), it's not wise to design an evaluation to test for these changes.

Program reach is also an important consideration. PSI's use of CEM has been designed for implementation with population-based surveys that capture a large population. The sampling and assumptions behind CEM work well when data collection can pick up large numbers of people who have been exposed to the program and large numbers of people who have not been exposed. This is usually most effective when the program has achieved geographic scale and is reaching large numbers of people. Additionally, if it is difficult to identify people who have not been reached by the program or if they cannot be captured in the same data collection approach as for those who have been reached, CEM will not be a good evaluation approach.

CEM works well for situations when comparable data from prior to the start of program do not exist or when baseline data collection occurred while the program was being rolled out (making a comparison difficult). CEM makes comparisons between exposed and unexposed at one point in time, and therefore assessments of change over time are not needed. If the program is focused on getting good baseline data to be able to understand patterns in the population prior to implementation, consider whether designating control groups in advance is feasible in terms of logistics and budget.

Another consideration is the extent that the PSI intervention is the sole or primary intervention in an area and therefore differences in outcomes between exposed and unexposed individuals can be attributed to the intervention. Sometimes PSI programs operate in conjunction with other complimentary programs and therefore it is hard to isolate the effect of the PSI intervention. In this case, CEM is not a good evaluation approach. Additionally, current versions of the CEM algorithm can only test exposed vs. non-exposed and don't have a capacity to easily test a dose-response effect of program exposure (Reynolds 1998). If a dose-response analysis is desired for program design reasons, another approach should be used.

The CEM approach discussed in this document focuses on retrospective matching, where matches are made after data collection has been conducted, as opposed to creating matches prospectively. Several other research designs are available for situations when CEM is not an appropriate approach. **Table 1** summarizes the areas to consider when deciding whether CEM or another analysis approach should be used.

TABLE 1. WHEN TO USE AND NOT USE CEM FOR IMPACT EVALUATION

DIMENSION	CONSIDER CEM	CONSIDER OTHER METHODS
Key objective of evaluation	To assess whether intervention is effective in changing health outcomes and behaviors	Program effectiveness not a primary research objective of program managers or stakeholders Dose-response analysis to desired for program design purposes
Temporality	Sufficient time for program to result in changed behaviors related to desired outcomes (time will vary based on program)	Program is in implementation stage or insufficient time for program to result in changed behaviors
Program size	Of sufficient scale to be captured in population-based surveys	Program operations have a very targeted geographic focus or reach a small number of people
Identifying exposed and unexposed	Both exposed and unexposed individuals will be captured in a population-based survey	Collecting data on people who have not been reached by program activities will require a separate strategy for data collection
Baseline data	No baseline or baseline data are messy	Good baseline data are collected to understand patterns over time
Complimentary interventions	Similar initiatives are not operating in the area where data will be collected or it is possible to isolate the specific program being evaluated from other initiatives	PSI operations occur in conjunction with other stakeholders' programming to the extent that it is hard to isolate specific effects

SECTION 2. PLANNING PHASE

A REMINDER: START EARLY

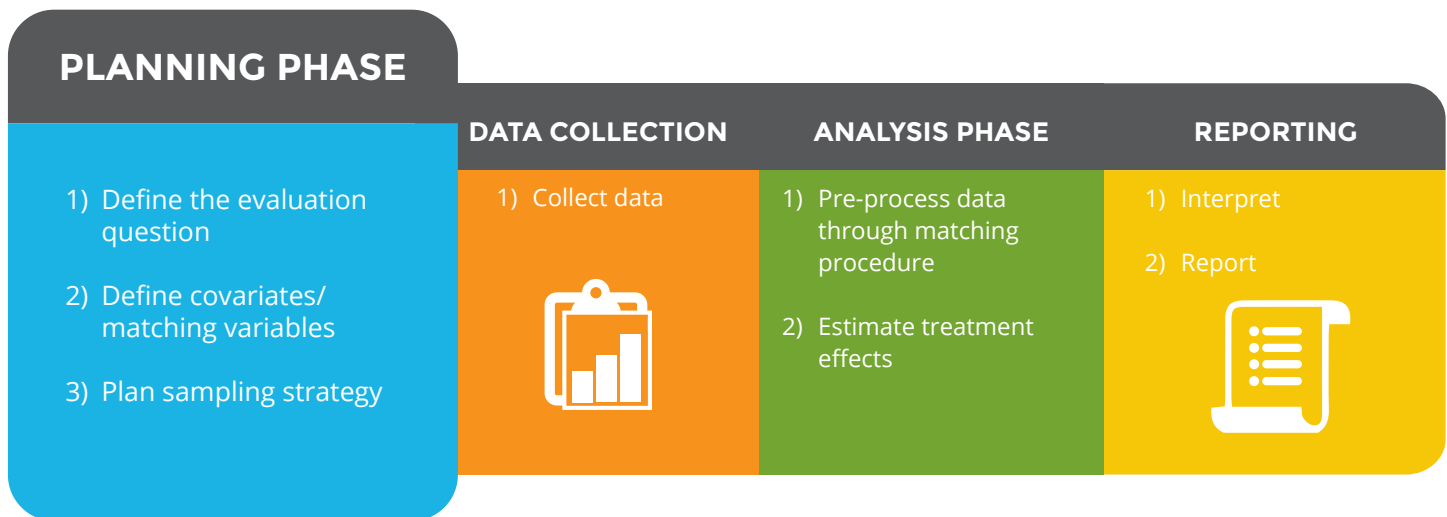
As a general rule, impact evaluation studies are best constructed when they are planned early in the program design and implementation stage. Ideally, the initial evaluation design should be identified during project development, before any implementation of project activities begins.

CEM has specifically been identified as a rigorous evaluation method for PSI programs that are already operating as some level of implementation, when it is not feasible to establish a priori control groups. Nonetheless, early evaluation planning is important in yielding a better study. A rule of thumb is to start planning at least 6 months before data collection is scheduled to begin.

Starting early with a CEM study means that program and research teams:

- Agree that CEM results are needed before a quantitative survey is designed. See the [study planning toolkit](#).
- Develop a common definition of program exposure based on how program operations have been implemented. Researchers use this definition to design an exposure section of the questionnaire to capture this definition.
- Develop a common understanding of what factors are likely to influence exposure to the program and confound the ability of the study to develop an unbiased estimate of the program's effects. Researchers use this understanding to incorporate questions that measure these factors into questionnaire design.
- Understand the geography of program operations, and researchers use this knowledge to design a sampling strategy that captures both exposed and unexposed cases.

The following sections provide guidelines and technical advice to facilitate this planning.



STEP 1: DEFINE THE EVALUATION QUESTION

A good impact evaluation should start with clear identification of the outcomes the program being tested is expected to achieve, an understanding of what constitutes exposure to the program, and a rationale or theory of change for why the program is expected to have these effects (Windsor et al. 1994).

1a. Defining outcomes

As a first step, it is important to establish the outcomes that the evaluation will be investigating. What does the program expect to have achieved and what changes are expected in the target population?

The project logframe is a good place to start in identifying outcomes to measure, particularly purpose-level indicators, such as behavioral or use indicators. Using **standard indicators** is recommended, and PSI's standard questionnaires are designed to measure most donor-recommended indicators of behavior change and product/service use.

Project logframes capture intended outcomes, but a program may also have unintended outcomes, which could either be positive or negative. For example, an HIV prevention program for men who have sex with men may have focused messaging on the use of water-based lubricant with condoms. An evaluation of the program might wish to also assess whether men participating in the program are more likely to use all forms of lubricants, including oil-based lubricants that are not recommended for use with condoms.

1b. Defining exposure

CEM requires program exposure to be well defined because the statistical procedure for CEM uses a binary exposure variable to match exposed and unexposed cases.

With PSI's programs the task of identifying participation or exposure can be very challenging. When evaluating the program after implementation has already started, we generally have to rely on an individual's recall of having seen a PSI communication, interaction with PSI or partner staff, or entered a PSI-operated facility (in some cases there are program records that can be used). Data collection for the evaluation must use cues to help survey respondents recall their exposure to the program, without biasing their recall. And the extent to which an individual needs to interact with program activities in order to change outcomes may be unknown or unclear.

Nonetheless, CEM requires the PSI evaluator to put survey respondents into one of two boxes: exposed (treated) or unexposed (controls). There are a variety of ways in which this can be done, but the definition of exposure requires program and research teams to have a shared and in-depth understanding of how the program operates (Trochim 2006; California State University n.d.).

BOX 5 | EXAMPLES OF EXPOSURE TYPES AND VARIABLES

- Exposure to any program activity (e.g. any recall of PSI-branded mass media)
- Exposure to a specific channel of communication (e.g. any recall of interaction with a PSI IPC agent/outreach workers)
- Receipt of a specific services (e.g. any recall of a visit to a PSI-branded drop-in center, social franchise, testing center, use of a PSI-branded voucher)
- Exposure to a specified package of services and/or communications (e.g. recall of mass media messaging plus interaction with an IPC agent/outreach worker)

A good definition of exposure requires:

Plausibility: There must be a reasonable justification for why program exposure is likely to lead to a change in behavior among the people reached by the program. This justification can be based on existing literature or programmatic experience. Consider, for example, whether it is plausible for one conversation with an IPC agent to lead a woman to have an IUD inserted or whether one television spot could lead to consistent condom use.

Temporality: Program exposure must occur before the outcome occurred. For example, if the evaluation study asks about receiving an HIV test in the past 12 months, but a drop-in center only started rapid testing within the past 3 months, then the drop-in center may not be the primary factor driving people's willingness to seek out HIV testing services.

Box 6 contains questions that should be discussed at a planning meeting during the design phase of an impact evaluation study. Incorporate inputs from this planning meeting into a section of the questionnaire on program exposure that enables the research team to capture different definitions of program exposure to be tested. These planning conversations will help the research team to get to a measureable evaluation question that is in line with program priorities.

BOX 6 | DISCUSSION QUESTIONS FOR DEFINING EXPOSURE DURING PLANNING STAGE

- What is the full range of program activities?
- What channels of communication will the program use to reach people in the target population? (e.g., IPC, Mid media, Mass media, Drop-in centers, Clinics)
- Where will activities be implemented during the projected period of data collection?
- Is there a minimum level of exposure to activities that programmers think is necessary before any changes in outcomes would be likely to occur?
- Does an intended beneficiary need to be reached by multiple channels before a change would be likely to occur? What combination of channels is likely to be required for behavior change to occur?
- What level of exposure and/or combination of channels is thought to be optimal in order to change behaviors/outcomes?
- Is planned dose of exposure thought to be necessary to change behaviors? For example, are there a minimum number of program contacts needed before change may occur?
- Can exposure be verified in advance through the project MIS, such as through vouchers or referral coupons?

By carefully defining the evaluation question, researchers can move from a general interest in the program's effectiveness to specifically operationalized research questions. An example is given in **Box 7**.

BOX 7 | GOING FROM AN IDEA TO A MEASURABLE EVALUATION QUESTION

Concept: We need to tell our donor whether our HIV prevention program with transgender women works

Measurable evaluation question: Is exposure to an HIV prevention program, including outreach and drop-in center, associated with consistent condom use or HIV testing among transgender women?

STEP 2: DEFINE COVARIATES FOR MATCHING

After defining the evaluation question, the next step for a CEM evaluation is to think about other covariates the evaluation will need to measure what will act as potential matching variables during analysis, keeping in mind the importance of including potential confounders in the questionnaire to avoid omitted variable bias. The aim in this stage of planning is to ensure that all potential covariates have been identified in advance during planning and are then incorporated into the study questionnaire. The section below provides advice on issues to consider and methods for identifying these covariates.

2a. Identify risks of selection bias

Many factors influence an individual's likelihood of being exposed to a PSI program, and these factors will vary substantially from program to program. Program exposure is non-random and based on an individual's voluntary willingness to participate in the program as well as implementation decisions on how to reach the target population. So, the factors that influence who actually receive the program, rather than the program itself, could be what is leading people to adopt the outcome of interest.

The best way to address this risk in a CEM evaluation is to think through and identify these factors in advance and ensure that they are captured on the questionnaire and thus be measured. Any factors not measured cannot be matched on during data analysis. For example, mass media campaigns using billboards require viewers to have enough literacy to understand the billboard. The evaluation may therefore need to measure literacy or education levels to understand who of the target population was and was not exposed to the billboard. An evaluation of a program using drop-in centers may need to ask about types of occupation and hours worked during the day because the drop-in center only operates at certain times of the day, which may preclude some people from using these services. **Box 8** contains questions that to be asked when considering covariates to rule out selection bias and identify potential factors for matching.

BOX 8 | QUESTIONS TO IDENTIFY POTENTIAL COVARIATES RELATED TO SELECTION BIAS

- When and where do program staff interact with the target population?
- How does the program contact the target population?
- Does the program assume that the target population is literate?
- Does the program assume that the target population consumes some form of mass media?
- Is the program more likely to reach men than women? or more likely to reach specific population groups than others?
- Are outreach strategies based on membership in particular social networks?
- Does outreach occur at specific times of day that would make some people more likely to interact with an IPC agent than others? For example, if IPC agents interact with household members during the day, they may be more likely to interact with individuals who are not formally employed.
- Are services at any program facility open at specific times of day or days of the week?
- Does the target population work specific hours?
- Are there other similar programs operating at the same time or in the same location as the program?

2b. Define potential confounding factors of the relationship between program exposure and the outcomes of interest

Different types of programs may have different causal pathways by which a program can have an effect on behavior. For example, mass media campaigns differ from interpersonal visits and therefore the covariates to match on would be different. After thinking through the factors that influence selection into program exposure, the next step is to think about confounding factors that can be measured for matching purposes. Many researchers will look first to socio-demographic factors often accounted for during analysis, such as age, gender, or socio-economic status. Avoid the temptation to match on these factors because they are always accounted for in analysis. Matching is successful when the correct potential confounders are in the analysis, not just the usual suspects.

Do not include factors that occur after program exposure as potential confounders, for example, factors that are potential consequences of the program. This is critical because matching on (or otherwise adjusting for) post-treatment variables can bias all subsequent inferences. For example, if we are measuring any exposure to a PSI message, we would not consider whether a survey respondent understood the message as a potential covariate to match on.

After developing a list of potential covariates for matching, start to prioritize these factors by breaking them into two groups:

Must-haves: These are factors that are absolutely critical to how the program reaches the target population, and these factors must be included in evaluation of the program. For example, if a program reaches its audience through radio spots, then the evaluation must account for how much radio an individual consumes. This could be structured as a yes/no question on the questionnaire, or the question of radio consumption could be more detailed, such as how often a survey respondent listened to the radio in the past week or month, or days of the week/times of day when a respondent is likely to listen to the radio. The specific structure of how radio consumption is asked on the questionnaire will depend on the program's media buy for that specific radio spot.

Nice-to-haves: These are factors that programmers suspect may have an influence on program reach or how the target population interacts with program activities, but program staff are less certain how important they may be. For example, a program reaching female sex workers with HIV prevention messages in Central America may suspect that if sex workers migrated to a neighboring country for work, they might be harder for the program to reach if they are in the country illegally. Asking about nationality on the questionnaire could account for this concern.

Finally, ensure that these factors have been incorporated into the questionnaire. Not all of the covariates identified with the program team may be used during data analysis, but the best practice is to ensure that as many as possible have been measured during data collection.

One area of confounding to consider is exposure to other programs. When designing and analyzing data for an evaluation conducted with CEM, it is important for programmers and evaluators to think about the effects that other programs may have on the ability to estimate the relationship between the PSI program and outcomes. In many instances, it may not be worthwhile to try to isolate the effects of the PSI program. If PSI works heavily in partnership with other organizations and/or if several other similar initiatives are operational in the same areas where PSI is working, an evaluation design that assesses the collective effects of these other initiatives may be more worthwhile than just looking at PSI-related effects (Victora, Black & Bryce 2007).

If it is worthwhile to tease out the effects of PSI's program alone, other programs may act as a confounding factor or reflect underlying selection bias. With regards to confounding, it is possible that the other program may also influence people's behaviors and the outcomes of interest for the evaluation. Further, there may be overlap in who receives the PSI program and the other program. This situation meets the definition of confounding, and participation in other programs should be measured and considered for matching. For example, ask about the presence of other programs promoting similar products or services in the same geographic areas. In terms of selection bias, it is possible that some other factor or factors motivate people to be exposed to both the other program and the PSI program. It is useful to think about what these factors could be and ensure they are measured and considered for matching. For example, education level may influence both individual's media consumption and his/her condom use.

STEP 3: PLAN SAMPLING STRATEGY

3a. Ensuring sufficient exposed and unexposed cases

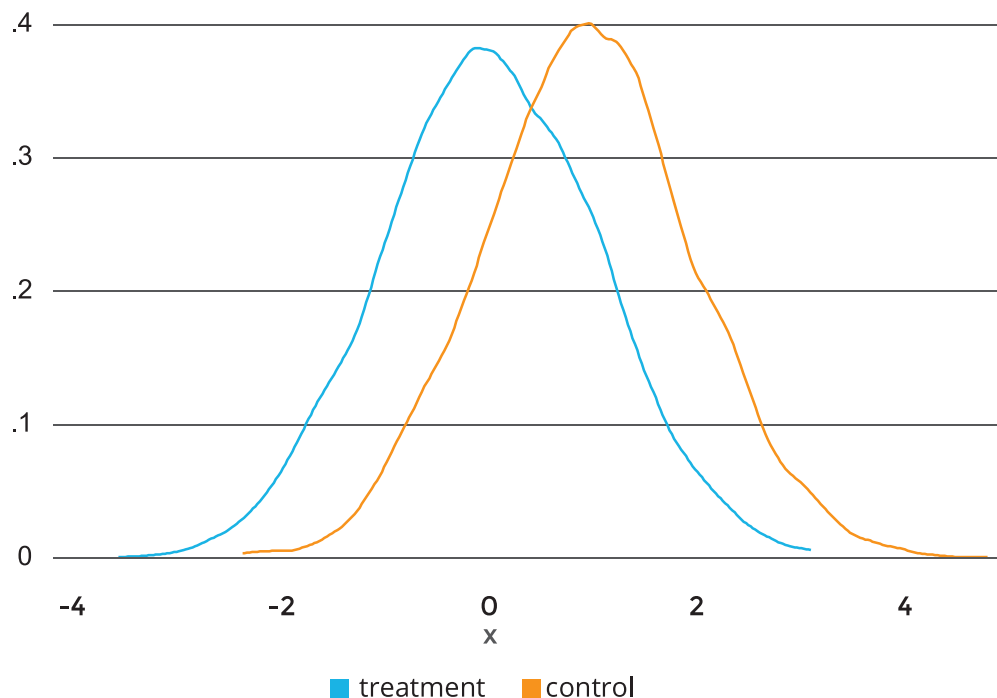
In most cases at PSI, evaluations using CEM will be based off of population-based surveys. However, it is worth remembering that statistical matching is an analysis technique that can be used with any type of data, including routinely collected data. The primary data requirement for statistical matching is that there be a sufficiently large sample from which to match exposed and unexposed cases. Statistical matching is a data analysis technique developed by economists and social scientists, who often do not have control over how data are collected. Guidance on sampling strategies for studies involving statistical matching is therefore not well developed. At best, we can flag issues for a research team to consider during planning.

The area of common support:

A major assumption of statistical matching is that people with the same values of the covariates used for matching can be found in both the exposed and unexposed groups (Heckman, LaLonde & Smith 1999). As such, one must assume that people who received the program have a combined distribution of a set of covariates used during matching, and that this combination of covariates is normally distributed. People who did not receive the program will also have a distribution of these covariates as well (**Figure 2**). This is called the **area of common support**. Analysis of matched data compares only comparable individuals. This means we can test that the program works among those individuals whose characteristics appear in both exposed and unexposed groups. More simply defined, common support describes a concept that there is “overlap” between the treated and control groups (exposed and unexposed) in terms of the covariates matched on. That is, the treated and control groups should have the same ranges of their distributions.

As an example, suppose that the treated group had extremely wealthy respondents, but the wealth distribution for the control group only went up to the middle class. In this example, we could not assume an area of common support. Once we use CEM, we would drop the wealthiest individuals. As a result, the two distributions would have the same support.

FIGURE 2. COMBINED DISTRIBUTION OF A SET OF COVARIATES TO MATCH ON, SHOWING AREA OF COMMON SUPPORT



This process is necessary for successful matching during analysis. However, it also means that treated and un-treated cases that have been matched during analysis resemble as much as possible the study population. Our task when designing the initial sampling strategy for the study is to ensure that we have a sufficiently large enough area of common support to be able to estimate a difference between the exposed and the unexposed group, taking only exposed and unexposed cases from within the area of common support. We are not going to know in advance how the covariates will be distributed in our exposed and unexposed groups.

Trimming of unmatched cases during analysis:

A second implication of matching is that it reduces the number of observations available for analysis, since unmatched cases are taken out of the sample being analyzed. The evaluation will therefore need to plan during the design stage for the possibility of a smaller sample during analysis. If we knew the proportion of cases that would be trimmed out during matching, this would be simple. However, as noted above, we will not know in advance how the covariates to match on will be distributed among exposed and unexposed cases. The magnitude of the cases trimmed out during matching out depends on several factors: geographic proximity of treated/exposed and un-treated/unexposed cases, the number of matching variables and the potential for unusual (impossible to match on) combinations of matching variables.

As an example, a recent [PSI evaluation](#) using CEM to assess the effectiveness of HIV prevention for transgender women in Thailand started with a sample of 308 cases (Pawa et al. 2013). After matching, the study worked with a smaller sample. This was still sufficient for the evaluation to be able to detect a difference between exposed and unexposed cases on several key outcomes.

Sampling approaches to ensure common support and successful matching:

Several steps can be taken while designing a study sample to address these issues. Again, because statistical matching was initially designed as a technique used after data had been collected, there is little available guidance on sample size strategies specifically accounting for statistical matching. The guidance provided here should be considered as suggestions to consider in light of considerations made while developing a sampling strategy.

- To avoid a larger number of un-matched observations (whether in the treated or un-treated groups), both groups should come from the same general population. We should ensure that the probability of being exposed (or not exposed) does not depend on something so particular that only a small minority of the population has this characteristic.
- Consider sampling from areas with and without the program that are close to each other geographically. This tactic will be easier to achieve when evaluating a program that uses IPC, as opposed to a program with a mass media component, where there is minimal programmatic control over geographic coverage.
- Researchers should be aware of how potential covariates to match on may be distributed in the population from which the sample will be drawn. However, it is not necessary to power a sample to detect a specific level of a covariate used in matching. Focus on powering the study 1) to detect differences between the exposed and unexposed groups and 2) to ensure sufficient numbers of exposed and unexposed cases. If a potential covariate is very rare or hard to measure, it would be advisable to drop that factor from the list of covariate to match on, rather than boost the evaluation's sample size.
- Consider oversampling by exposure area through stratification. A possible risk with sampling for CEM is when one of the groups (exposed or unexposed) is too small. In this case, random sampling of the population to ensure adequate numbers of cases in both groups would be difficult. When this is the case, it is much more cost-effective to stratify the population by treatment group, drawing a sample from the treated group and another from the control group. However, this requires knowing who belongs to the treatment group and who belongs to the control group. In most settings, ensuring a sufficiently large pool of potentially exposed cases is a greater challenge than locating unexposed cases. Small-area (district-level, enumeration area-level) documentation of program implementation may be needed from program teams to develop a sampling strategy that is structured to ensure that these areas are reached.

For a probability sample, which uses some form of random selection, consider the following:

- Stratify the sample into intervention and non-intervention areas.
- Calculate required sample size per group (exposed and non-exposed) assuming 1 exposed for 1 unexposed.
- All people in intervention area are exposed.
- To calculate the per-strata sample size, you will need (as usual) estimates of the % of the total population that is an eligible program participant (for example, % of total women that are women of reproductive age). Select stratum-specific sample based on sample selection strategy.

For non-probability sampling, specifically respondent driven sampling (RDS) or time-location sampling (TLS):

- It is not possible to systematically stratify on exposure.
- For RDS, consider selecting cities and/or social networks where the program is operating and where the program is not operating. Population size estimates for these cities will improve the evaluation, but are not necessary. It is also advisable to select some seeds, or initial respondents who recruit other respondents, who have a connection to the program and some who do not. This increases the chance that respondents recruited will include both exposed and unexposed cases.
- For TLS, consider selecting locations where the program is likely to be operating and where program is not likely to be operating. See sampling guidance for best practices in designing studies with RDS and TLS.
- One possible strategy to ensure a sufficient number of exposed and unexposed cases would be simply to inflate the study's required sample size by a specific percentage. This is an option, but before considering it, the research team should check with the program team to learn how the probability of program exposure is likely to be distributed in the population being sampled. If the program targeted operations in a focused geographic area or worked on a very small scale, boosting the required sample size along may not ensure that sufficient numbers of exposed cases are recruited into the study.

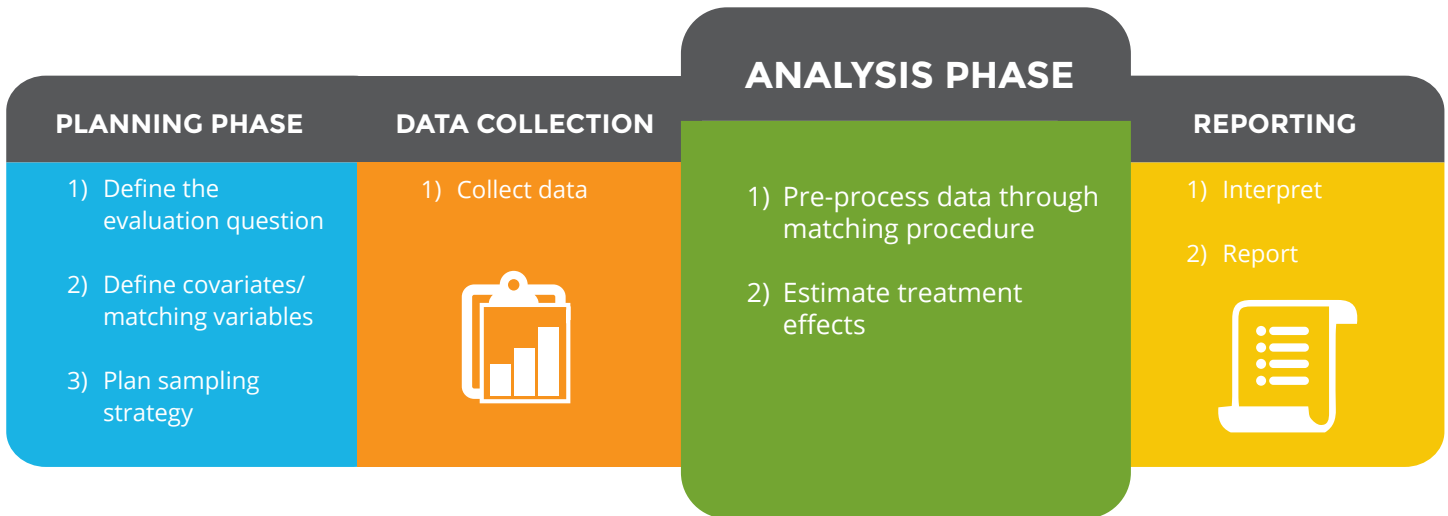
There are no hard-and-fast rules about constructing a sampling strategy when statistical matching is planned for analysis. If you have questions, consider consulting your regional researcher or a statistician for technical support.

TIP



SECTION 3. ANALYSIS PHASE

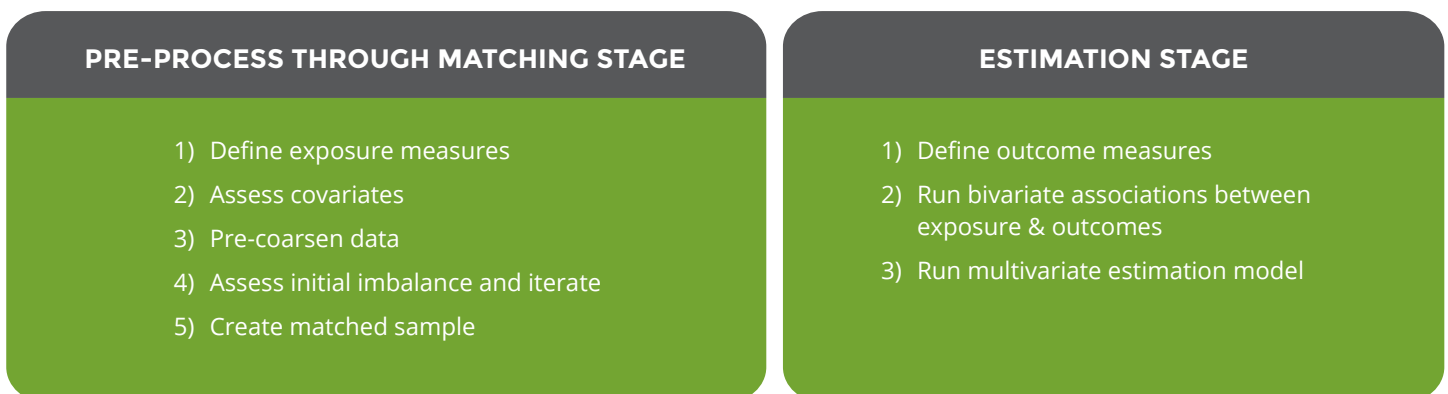
This guide lays out four key phases in conducting a program evaluation: planning, data collection, analysis, reporting. However, in this next section we move to the analysis phase, which is a critical steps for evaluations that use CEM. Guidance on effective management of data collection can be found in PSI's [research standard operating procedures](#) and the [operating guidelines for research](#).



After the evaluation has been designed and data have been collected, the research team moves into the analysis phase. This section is structured to help the research team, and specifically the data analyst working with the data, to understand what decisions need to be made to successfully analyze program evaluation data using coarsened exact matching in a way that allows the research team to explain how it arrived at its conclusions and what they mean for the program.

CEM is conducted in two separate stages: the first stage is the matching stage and the second stage is the estimation stage. After matching and estimation, the research team should re-convene with the program team to interpret results. It is important to document all the different programming steps undertaken in a `do.file`, as many steps in the CEM are iterative. Keeping a complete record of the decision making process will be helpful.

The first two stages of the CEM analysis process can be broken down into the following steps:



The matching stage consists of defining the measures of exposure(s) established in the design phase, selecting the covariates of interest to match on, and then matching samples along these covariates. If exposure is measured in more than one way, then different datasets are analyzed for each exposure definition. Outcome measures are not used during the matching stage. The relationship between the outcome and the exposure is only specified during the estimation stage (Yanovitsky, Zanutto & Hornik 2005).

At the end of the matching stage, the analyst should have two exposure groups (those exposed to the program, and those unexposed) and these should be statistically equivalent in terms of the distribution of the covariates that were used for matching. Cases that remained unmatched because they were not supported by exposed or unexposed counterparts are removed from the analysis prior to the estimation stage. In this way, the analyst can enter the estimation stage and examine comparisons between the groups, which are supported by the data (Iacus, King & Porro 2011; Blackwell et al. 2009).

The following section outlines the steps needed to conduct both the matching and the estimation stages of CEM using data collected as part of a HIV risk-reduction program in Thailand (**Box 9**). The case study here was developed in Stata. Additional resources on using CEM in SPSS can be found in **Appendix 3**.

BOX 9 | CASE STUDY

Sisters HIV prevention for transgender women in Pattaya, Thailand

Since 2004 Population Services International (PSI) has implemented the Sisters program to prevent HIV among transgender women in the city of Pattaya in eastern Thailand. Sisters program addresses HIV prevention within the context of providing a safe haven and broader social support to transgender women. Staff, peer educators, and volunteers are all transgender women. The program operates through three service delivery channels 1) A drop-in center (DiC) that provides counseling, social services, and on-site HIV and STI testing; 2) Peer-led interpersonal communication activities in transgender bars, clubs, and parks to reach transgender women with messages on safe sexual behaviors and HTC; and 3) peer-led interpersonal communication activities through home visits to provide psychosocial/emotional support, and information on gender reassignment, hormone therapy, and cosmetic surgery (Berry, Escobar & Pitorak 2012). The evaluation by Pawa et al. (2013) was published [here](#).

Table 2 below presents a listing of all of the variables used in the model analysis presented for this example from Thailand.

TABLE 2: VARIABLES USED IN THAILAND CEM EXAMPLE

VARIABLE NAME	VARIABLE DESCRIPTION
ltsx_con	Condom use at last sex
ccu_cm	Consistent condom use, commercial ptnrs
ccu_cas	Consistent condom use, casual ptnrs
ccu_reg	Consistent condom use, regular ptnrs
clb_cm	Condom/lube use, commercial ptnrs
clb_cas	Condom/lube use, casual ptnrs
clb_reg	Condom/lube use, regular ptnrs
htc_6	HIV testing, 6 mos
ss_any	Any Sisters service
dic	Drop-in center
outreach	Contact with outreach worker
hv	Received a home visit from peer educator
age	Age
edu	Education (primary, secondary, high school diploma and above)
dur_pat	Duration of residence in Pattaya
incm	Monthly income (baht)
tgno	# transgender friends
emp_entv	Work in entertainment venue
sxwk	Involved in sex work
pt3mos	# sex partners
sx_role	Sexual role (receptive only, penetrative only, receptive and penetrative)
sx_alcdr	Use drugs/alcohol before sex

THE MATCHING STAGE

PRE-PROCESS THROUGH MATCHING STAGE

- 1) Define exposure measures
- 2) Assess covariates
- 3) Pre-coarsen data
- 4) Assess initial imbalance and iterate
- 5) Create matched sample

ESTIMATION STAGE

- 1) Define outcome measures
- 2) Run bivariate associations between exposure & outcomes
- 3) Run multivariate estimation model

STEP 1: DEFINE EXPOSURE MEASURES

Exposure measures should be defined in conjunction with program teams as outlined in section 2 of this guide. As previously discussed, any exposure measures used by the evaluation should be binary (0/1) variables with exposed individuals coded as 1 and unexposed individuals coded as 0.

Case study: Defining exposure

Exposure to the Sisters program was measured during the matching stage as receipt of any Sisters services in the past 12 months (variable `a29d`). Other, more specific measures of exposure were available (variables `dic`, `outreach` and `hv`). Matching on exposure to any of the Sisters' exposure channels allowed the evaluation to incorporate the complete set of possibilities for engaging with the program into the matching process. The effects of specific exposure measures could then be considered as part of the estimation stage.

```
tab1 ss_any
```

participate with any sisters in the past 12 months	Freq.	Percent	Cum.
no	75	24.35	24.35
yes	233	75.65	100.00
Total	308	100.00	

STEP 2: SELECT COVARIATES

Remember that the CEM procedure aims to match exposed and unexposed cases by finding exact matches for grouped (i.e. coarsened) values of covariates. Matching covariates can be identified theoretically, based on an understanding of a correlation between covariates and program participation. It may be helpful to narrow down the potential covariates to consider in the dataset being analyzed by running a correlation matrix between the exposure variable and possible covariates, where highly correlated relationships would indicate a potential variable for matching. Covariates can be binary, categorical or continuous variables. As described earlier, covariates are factors that are thought to influence selection into exposure to the program. Remember that variables that are post-treatment (i.e., that are potentially affected by the exposure of interest) should not be included in the set of matching covariates (Stuart 2010). Covariates should be selected primarily on theoretical grounds, because there is a plausible explanation that these covariates influence an individual's probability of being exposed to the program. The research team should very deliberate about what covariates to match on. Particularly with small samples, matching along a large set of covariates makes it difficult to find appropriate matches.

When considering appropriate matching covariates, assess correlations and associations with exposure with all possible covariates. This process helps identify covariates that are important to include in the matching procedure and also assesses the initial imbalance in the covariates. Understanding how variables are distributed will also provide the data analyst with some clues about how the variable can be coarsened during the matching process. This imbalance can be assessed by running cross-tabulations for categorical variables, with exposure as the dependent variable and each covariate as the independent variable. For continuous variables, t-tests for differences in means can be conducted.

Case study: Selecting covariates and assessing initial imbalance in confounders

Five variables were initially chosen as potential matching covariates for the Sisters program. These included: 1) length of residence in Pattaya; 2) average monthly income in the past 12 months; 3) self-reported occupation as a sex worker; 4) number of transgender friends; and 5) whether the respondents worked in an entertainment venue. The selection of these covariates focused on transgender women's involvement in commercial sex activities, which would influence likelihood of being reached by an outreach worker, plus factors reflecting likelihood of participating in the transgender community in Pattaya, which would influence opportunities to attend the drop-in center.

The association between covariates and exposure was assessed with tabulations and chi-square tests for categorical variables. A sample output between the matching variable occupation is a sex worker, and exposure to the program is:

```
tab sxwk ss_any, row col chi2
```

occupation is sexworker	participate with any sisters in the past 12 months		Total
	no	yes	
no	4 40.00 5.33	6 60.00 2.58	10 100.00 3.25
yes	71 23.83 94.67	227 76.17 97.42	298 100.00 96.75
Total	75 24.35 100.00	233 75.65 100.00	308 100.00 100.00

Two sample t-tests comparing means can be used for continuous variables. An example is the duration of residence in Pattaya.

```
mean dur_pat, over(ss_any) test [no]=[yes]
```

Mean estimation Number of obs = **308**

no: ss_any = no
yes: ss_any = yes

Over	Mean	Std. Err.	[95% Conf. Interval]	
dur_pat				
no	2.087778	.5166567	1.071141	3.104414
yes	2.756795	.2447022	2.27529	3.238301

```
. test [no]=[yes]
```

```
( 1) [dur_pat]no - [dur_pat]yes = 0
```

```
F( 1, 307) = 1.37  
Prob > F = 0.2428
```

Exposed had longer duration but not statistically significant

When using this process, the results show significant associations between exposure to the program and: 1) working in entertainment venues; and 2) monthly income. This indicates that the matching process should definitely include these variables as covariates. All other variables should also be included, as theoretically they are also associated with the outcome.

When examining the distributions of the covariates, it is important to flag those covariates that have very small cell sizes or highly unequal distributions among exposed and unexposed cases. For small cell sizes, it may be important to determine whether inclusion of the covariate eliminates a large number of cases that cannot be matched. This can be tested and examined further on during the analysis. For an unequal distribution, it may be important to pre-coarsen the covariate, particularly if there are empty or near-empty cells in a contingency table for either exposed or unexposed individuals.

TIP



STEP 3: PRECOARSENING THE DATA

Stata's CEM command has a functionality in which the data can be coarsened by the program into fixed bin sizes (i.e., the coarsening is automated); however, there are cases where pre-coarsening the data manually, that is, re-categorizing or re-coding a variable, may be advisable. This can be done when there are natural breaks in continuous variables that have a logical grouping (for example, changing years of education into level of education) or when the categories in a categorical variable are not ordinal (for example, scales with "neutral" or "no opinion" codes). The decision as to whether to pre-coarsen will likely have to do with the types of covariates that are considered in the previous step and their distributions.

If few cases among exposed or unexposed are found in a few categories, it may be worthwhile to pre-coarsen the data. In this way, the bins in which exposed and unexposed cases are matched have a logical and meaningful division, rather than those created by the CEM algorithm. It is worth noting that creating new categories for a variable changes what data will be coarsened by the CEM algorithm. As such, pre-coarsening through the creation of new categories should be done when the research team has a clear reason, based on theory or data, to do so.

It is also important for the research team to decide what variables should not be coarsened. Within the CEM command, a user can specify with a #0 what variables must be kept with their original category responses. This might be important if, for example, there is a categorical variable for city and/or region, and the analyst would like to ensure that all matches are drawn from the same city or region.

See **Box 4** in Section 1 for examples of coarsened variables. For the Sisters' CEM evaluation, that data were not pre-coarsened; however, it would have been possible to look at different income categories based on population quintiles.

STEP 4: ASSESS THE INITIAL IMBALANCE OF COVARIATES

As discussed earlier, imbalance is a measure of how covariates differ between the exposed and unexposed groups. In the CEM algorithm, imbalance is measured by the **L1 statistic**—a summary measure of global imbalance calculated by comparing the differences between all the covariates at once. When perfect balance between exposed and unexposed groups is achieved, L1 is equal to zero. When there is perfect imbalance, L1 is equal to 1. The L1 statistic, therefore, takes on values that range from 0 to 1, where the higher the number the less balanced the exposed and unexposed groups are. By matching, our objective is to make the L1 statistic smaller (i.e., make the groups more similar to one another and therefore the comparisons more valid). For PSI purposes, an L1 statistic of 0.2 or lower is considered acceptable.

To assess how matching is improving the balance between covariates, it is important to first determine the initial imbalance between covariates. If matching is successful, the L1 statistic should become smaller in the matched samples as compared to the original sample.

To determine the original imbalance, Stata uses a command called `imbalance` (abbreviated `imb`) to calculate the initial imbalance between confounders. The syntax for this command is:

```
imbalance covariates, treatment(treated)
```

Where **covariates** is the list of covariates separated by spaces, and `treated` indicates the binary treatment variable (where 1 indicates that the individual was exposed and 0 indicates that the individual is unexposed. This is the exposure variable previously identified to use in matching.

Case study: Assessing initial imbalance

To determine the initial imbalance of covariates for the Sisters program, the following syntax was used:

```
imb dur_pat incm sxwk tgnv emp_entv, treatment (ss_any)
```

The syntax yields the following output:

Multivariate L1 distance: .57144492

Univariate imbalance:

	L1	mean	min	25%	50%	75%	max
dur_pat	.25431	.66902	0	.25	.83333	1.8333	-7
incm	.22524	6295.4	-1500	5000	5000	5000	40000
sxwk	.02758	.02758	0	0	0	0	0
tgnv	.07285	4.1276	0	1	3	2	270
emp_entv	.17894	.17894	0	0	1	0	0

The L1 statistic of 0.5714 indicates that there is a high level imbalance between the two groups, with an L1 greater than the threshold of 0.2. This measure takes into account the imbalance of all the matching variables and their interactions.

The first column in the output table that is labeled L1 is the L1 statistic for each one of the variables (which does not include interactions). The second column in the table labeled “mean,” reports the difference in means between exposed and unexposed cases. The remaining columns in the table report the difference in the quantiles of the distributions of the two groups for the 0th (min), 25th, 50th, 75th, and 100th (max) percentiles for each variable. In this case, we see a high imbalance in the mean values of variable `dur_pat` and `incm` and also high imbalance in the quantiles of variable `a3`, and in the max quantile of all three continuous variables. This points to a need to closely examine the distribution of the upper bounds of the continuous variables in the exposed and unexposed groups.

STEP 5: CREATING A MATCHED SAMPLE

The next step in the process is to execute the initial matching command for CEM. The syntax for CEM in Stata is:

```
cem varname1 [(cutpoints1)] [varname2 [(cutpoints2)]] ... ,treatment(varname)
```

Where `varname#` immediately following the `cem` command corresponds to a covariate selected for matching and `treatment(varname)` corresponds to the exposure variable. Cutpoints placed in parenthesis following the matching variable are an option that can be specified by the analyst to coarsen the variable with user-specified cutpoints. When cutpoints are not specified, Stata uses a default binning algorithm, essentially creating a set of computer-generated categories or bins of the matching variable.

There are two analytic goals to keep in mind while the analyst creates a matched sample:

- 1) Minimize imbalance between exposed and unexposed group
- 2) Maximize sample size

The analyst’s task is to establish a matching solution with the identified exposure variable of interest and the set of prioritized covariates to use in matching. To create this matching solution, the analyst must consider which covariates to include in the matching algorithm and how they are coarsened, such that minimal imbalance is present between exposed and unexposed groups (i.e. there is a substantial area of common support) without sacrificing sample size.

An advantage of CEM is the ability to specify cutpoints (Blackwell et al 2009). To specify cutpoints for a variable, place a list of numbers in parentheses after the variable's name. The user can also ask the command to create a number of equally spaced cutpoints (for example, if you want to specify deciles of an age variable you would place "#10" in the parentheses). If the user does not want the variable coarsened, specify (#0) after the variable.

Specifying cutpoints is an iterative process. The analyst must try to ensure that the matches are as close as possible (i.e., that the L1 statistic is as close to zero as possible), but then also ensure that a sufficient sample (particularly when samples are very small to begin with) is retained in order to conduct the estimation phase of the analysis. There are no hard and fast rules for what constitutes good matching and what constitutes an unacceptable sacrifice of sampled cases. The goal is to get to a low L1, signifying a balanced match, without losing too many cases, because this influences the representativeness of the results.

It is possible that the analyst will need to try several different specifications of the cutpoints to ensure that this balance between L1 and sample size is achieved.

It is usually helpful to run the cem command with the Stata-specified cutpoint algorithm to determine what occurs with the L1 and matched sample size. To do so, the syntax used is:

```
cem matching_variables, treatment(exposure_variable) sh
```

The sh on the end of the syntax indicates that Stata should show the cutpoints that are used for each variable. This syntax should include (#0) after every matching variable that: 1) is dichotomous and thus does not require coarsening; or 2) the analyst that decided that matches should be exact on the specified matching variable (as may be the case with a geographical matching variable, if the analyst would like to match cases within region, city or province.)

Case study: Creating a matched sample using the CEM algorithm

The following syntax was used to create a matched sample for the Sisters program using the cem algorithm. The default algorithm for the program is called the Sturge's algorithm.

```
cem dur_pat incm sxwk(#0) tgn0 emp_entv(#0), treatment (ss_any) sh
```

The first section of the output presents the bins created by the Stata algorithm for each of the matching variables.

Cutpoints:

dur_pat: (sturges)

1

1	.083333358
2	3.518518521
3	6.953703706
4	10.38888889
5	13.82407408
6	17.25925926
7	20.69444445
8	24.12962963
9	27.56481482
10	31

The lower part of the output presents the CEM results.

Matching Summary:

Number of strata: 64
Number of matched strata: 16

	0	1
All	75	233
Matched	70	167
Unmatched	5	66

Only 16 of the 64 strata had matched cases

Multivariate L1 distance: .64285714

High imbalance

Univariate imbalance:

	L1	mean	min	25%	50%	75%	max
dur_pat	.21953	.34174	0	0	.16667	1	2
incm	.16895	976.48	0	0	0	0	0
sxwk	1.7e-18	0	0	0	0	0	0
tgno	.08169	.90024	0	1	1	2	0
emp_entv	3.3e-16	3.3e-16	0	0	0	0	0

The matching summary at the top of the output presents the combinations of coarsened response choices for all matching variables specified in the command. In this case, there are 64 different combinations, among them only 16 have at least one treated individual and one untreated individual (i.e., matched strata). Below that, we see that 70/75 of unexposed cases and 167/233 exposed cases were matched. The multivariate L1 distance shows a high imbalance in the matches, with a value of 0.6429. Examining the individual variable L1 statistics shows that the continuous variables, particularly `dur_pat` and `incm`, show a high L1. The variable `incm`, which measure income in the past 6 months, also presents a very high mean difference between exposed and unexposed individuals.

In cases where the Stata algorithm for binning does not produce an adequate L1 statistic (less than 0.2), the analyst must determine how to achieve a better matching solution. There are two options available, 1) to eliminate covariates; and 2) to coarsen the variables differently, to improve the match.

ELIMINATE UNNECESSARY COVARIATES

To improve matches, the analyst must first re-examine the covariates and identify any matching covariates that: 1) may have a weak conceptual link to exposure, and 2) are not significantly associated with exposure. These covariates can be eliminated from the matching procedure. Once the set of final covariates is established, re-run the `cem` matching procedure and assess the L1 statistic again.

CUSTOMIZING COARSENING TO IMPROVE THE L1 STATISTIC

If after eliminating unnecessary covariates there is still a high L1 statistic, the analyst must manually re-coarsen the variables, one at a time, to try to improve the quality of the matches.

Binning for variables should be customized according to the distribution of the variable. A first step is to look at the distribution of the variables that have high individual L1 statistics by exposure group. Here, the analyst can confirm whether the distributions between exposed groups are skewed and what values of the matching variable may lack matches for either group. Bins for the variable can be created as to ensure the closest matches possible without losing a large proportion of the sample. CEM will be able to maintain a higher number of cases with fewer, wider bins.

This process can then be repeated for each of the variables that require coarsening. It is important to recognize that the global L1 statistics includes the imbalance in all the variables (jointly) and their interactions. While the individual L1 may be lowered by this process, it is the global imbalance (the multivariate L1 statistic) that should be considered when assessing the quality of a match.

Case study: Manually coarsening data

In the Sisters example, variable `dur_pat`, corresponding to the length of residence in Pattaya, showed a high individual L1 statistic, and a mean difference between the exposed and unexposed groups in the highest quantiles. As a first step, it is critical to look at the actual distribution of the variable between the two groups. This can be done by both looking at the descriptive statistics by exposure category, and by looking at the tabulation of the continuous variable by exposure.

```
sort ss_any by ss_any: summarize dur_pat, detail
```

→ `ss_any = no`

length of residence in chonburi			
Percentiles	Smallest		
1% .0833333	.0833333		
5% .1666667	.0833333		
10% .1666667	.1666667	Obs	75
25% .3333333	.1666667	Sum of Wgt.	75
50% .6666667		Mean	2.087778
		Std. Dev.	4.474378
75% 1.166667	8	Variance	20.02006
90% 7	8	Skewness	4.659608
95% 8	20	Kurtosis	27.59457
99% 31	31		

→ `ss_any = yes`

length of residence in chonburi			
Percentiles	Smallest		
1% .0833333	.0833333		
5% .25	.0833333		
10% .25	.0833333	Obs	233
25% .5833333	.0833333	Sum of Wgt.	233
50% 1.5		Mean	2.756795
		Std. Dev.	3.735218
75% 3	20	Variance	13.95185
90% 6	21	Skewness	3.21526
95% 10	23	Kurtosis	15.56685
99% 21	24		

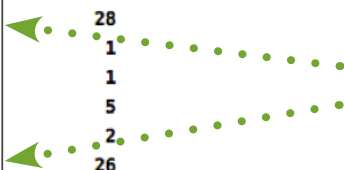
Differences
between treated
and untreated

This distribution shows that the values at the 25th, 50th, and 75th percentiles are higher among the exposed group than the unexposed group. The overall mean value is also higher among the exposed group, and the variance is higher among the unexposed group.

A tabulation of the matching variable by exposure category shows that there are many values for which there are no matches among the unexposed. The bins created in the CEM command should thus maximize the proximity of the observations where values overlap, and create wider bins for the ranges of values that require a more coarsened match.

```
tab dur_pat ss_any
```

length of residence in chonburi	participate with any sisters in the past 12 months		Total
	no	yes	
.0833333	2	4	6
.1666667	7	7	14
.25	7	19	26
.3333333	5	6	11
.4166667	6	7	13
.5	2	8	10
.5833333	5	10	15
.6666667	4	9	13
.75	1	4	5
.8333333	2	1	3
1	14	33	47
1.166667	2	4	6
1.25	0	1	1
1.333333	0	1	1
1.416667	0	2	2
1.5	0	3	3
1.583333	0	1	1
1.75	0	1	1
2	3	25	28
2.25	0	1	1
2.416667	0	1	1
2.5	2	3	5
2.666667	0	2	2
3	3	23	26



Few matches between exposed and unexposed

This tabulation also shows that there are certain values that fall out of range. Creating finer bins in areas of great overlap and wider bins in areas of less overlap can help improve matches for the majority of the sample and retain cases where there is less overlap. Manually coarsening covariates to improve the L1 is the same as precoarsening data, only it is done after an initial match has been tried.

DROPPING PROBLEMATIC VARIABLES AS A LAST RESORT

The process of manually coarsening variables can be repeated for every covariate in the CEM procedure. If this process yields a high L1 or a large loss of sample size, the analyst may choose to drop problematic variables and adjust for them in the estimation stage. This should be avoided, if possible, among those variables that are conceptually very important to exposure to the program.

Once the final matching solution is chosen and the most efficient coarsening cutpoints are established, run the final CEM model and continue to the estimation stage.

TIP



Documentation of decisions taken during data analysis is always a best practice. Because CEM may require several iterations before achieving a successful matching solution, it is essential that all decision points in the analysis process be thoroughly documented in the do-file.

Case study: Running the final estimation model

The final set of the Sisters' CEM matching variables required the removal of variable `dur_pat`, the length of residence in Pattaya, to ensure a small enough L1 statistic and conservation of the sample. For the `dur_pat` variable, several binning options were attempted, including a specification of only two bins. In this case, the L1 statistic still remained unacceptable. Since the variable was not significantly associated with exposure, it was eliminated from the final `cem` set of matching covariates.

The final `cem` procedure was run as follows:

```
cem incm(5000 10000 15000 20000 25000) sxwk(#0) tgn0 (5 10 15 20) emp_entv(#0),
treatment (ss_any)
```

Number of strata: 57

Number of matched strata: 29

	0	1
All	75	233
Matched	64	200
Unmatched	11	33

Multivariate L1 distance: .28666667

Univariate imbalance:

	L1	mean	min	25%	50%	75%	max
incm	.14083	2080.4	0	0	0	0	40000
sxwk	4.3e-19	0	0	0	0	0	0
tgn0	.0375	3.0396	0	0	-2	0	270
emp_entv	1.4e-16	-2.2e-16	0	0	0	0	0

Improvements in matched strata and L1 compared to initial match

Here the procedure presents an L1 statistic of 0.28. While not ideal (not under 0.2), the reduction in the L1 from the original imbalance of 0.54 is sizable. With this matching specification, eliminating the problematic variable `dur_pat`, we were able to retain a large proportion of the sample for analysis, where otherwise nearly half of the sample would have been eliminated. This is a reminder that a good understanding of the evaluation question and of the data being analyzed should drive data analysis decisions, rather than making decisions based on specified cut-offs.

The CEM procedure adds three new variables to your original dataset every time the CEM procedure runs. The new variables are the following:

- `cem_strata`: The stratum to which each case was assigned by CEM.
- `cem_matched`: Indicates whether or not this observation was matched. This variable is coded 0 (not matched) and 1 (matched).
- `cem_weights`: Provides weights for each case based on the most recent matching solution. The weight is specific to stratum to which the case has been assigned and representative of the proportion of all members present in the stratum. Unmatched cases have a weight of 0.

These variables will be over-written each time a new CEM command is processed. Make sure that the correct CEM variables are saved to the dataset before proceeding to estimation. Save each matching procedure for each treatment separately, and always match from the original dataset.

PRE-PROCESS THROUGH MATCHING STAGE

- 1) Define exposure measures
- 2) Assess covariates
- 3) Pre-coarsen data
- 4) Assess initial imbalance and iterate
- 5) Create matched sample

ESTIMATION STAGE

- 1) Define outcome measures
- 2) Run bivariate associations between exposure & outcomes
- 3) Run multivariate estimation model

THE ESTIMATION STAGE

In the estimation stage, the data analyst will formally test a hypothesis that program exposure is associated the evaluation's outcomes of interest (Privitera 2012). Specifically, the analyst will be looking for evidence to disprove the null hypothesis that there is no difference between exposed and unexposed groups for any outcome of interest.

CEM was developed in part was to avoid **model dependence**, where findings are conditional on how variables are specified during modeling (Blackwell et al. 2009). Following the logic of statistical matching, if the analyst has achieved a quality match with minimal imbalance, any remaining difference between exposed and unexposed cases should solely be attributable to the program cases were exposed. As such, the estimation stage should use simple models that maximize the previously conducted matching to reduce bias in the estimation of program effects. Deriving an unbiased estimate of the effects of the program is the goal, rather than developing a model with the best model fit (Harrell 2002).

MODELING STRATEGY

The goal of this type of evaluation is to estimate average treatment effects –that is, the strength of the relationship between program participation and outcomes of interest. Once we’ve accounted for effects of selection bias and risks of confounding through statistical matching, we need to develop an estimation model to estimate program effects. Because of the power of matching, the modeling strategy should be minimal and straightforward, without extensive adjustment for other factors.

Regression models are recommended for estimating program effects. Statistical regression models – whether for continuous outcomes or for categorical outcomes – are a highly powerful tool that estimate the nature of the relationship between two variables, thus giving an estimate of how large a difference program participation makes in encouraging people to adopt or practice the outcome of interest, holding all other factors that could influence that outcome constant.

At PSI, most behavioral outcomes of interest are dichotomous in nature – that is, they are yes/no variables. We therefore use logistic regression in the case study presented here. For more information on the properties of logistic regression and fitting logistic regression models in Stata, see: http://www.ats.ucla.edu/stat/stata/topics/logistic_regression.htm.

For continuous outcomes, ordinary least squares (linear) regression is recommended. OLS regression provides estimates on the magnitude of the difference between exposed and unexposed groups (unlike Analysis of Variance, which provides adjusted means for the two groups). Materials on OLS regression are available at: <http://www.ats.ucla.edu/stat/stata/topics/regression.htm>.

Before beginning a modeling strategy, it is worth reviewing the study’s sampling strategy to address how it will be accounted for during analysis. All of the analysis requirements for handling survey data collected from stratified probability, RDS, or TLS sampling strategies will apply when analyzing data to which CEM has been applied (Amon et al. 2000; United Nations 2008; Heckathorn 2012). Weights, clustering, and stratification variables should be available in the dataset and applied accordingly. For simplicity of focusing on the matching process, the examples provided here do not delve into survey data analysis in detail.

Because CEM trims the number of cases used in the analytic sample to only those cases which can be matched, the analyst should be mindful about the representativeness of the subsequent sample.

Make sure that sampling weights are factored into the estimation process after matching. A new weight variable should be generated that factors in both sampling weights (`wt`) and CEM weights (`cem_weights`):

```
gen new_wt=wt*cem_weights
```

Use this new weight variable (`new_wt`) in subsequent analyses.

TIP



STEP 1: DEFINE OUTCOME MEASURES

After completing matching, the analyst can return to the evaluation's outcomes of interest, which should have been identified and defined during the evaluation's planning phase. As discussed previously, outcomes should follow the correct technical specifications of any required donor indicators.

Outcomes should be sorted by whether they are continuous or categorical variables, as modeling techniques will differ according to the functional form of the variable being modeled.

Be mindful of whether any of the outcomes of interest may lay on a theory of change that leads to another outcome. In other words, think about whether any of the outcomes identified for the evaluation might theoretically cause any of the other outcomes to occur. For example, condom use self-efficacy is often identified as a factor that contributes to consistent condom use (Schiavo 2007). If this is the case, the research team should be aware of any evidence of relationships between these outcomes, before developing a modeling strategy.

Standard univariate descriptive statistics should be performed for each outcome variable before proceeding to the next step in order to understand how the variables are distributed.

```
tab htc_6
```

received hiv testing in the past 6 months	Freq.	Percent	Cum.
no	142	46.10	46.10
yes	166	53.90	100.00
Total	308	100.00	

```
tab htc_6 [iweight = cem_weights]
```

received hiv testing in the past 6 months	Freq.	Percent	Cum.
no	118.013333	44.70	44.70
yes	145.986667	55.30	100.00
Total	264	100.00	

STEP 2: RUN BIVARIATE ASSOCIATIONS BETWEEN EXPOSURE AND OUTCOMES

Start first by assessing bivariate relationships between the exposure variable and outcome of interest. In the example below, the variable `dic` is an exposure variable indicating use of a drop-in center. Assessment of relationships between exposure and outcome should be conducted in both the full and the matched sample, to understand how application of the matching procedure changes the inference that can be made about this relationship.

If both the exposure and the outcome variable are categorical, this can be done through a chi-squared test of independence.

```
tab2 htc_6 dic, row col chi2
```

received hiv testing in the past 6 months	dic		Total
	0	1	
no	106	36	142
	74.65	25.35	100.00
	56.99	29.51	46.10
yes	80	86	166
	48.19	51.81	100.00
	43.01	70.49	53.90
Total	186	122	308
	60.39	39.61	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 22.3920 Pr = 0.000

In Stata, the matching procedure is applied through a weight option in the command line (see **Appendix 3** for guidance on how to apply CEM weights in SPSS). However, the particular weight option to apply CEM weights is not available for Stata's tabulation commands for chi-square tests. We will therefore proceed to bivariate logistic regression models, since these provide information equivalent to chi-square tests and useful as a baseline before moving to multivariate models.

BOX 10 | WEIGHTS IN STATA

Stata has a variety of weighting options that can be applied in its commands. Stata documentation on weights provides useful documentation on the types of weights available and when they should be applied. Documentation on `iweights` is provided [here](#).

`iweights`, or importance weights, are weights that indicate the “importance” of the observation in some vague sense. `iweights` have no formal statistical definition; any command that supports `iweights` will define exactly how they are treated. Usually, they are intended for use by programmers who want to produce a certain computation.

```
logistic htc_6 dic
```

```
Logistic regression      Number of obs =      308
                        LR chi2(1)  =      22.88
                        Prob > chi2  =      0.0000
Log likelihood = -201.11222 Pseudo R2 =      0.0538
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
dic	3.165278	.7839393	4.65	0.000	1.948032	5.143131
_cons	.754717	.1117745	-1.90	0.057	.5645731	1.0089

```
logistic htc_6 dic [iweight=cem_weights]
```

```
Logistic regression      Number of obs =      264
                        LR chi2(1)  =      13.23
                        Prob > chi2  =      0.0003
Log likelihood = -174.88864 Pseudo R2 =      0.0365
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
dic	2.585739	.6899442	3.56	0.000	1.532715	4.362225
_cons	.8732761	.1371146	-0.86	0.388	.6419526	1.187956

Relation
strong in
models

STEP 3: RUN MULTIVARIATE ESTIMATION MODEL

Building a multivariate model requires considerable thought about the sequencing of variables to add into a model. The aim is to develop an unbiased estimate of the difference in outcomes attributable to the program or specific program channels. Decision-making about modeling strategy will require discretion by the data analyst, based on a solid understanding of how the program has been implemented. As such, this section can provide guidance, but ultimately, the quality of the model and approach to deriving program estimates is in the hands of the data analyst. Remember to document all decisions taken in the do file.

Begin with the exposure measure used in the matching procedure. As with the bivariate models, multivariate models should be estimated in both the full and the matched samples to assess how application of the matching procedure changes the estimation.

CONSIDER ADJUSTING FOR COVARIATES USED IN MATCHING

Next, consider including the covariates matched on in the model, as a sensitivity analysis to understand the effects of the matching. If the effect size of the exposure variable changes substantially in magnitude, direction of effect, and/or statistical significance, it may indicate that the matching solution requires further refinement, either by changing how variables were coarsened or removing variables from the matching procedure.

```
logistic htc_6 dic tgn0
```

```
Logistic regression      Number of obs =      308
                        LR chi2(2) =      27.54
                        Prob > chi2 =      0.0000
Log likelihood = -198.78338 Pseudo R2 =      0.0648
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
dic	3.074416	.7675593	4.50	0.000	1.88474 5.015033
tgn0	1.015417	.0086822	1.79	0.074	.9985421 1.032577
_cons	.6211508	.1126293	-2.63	0.009	.4353657 .8862167

OR is similar but
statistical significance
lessened in matched

```
logistic htc_6 dic tgn0 [iweight=cem_weights]
```

```
Logistic regression      Number of obs =      264
                        LR chi2(2) =      16.25
                        Prob > chi2 =      0.0003
Log likelihood = -173.37992 Pseudo R2 =      0.0448
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
dic	2.548722	.6837719	3.49	0.000	1.506475 4.312042
tgn0	1.012023	.0084587	1.43	0.153	.9955794 1.028739
_cons	.7480003	.1402774	-1.55	0.122	.5179292 1.080272

ADDRESS ADDITIONAL REMAINING COVARIATES

Add other remaining covariates, one-by-one. Think carefully about which covariates to adjust for, as most of these decisions should have been made during the matching process. The two main categories of covariates to consider for adjustment should be 1) conceptually important covariates removed from matching and 2) covariates related to the study's sampling and the population it seeks to represent.

Covariates removed from matching because of challenges in reducing imbalance – likely due to small cell sizes or a skewed distribution between exposed and unexposed cases – may still be considered for adjustment. However, there should be a strong theoretical case for including these variables.

Case study: Examining age covariate

In the example below, the age covariate is considered for inclusion in the model examining the effect of the drop in center, both with and without weights. While both models show that age is not a statistically significant covariate, one can see that application of the CEM weights attenuates the results of the relationship between the outcome and exposure.

```
logistic htc_6 dic age
```

```
Logistic regression          Number of obs   =      308
                             LR chi2(2)           =      22.94
                             Prob > chi2          =      0.0000
Log likelihood = -201.08299   Pseudo R2       =      0.0540
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
dic	3.168604	.7850046	4.66	0.000	1.949788	5.149305
age	1.007123	.029566	0.24	0.809	.9508101	1.066771
_cons	.6337591	.4675618	-0.62	0.536	.1492595	2.690954

Consider how the odds ratio for DiC changes when CEM weights are applied to the matched sample.

```
logistic htc_6 dic age [iweight=cem_weights]
```

```
Logistic regression          Number of obs   =      264
                             LR chi2(2)           =      14.13
                             Prob > chi2          =      0.0009
Log likelihood = -174.44114   Pseudo R2       =      0.0389
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
dic	2.585706	.6910665	3.55	0.000	1.531381	4.36591
age	.971467	.0297904	-0.94	0.345	.9147989	1.031646
_cons	1.784621	1.378725	0.75	0.453	.3925928	8.112405

The odds ratio for DiC is attenuated but still statistically significant. However, the odds ratio for age is small and not statistically significant. Compare the model above to a model applying CEM weights that does not adjust for age.

```
logistic htc_6 dic [iweight=cem_weights]
```

```
Logistic regression          Number of obs   =      264
                             LR chi2(1)           =      13.23
                             Prob > chi2          =      0.0003
Log likelihood = -174.88864   Pseudo R2       =      0.0365
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
dic	2.585739	.6899442	3.56	0.000	1.532715	4.362225
_cons	.8732761	.1371146	-0.86	0.388	.6419526	1.187956

In going back to the unadjusted model, it becomes clear that adjusting for age did not substantially influence estimation of the odds ratio for DiC. Rather, application of the CEM, effectively accounting for all of the variables in the matching procedure, was more influential.

One common covariate included in the matching stage is location; however, it is possible to make a case for not matching on covariates related to the geography of the study's sampling. Study location, for example, city or country of residence, may be important for influencing an individual's probability of having been exposed to the program, but this variable also influences researchers' understanding of the study population and the representativeness of the study. Matching on these variables may influence this, if cases are dropped differentially across study locations and if sampling weights are not accounted for. The evaluation may also wish to isolate the effects of the program in specific locations. If this is the case, it may be more appropriate to adjust for study location during estimation, rather than match on these factors during matching.

ADDRESS RELATIONSHIPS BETWEEN EXPOSURE VARIABLES

Consider adding other appropriate exposure variables to test independent and additive effects of channels. Only add other exposure variables for which the matching procedure is appropriate.

In the Thailand example, the research team was interested in whether exposure to any component of Sisters influenced the outcomes of interest. Transgender women who reported any exposure to Sisters could have been exposed to any combination of contacts with outreach workers, visits to the drop-in center and/or home visits by Sisters staff. The Sisters team wanted to know which of these specific components might be most effective at achieving condom and lubricant use or HIV testing. This analysis tested each program component to assess their independent effects, having matched on any program exposure.

In the models below, one can see that when analyzed separately, both the drop in center and outreach appear to have a statistically significant relationship with the outcome; however, when both the drop in center and outreach are included in the model, the outreach variable is not statistically significant on its own.

```
logistic htc_6 dic [iweight= cem weights]
```

```
Logistic regression      Number of obs   =      264
                        LR chi2(1)      =      13.23
                        Prob > chi2     =      0.0003
Log likelihood = -174.88864   Pseudo R2      =      0.0365
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
dic	2.585739	.6899442	3.56	0.000	1.532715 4.362225
_cons	.8732761	.1371146	-0.86	0.388	.6419526 1.187956

Statistically significant

```
logistic htc_6 outreach [iweight = cem_weights]
```

```
Logistic regression      Number of obs   =      264
                        LR chi2(1)      =      5.05
                        Prob > chi2     =      0.0247
Log likelihood = -178.98238   Pseudo R2      =      0.0139
```

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
outreach	1.816953	.4848441	2.24	0.025	1.076974 3.065365
_cons	.8255578	.1809796	-0.87	0.382	.5372126 1.26867

Statistically significant

```
logistic htc_6 outreach [iweight = cem_weights]
```

Logistic regression

Number of obs	=	264
LR chi2(2)	=	15.71
Prob > chi2	=	0.0004
Pseudo R2	=	0.0433

Log likelihood = -173.64859

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
dic	2.39361	.6501442	3.21	0.001	1.405572 4.076185
outreach	1.543275	.4253434	1.57	0.115	.8991724 2.648767
_cons	.6697115	.1553723	-1.73	0.084	.4250206 1.055275

DiC statistically significant;
Outreach not

This analysis could have also investigated whether having been exposed to more than one program component influence condom and lubricant use or HIV testing. For example, it is possible to create exposure variables of outreach plus drop-in center or drop-in center plus home visit. Because of small sample sizes, this was not investigated in detail.

Finally, it is worth considering statistical interactions between specific program components, which would indicate that the program's effects were boosted beyond simply an additive effect of having received two or more components. The analysis should only test for interactions, if each main effect term for exposure is statistically significant (Mitchell & Chen 2005).

```
logit htc_6 i.dic#outreach
```

Logistic regression

Number of obs	=	308
LR chi2(3)	=	29.39
Prob > chi2	=	0.0000
Pseudo R2	=	0.0691

Log likelihood = -197.8568

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
1.dic	4.25	2.119785	2.90	0.004	1.598942 11.29654
1.outreach	2.163636	.6749487	2.47	0.013	1.173554 3.987654
dic#outreach 1 1	.577731	.3342929	-0.95	0.343	.1858636 1.795796
_cons	.4705882	.1164879	-3.05	0.002	.2896925 .7644425

Outreach not statistically significant

Interaction effect

Interaction not statistically significant

```
logit htc_6 i.dic#outreach [iweight = cem_weights]
```

Logistic regression

Number of obs	=	264
LR chi2(3)	=	15.89
Prob > chi2	=	0.0012
Pseudo R2	=	0.0438

Log likelihood = -173.55999

htc_6	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
1.dic	2.89968	1.549183	1.99	0.046	1.017623 8.26253
1.outreach	1.658951	.5403569	1.55	0.120	.8761504 3.14115
dic#outreach 1 1	.7708762	.4783692	-0.42	0.675	.2284385 2.601357
_cons	.6404648	.164106	-1.74	0.082	.3876076 1.058274

Outreach not statistically significant

Interaction effect

Interaction not statistically significant

After reviewing and testing the considerations discussed here, the analyst should interact to re-fit models, while keeping all primary evaluation hypotheses in mind.

Repeat these steps for each outcome of interest. Potential covariates to consider will vary according to the exposure and outcome variables being tested.

SECTION 4. DATA INTERPRETATION AND PRESENTATION

The next stage of the process is to review results, determine how to report them, and assess what they mean with the team implementing the program that has been evaluated.

REPORTING RESULTS

Once all hypotheses related to the evaluation have been tested and models have been finalized, the research team should consider how to report their results. The exact nature of how to report results is left to the discretion of the research team, which should decide how best to convey main findings according to the audiences for evaluation results. This section provides suggestions, best practices and examples. Best practices in data visualization can also help the research team convey the evaluation's key findings (Institute for Digital Research and Education 2014).

1. Present results from the full and the matched samples

For purposes of transparency, it is advised to show results from both the full and the matched samples. Primary findings about program effectiveness should be drawn from the matched sample, but it is advisable to show how the matching process influenced the conclusions that can be drawn about the program. By limiting the sample to the area of common support, the matching procedure may influence the population-representativeness of the findings.

2. Provide descriptive statistics for the study before presenting program effects

Present descriptive statistics for variables used in matching, as well as the exposure and outcome variables. The two tables below show findings presented from the evaluation of the Sisters program, including key characteristics of the sample. A comparison of descriptive statistics in the full and the matched samples can support an assessment of whether the matching process influenced the population representative-ness of estimates.

TABLE 3. POPULATION CHARACTERISTICS BY SAMPLE

SOCIO-DEMOGRAPHIC CHARACTERISTICS	FULL SAMPLE (N=308), %	MATCHED SAMPLE (N=238), %
Age, mean (sd)	24.56 (0.23)	24.55 (0.27)
Education		
Primary school and lower	6.17	7.23
Secondary school	39.29	43.49
High school	43.18	39.21
Diploma and higher	11.36	10.07
Duration of residence in Pattaya		
Years, mean (sd)	2.59 (0.22)	2.65 (0.25)
≤ one year, %	52.92	48.92
> one year, %	47.08	51.08
Occupation		
Employment in evening entertainment venue	60.06	59.16

Freelance sex worker	39.29	42.00
Other employment	8.12	5.55
Monthly income		
0 - 10,000 THB	19.81	15.00
10,001 - 20,000 THB	36.69	36.08
20,001-30,000 THB	25.32	26.48
> 30,000 THB	18.18	22.44
Transgender friends, mean (sd)	14.28 (1.29)	17.29 (1.67)
Sexual behaviors and practices		
Sexual partners in the past 12 months		
Mean number of partners, any type (sd)	41.21 (2.82)	44.01 (3.49)
Commercial partners		
%	93.83	97.90
Mean number of partners (sd)	38.63 (2.84)	39.47 (3.41)
Casual partners		
%	13.96	12.18
Mean number of partners (sd)	8.86 (1.44)	9.43 (1.91)
Regular partners		
%	17.86	13.45
Mean number of partners (sd)	1.15 (0.054)	1.11 (0.06)
Sexual role		
Always practice receptive sex	38.31	41.08
Always practice penetrative sex	0.97	1.84
Practice both receptive and penetrative sex	60.71	57.08
Had sex after using drugs or alcohol in past 3 months	52.27	51.09

TABLE 4. BEHAVIORAL OUTCOMES AND PROGRAM PARTICIPATION, BY SAMPLE

BEHAVIORAL OUTCOMES	FULL SAMPLE (N=308), %	MATCHED SAMPLE (N=238), %
Condom use at last sex with any partner	93.18	92.44
Commercial partners		
<i>Consistent condom use in the past 3 months</i>	85.81	84.77
<i>Consistent condom and water-based lubricant use in past 3 months</i>	75.09	72.72
Casual partners		
<i>Consistent condom use in the past 3 months</i>	72.09	61.75
<i>Consistent condom/water-based lubricant use in the past 3 months</i>	62.79	55.44
Regular partners		
<i>Consistent condom use in the past 3 months</i>	60.00	61.73
<i>Consistent condom/water-based lubricant use in the past 3 months</i>	49.09	36.32
HIV testing		
Received an HIV test in the past 6 months	53.90	54.68
Program participation in the past 12 months		
Received any Sisters service	75.65	72.31
Visited Sisters drop-in center	39.61	39.92
Received Sisters outreach	67.86	64.29
Received Sisters home visit	17.53	16.81

3. Effect sizes should be discussed and interpreted based on the size of estimate, direction of the estimate, and statistical significance

When presenting estimates about the effectiveness of the program itself and making decisions about what results should be highlighted and shared with the program team and the evaluation's stakeholders, consider not just whether the estimate is statistically significant. Consider as well is the magnitude of the effect. Does the size of the estimate make sense, based previous information about the program or similar evaluations of other programs? The direction of the estimate, namely whether the estimate is positive or negative, should be also considered. Is the direction expected or unexpected? If unexpected, what could explain the finding?

4. Consider how to present statistical significance in a manner best understood by the evaluation's stakeholders

Statistical significance of an estimate is an important factor in making decisions about what to report. Statistical significance can be reported in a variety of ways, and the research team should consider the audience that the evaluation is trying to reach when deciding how to report on statistical significance. How comfortable are program implementers and other stakeholders for the evaluation with the concept of statistical significance and their ability to interpret quantitative findings. It is the research team's responsibility to present findings in a manner that facilitates interpretation and discussion.

Statistical significance can be reported through p-values, through symbols for different p-values (e.g. * for $p < 0.05$), or through confidence intervals. Confidence intervals are often valuable because they provide additional information about the statistical certainty of the estimate through the relative width of the interval. For example, suppose we are looking at an effect estimate on the odds ratio scale of 1.95. A narrow 95% confidence interval, for example 1.92-1.98, tells us something different than a wide 95% confidence interval, such as 1.80-2.10. In both instances, we know that the initial effect estimate of OR=1.95 is statistically significant at $p < 0.05$. However, the narrow confidence interval indicates that there is greater statistical confidence that the true population-level effect size is OR=1.95, than the wider confidence interval. A difference in the width of a confidence interval is generally a feature of the study's sampling error (Stanley 2011). Some audiences may wish to see confidence intervals reported, while other audience may be content with p-values. The research team should keep their audience in mind when deciding how to present evaluation results.

5. Consider findings of no statistical significance and of findings going in an unexpected direction

The temptation is often strong just to focus on statistically significant findings that move in the expected direction. When we have a result that is positive and statistically significant, it suggests that our hypothesis has been tested, and we have information in support of our supposition that the program is effective. However, it is important to also consider results that are not statistically significant as well as results that are unexpected.

For purposes of transparency, the evaluation should also consider reporting when there are null findings, i.e. when results are not statistically significant. This is particularly important if these findings are in an area of particular focus for the program. Results that are not statistically significant can be harder to interpret than those that are. The findings could be related to challenges in how the study was implemented, programmatic implementation problems, or a theory of change about the program and its expected effects that is not plausible. The research team will need to work in partnership with program implementers to develop an explanation for important null findings.

Similarly, the research team should be report findings that are statistically significant and not in the expected direction. While these findings may be disappointing, they require an explanation, since these findings can be challenging to communicate with program implementers and with the evaluation's stakeholders. The research should similarly work to assess whether unexpected results related to study implementation, program implementation, and/or an unsupportable theory of change.

The table on the next page presents results for condom use, condom and lubricant use, and HIV testing from the evaluation of the Sisters program. It should be noted that these results are taken from a peer-reviewed publication. For a different audience, these reports might be presented in a very different manner.

TABLE 5. LOGISTIC REGRESSION ESTIMATES OF PROGRAM PARTICIPATION ASSOCIATIONS WITH CONDOM USE, CONDOM/LUBRICANT USE, AND HIV TESTING IN MATCHED AND UNMATCHED SAMPLES OF TRANSGENDER WOMEN IN PATTAYA, THAILAND, 2011*

FULL SAMPLE (N=308)							MATCHED SAMPLE (N=238)						
	OR, 95% CI	p-value	Model 1 Adjusted OR, 95% CI	p-value	Model 2 Adjusted OR, 95% CI	p-value		OR, 95% CI	p-value	Model 1 Adjusted OR, 95% CI	p-value	Model 2 Adjusted OR, 95% CI	p-value
CONDOM USE AT LAST SEX, ANY PARTNER													
Any service	2.51 (1.01-6.22)	0.047						3.75 (1.41-9.97)	0.008				
Drop-in center	1.70 (0.64-4.50)	0.289	1.48 (0.54-4.02)	0.445				2.48 (0.79-7.74)	0.121	1.92 (0.59-6.24)	0.280		
Outreach	2.02 (0.83-4.94)	0.122	1.86 (0.75-4.65)	0.182				3.10 (1.15-8.32)	0.025	2.68 (0.97-7.40)	0.057		
Home visit	4.53 (0.59-34.50)	0.145						3.66 (0.47-28.34)	0.214				
COMMERCIAL PARTNER (FULL SAMPLE N=289, MATCHED SAMPLE N=233)													
<i>Consistent condom use in the past 3 months</i>													
Any service	1.06 (0.49-2.28)	0.888						1.58 (0.74-3.38)	0.238				
Drop-in center	1.06 (0.54-2.08)	0.875						1.36 (0.64-2.88)	0.422				
Outreach	0.90 (0.44-1.86)	0.780						1.30 (0.63-2.70)	0.483				
Home visit	0.91 (0.40-2.11)	0.834						0.82 (0.33-2.02)	0.661				
<i>Consistent condom/water-based lubricant use in past 3 months</i>													
Any service	1.23 (0.67-2.27)	0.510						2.37 (1.28-4.41)	0.006				
Drop-in center	0.79 (0.46-1.35)	0.390	0.70 (0.40-1.22)	0.210	0.70 (0.40-1.22)	0.211		1.03 (0.57-1.86)	0.912	0.79 (0.42-1.48)	0.460	0.80 (0.43-1.51)	0.490
Outreach	1.59 (0.91-2.78)	0.103	1.73 (0.97-3.09)	0.062	1.75 (0.95-3.21)	0.074		2.72 (1.50-4.92)	0.001	2.89 (1.56-5.37)	0.001	3.22 (1.64-6.31)	0.001
Home visit	1.16 (0.57-2.36)	0.672			0.98 (0.46-2.08)	0.951		1.15 (0.53-2.52)	0.723			0.68 (0.29-1.62)	0.381
CASUAL PARTNER (FULL SAMPLE N=43, MATCHED SAMPLE N=29)													
<i>Consistent condom and water-based lubricant use in the past 3 months</i>													
Any service	0.81 (0.17-3.80)	0.787						0.49 (0.03-7.28)	0.604				
REGULAR PARTNER (FULL SAMPLE N=55, MATCHED SAMPLE N=32)													
<i>Consistent condom use in the past 3 months</i>													
Any service	1.00 (0.30-3.36)	1.000						0.35 (0.06-1.91)	0.225				
<i>Consistent condom and water-based lubricant use in the past 3 months</i>													
Any service	1.14 (0.35-3.75)	0.826						1.95 (0.38-10.08)	0.79				
<i>Received an HIV test in the past 6 months</i>													
Any service	3.32 (1.90-5.76)	0.000						2.45 (1.36-4.39)	0.003				
Drop-in center	3.17 (1.95-5.14)	0.000	2.84 (1.73-4.66)	0.000	2.83 (1.72-4.65)	0.000		2.80 (1.62-4.83)	0.000	2.60 (1.48-4.56)	0.001	2.58 (1.47-4.52)	0.001
Outreach	2.24 (1.37-3.65)	0.001	1.84 (1.11-3.06)	0.018	1.64 (0.96-2.80)	0.068		1.72 (1.01-2.93)	0.047	1.38 (0.79-2.41)	0.263	1.29 (0.72-2.34)	0.392
Home visit	2.11 (1.13-3.94)	0.019			1.59 (0.81-3.13)	0.179		1.67 (0.83-3.39)	0.153			1.29 (0.60-2.78)	0.518

*Multiple logistic regression models only estimated when factors statistically significant at $p < 0.05$ in bivariate model

DISSEMINATING RESULTS

In the best of cases, sharing evaluation results with stakeholders should provoke a discussion and critical assessment of the program evaluated. Research teams develop a set of actionable recommendations about the program, based on the results of the evaluation, and then use the results and recommendations to foster dialogue about the program.

A first step is to review preliminary results with the program team.

Organize a review meeting between program and research teams to discuss preliminary results

The following points can shape the discussion:

- Does the final matching solution seem appropriate? Are any factors missing?
- Do differences in how variables are distributed in the full and the matched samples raise any concerns for the program team, based on their experience working with the target population?
- Which results are expected? Which are unexpected?
- How does the program team explain the results that are expected, based on their knowledge of how the program operates?
- How does the program team explain results that are unexpected? Could these findings be based on how data was collected? Could they be based on how the program operates?
- How could the program adapt their implementation strategy to incorporate positive findings and address areas for improvement?
- What are the most important points to highlight to the evaluation's stakeholders?

TIP



Use this review meeting to assess how results are being interpreted and understood by the program and to highlight areas where further data analysis may be needed.

The next step is to finalize analysis, reporting, and recommendations about the program. It may be useful for program implementers to review the recommendations about the program again to ensure that the recommendations are meaningful and actionable for the program.

The research team should also develop a dissemination plan for the evaluation, in conjunction with program implementers. This scope of this dissemination plan will happen on who the evaluation's stakeholder are.

A variety of channels and methods of dissemination are possible (Bennett & Jessani 2011; Gertler et al. 2011; Patton 2008). The results of the Sisters evaluation were shared through the following channels, and for the conferences were formatted as presentations:

- Asia-Pacific HIV/AIDS Conference
- American Evaluation Association Conference
- PLoS One journal article

BOX 11 | GETTING EVALUATION RESULTS INTO USE

The act of program evaluation can seem threatening to the people running the program being evaluated, especially if the evaluation is designed to test whether or not the program works.

The evaluation field has generally found that evaluation results are mostly likely to be used by program implementers and/or policy-makers when the evaluation was designed to address a particular evidence need or a decision (Weiss 1998; Patton 2008).

If the evaluation has been planned with key stakeholders in mind, both the program team and external stakeholders, it is more likely that results will be taken up and used.

Even when an evaluation has been designed in consultation with program implementers and other stakeholders, it is possible that some results may be received negatively. It is important that the research team be certain about how they have arrived at their conclusions and recommendations and be able to explain them to a non-research audience, including being able to discuss unexpected results.

CONCLUSIONS

This manual is intended to guide PSI programs managers and researchers through the many thought processes and steps needed in order to conduct an impact evaluation using CEM. While there are many issues to consider and research steps to take, the use of CEM for PSI evaluations is an accomplishable goal. The checklist presented below the main steps needed to successfully complete an evaluation using this approach.

Although the information presented here is important for a the design of a good evaluation using CEM, the manual is also intended to allow for flexibility in the design and reporting of findings so that different PSI programs can tailor their analyses and reports to their specific program and the needs of their funders and other stakeholders. Further examples on the use of CEM in the evaluation of PSI programs are presented in the Appendices. These are provided to help guide researchers, but are not necessarily expected to be used as templates.

BOX 12 | FINAL CONCLUSIONS ABOUT THE SISTERS EVALUATION

"Our findings suggested that an HIV prevention program targeted to transgender women can address HIV-related risks, evidence that is needed given the substantial HIV burden this population faces in Thailand and globally. Key elements appear to be making water-based lubricant accessible along with condoms in outreach activities, and embedding rapid HIV testing in community-based, transgender-friendly services..."

These findings also highlight the potential value of developing more broad-based HIV prevention programming specific to transgender women and independent of activities for MSM in Thailand. This response could include dedicated sampling of transgender women in integrated bio-behavioral surveys of HIV prevalence plus identification of transgender women as a specified key population in government plans for HIV/AIDS control and in reporting to the United Nations. These efforts would contribute to more effectively meeting the needs of this key population" (Pawa et al 2013).

CEM ANALYSIS STEPS CHECKLIST

- ☐ Define evaluation question
 - Define outcomes
 - Define exposure
- ☐ Define covariates for matching
 - Identify risks of selection bias
 - Identify potential confounders
- ☐ Incorporate relevant questions into survey based on steps 1 and 2
- ☐ Plan sampling strategy
 - Ensure sufficient overlap or areas of common support between exposed and unexposed
 - Assume sample trimming during analysis when planning sample size
- ☐ Collect and clean data
- ☐ Match sample
 - Operationalize exposure variable
 - Assess matching covariates for distributions and relationship with exposure
 - Precoarsen covariates, if applicable
 - Assess initial imbalance of covariates
 - Create a matched sample using CEM algorithm
 - Assess output and refine matching process until an acceptable balance is reached between the L1 statistic and sample size (iterative process)
- ☐ Estimate model(s)
 - Operationalize outcome variables and conduct univariate analyses
 - Examine bivariate relationships between exposure and outcomes
 - Run multivariate estimation models examining the inclusion of different covariates based on statistical and theoretical rationales for inclusion (iterative process)
- ☐ Interpret and present results
 - Provide descriptive statistics
 - Consider non-significant and unexpected results
 - Review results with team
 - Finalize and disseminate results

APPENDICES

APPENDIX 1: GLOSSARY OF TERMS

Area of common support: Overlap in distribution of covariates for the exposed and unexposed.

Bias: a systematic error in the design, conduct, or analysis of a study that leads to mistakes in estimates of results.

Coarsening: categorizing variables into broader groups for the purpose of matching.

Confounding variable/factor: an extraneous variable that correlates with both the dependent and independent variable; see omitted variable bias.

Control group: group in a scientific study that is not exposed to the intervention; see counterfactual and treatment group.

Counterfactual: what would have happened to the target population in the absence of the program; see control group.

Covariate: analysis variable that can affect the relationship between the dependent variable and the independent variable(s) of interest

Endogeneity: when there is a correlation between the outcome variable and error term of a statistical model.

Experiment: scientific procedure to test a hypothesis by randomly assigning individuals (or other units of analysis) to exposed and unexposed groups.

Imbalance: the extent to which exposed and unexposed groups are different from each other on covariates.

Impact evaluation: evaluations used to address questions of program effectiveness by looking for causal links between a specific program, policy, or intervention and outcomes.

Inference: a conclusion based on evidence and reasoning.

L1 statistic: a summary CEM measure of global imbalance calculated by comparing the differences between all the covariates at once.

Model dependence: where findings are highly dependent on variable specifications in the model rather than the true relationship between independent and outcome variables.

Omitted variable bias: when a model is incorrectly specified by leaving out an unmeasured causal variable; see confounding variable/factor.

Plausibility: a believable and reasonable justification for why exposure is can lead to a change in outcomes.

Program exposure: individual came into contact with program, such as saw a mass media communication or received a home visit.

Program participation: active engagement with a program where an individual volunteers to participate, such as attending a clinic or attending a meeting.

Quasi-experiment: a scientific study to test a hypothesis but lacks the feature of random assignment.

Sampling error: error that can occur by chance when statistically examining a sample rather than an entire population of interest.

Selection bias: the introduction of error due to systematic differences in the characteristics between those who were and were not exposed to a given program, so that the sample is not representative of population intended to be analyzed.

Temporality: when program exposure occurs before the outcome is measured.

Treatment group: group in a scientific study that is exposed to the intervention; see control group.

Validity: the extent that a concept, conclusion or measurement accurately corresponds to the real world.

APPENDIX 2: CEM INSTALLATION

How to install CEM in Stata

First-time users of CEM need to install the cem package in Stata before initiating any matching.

Under Stata 10 or later, type the following: `ssc install cem`

Users will need to be connected to internet to install the cem package successfully.

How to install CEM in SPSS

Pre-Installation steps

1. Match the version of R you will need with the version of SPSS you have
 - SPSS 18 = R 2.8
 - SPSS 19 = R 2.10
 - SPSS 20 = R 2.12
2. Figure out if you have Windows 32-bit or 64-bit
3. Make sure you have administrator rights to install software on your computer

Installation steps

1. Install R
 - a. Right click for "Run as administrator"
2. Open R
 - a. Install the CEM package by typing `install.packages("cem")`
 - b. Select a CRAN mirror near you (South Africa or UK) to download and install the CEM package
 - c. Make sure CEM installed properly library (cem)
3. Install the SPSS R-Essentials plug-in
 - a. Make sure you're connected to Internet first
 - b. Right click for "Run as administrator"
 - c. Select: "C:\Program Files\R\"
 - d. Select "C:\Program Files\IBM\SPSS\Statistics\"
4. Install the SPSS Python-Essentials plug-in
 - a. Right click for "Run as administrator"
 - b. Install Python 2.6 from this installer to "C:\Python26"
 - c. Select: "C:\Python26"
 - d. Select "C:\Program Files\IBM\SPSS\Statistics\"
5. Go to Start menu and move your mouse to the SPSS icon
 - a. Right click for "Run as administrator"
6. In SPSS, go to "Utilities > Extension Bundles > Install Extension Bundle..."
 - a. Direct the dialog box to the "cem.spe" file you downloaded
7. Restart SPSS

Troubleshooting

- Make sure the CEM package installed to R
- Make sure you have administrator rights
- If you run into trouble, uninstall everything and start over

APPENDIX 3: SPSS SYNTAX FOR USING CEM

CEM SPSS syntax: the basics

```
CEM TREATMENT= [exp var] VARIABLES= [match var] [match var] [match var].
CEM SPSS syntax: options
CEM TREATMENT=[exp var] VARIABLES=[match var] [match var] [match var]
/NOCOARSENING VARIABLES=[match var]
/CUTPOINTS "[match var]=(value)"
/GROUPING " [match var]=[(value),(value, value)]".
```

Coarsening with continuous variables

```
CEM TREATMENT= ss_any VARIABLES= sxwk emp_entv incm dur_pat tgn0 edu
/CUTPOINTS "tgn0=(4, 7, 18)".
```

Coarsening with categorical variables

```
CEM TREATMENT= q29d ss_any VARIABLES= sxwk emp_entv incm dur_pat tgn0 edu
/GROUPING "a2 = [(1), (2), (3, 4, 5)]".
```

Not allowing any coarsening

```
CEM TREATMENT=ss_any VARIABLES= sxwk emp_entv incm dur_pat tgn0 edu
/NOCOARSENING VARIABLES=edu.
```

To apply CEM weights just calculated in SPSS, first ensure that CEM weights are factored into existing sampling weights. A new weight variable can be calculated by multiplying the two or more weights together, as follows:

```
new_wt=wt*cem.weights.
```

Next, SPSS requires an analysis that filters out unmatched cases, follows:

```
filter by cem.matched.
```

Working now with a sub-set of the original dataset, the analyst should apply either the CEM weights or the new weighting variable just calculated, as follows:

```
weight by cem.weights.
```

The analyst should be mindful that all subsequent analysis steps will be conducted in the filtered dataset and with weights applied. Filters should be removed and weights turned off to return to the original dataset.

APPENDIX 4: SAMPLE EXPOSURE QUESTIONNAIRES

PSI-MADAGASCAR SURVEY ON CHILD NUTRITION

In the last 3 months have you participated in an individual talk about infant and young child nutrition in your community/household by a CHW? (Community Health Worker) (ACN or AC PCIMEC)	Yes No
In the last 3 months have you received/participated in a group activity/talk about infant and young child nutrition in your community by a CHW (ACN or AC PCIMEC)? <i>NB, this can be at a SEECALINE center or any other place.</i>	Yes No

PSI-MYANMAR SURVEY ON HIV RISK BEHAVIORS

Q1025	INTERVIEWER: In Yangon, only read out DIC details for Yangon. In Mandalay, only read out DIC details for Mandalay. Have you ever visited the TOP drop in center at DiCYangon at (no. 148-b 1/2, A1 street, 9 mile, MayangoneTsp) Have you ever visited the TOP drop in center at DiCMandalay at (no. nga3/53, Thumarlar Street, Between KhaingShweWar and ZalatSt, ChanMyaTharZiTsp)	Yes	If "No" OR "Don't know/don't remember", skip to Q1027
		No	
		Don't know/ Don't remember	
Q1026	When was the last time you visited that drop-in center?	Within the last 2 weeks	
		Between 2 weeks to 1 month ago	
		Between 1-3 months ago	
		Between 4-6 months ago	
		Longer than 6 months	
		Don't know/don't remember	

PSI-CHINA SURVEY ON HIV RISK BEHAVIORS AMONG INJECTING DRUG USERS

Q1101	In the past 12 months, have you visited a drop in center? If yes, ask: how many time have you visited the drop in center in the past 12 months? If no, write down "0".	1. HuxianghaoBa Drop-in Center	__times	If all answers are "0", skip to Q1105.
		2. Poplar TreeinGejiu	__times	
		3. Green Garden Drop-in Center in Gejiu	__times	
		4. Wangzhou Community in Nanning	__times	
		5. Red Rubbin Center in Nanning	__times	
		6. Yuxin Home in Luzhai	__times	
		7. Red Rubbin Center in Luzhai	__times	

PSI-CHINA SURVEY ON HIV RISK BEHAVIORS AMONG INJECTING DRUG USERS

Q1102	In the past 12 months, have you ever met with an outreach worker or peer educator or program staff who talked with you about your health, particularly HIV related issues? If yes, ask:how many times have you had contact with an outreach worker or peer educator outside of aDiC? If no, write down "0".	1. Huxianghao Ba in Kunming	__times	If all answers are "0", skip to Q1111
		2. Poplar Tree in Gejiu	__times	
		3. Sun Flower in Gejiu	__times	
		4. Green Garden in Gejiu	__times	
		5. Jinhudong DiC in Gejiu	__times	
		6. Red Rubbin Center in Nanning	__times	
		7. Wangzhou Community in Nanning	__times	
		8. Yuxin Homeland in Luzhai	__times	
		9. Red Rubbin Center in Luzhai	__times	
		10. PSI Guangxi office	__times	
		11. Nanning CDC	__times	
		12. Shuangguang Homeland	__times	
		13. Alliance Guangxi office	__times	
		14. FHI Guangxi office	__times	

PSI-ZIMBABWE SURVEY ON WATER TREATMENT

Q219	Have you ever seen or heard any message promoting boiling of water or water treatment using WaterGuard indicating that it protects your family's health, makes your water safe and prevents diarrheal diseases?	Yes No	1 0	IF 0, Skip to Q222
Q220	Where did you see or hear the message(s) about water treatment? (multiple responses possible)	Newspaper/ Magazine	0	1
		Radio	0	1
		TV	0	1
		Dramagroups/Roadshows	0	1
		Public clinic	0	1
		Poster	0	1
		Point of sale	0	1
		Leaflets	0	1
		Billboards	0	1
		In-store promotion	0	1
		Door to door campaign	0	1
		Other (Specify)	_____	

PASMO SURVEY ON HIV RISK BEHAVIORS

H5	INTERVIEWER: Show card 8 and 32 with picture of voucher. In the last year, how many times did you use a voucher like the one I am showing you to go to a doctor to check if you had an STI?	___times	Write the exact amount. Only write "0" if the person has never used a voucher.
Thinking about the last 12 months...			
I3	Would you please tell me how many times were you tested for HIV?	___ times	Write 0 if never, then jump to I7.
I4	Some people get tested but don't go back to get their results. How many of those times did you go back to pick yours?	___ times	Write 0 if never picked up results, then jump to I7. The number should be equal or lower than I3.
I5	INTERVIEWER: Show card 8 and 32 with picture of voucher. Of those occasions when you were tested and picked your results, how many times did you use a voucher like this?	___ times	Write 0 if never and jump to I7. The number should be equal or lower than I4.

PASMO SURVEY ON HIV RISK BEHAVIORS

			Never	Less than once a week	At least once a week	Every day	Don't know	No answer
J1-7	In the last 12 months, how often would you say that you...	Watched national TV	0	1	2	3	-1	-2
		Watched cable TV	0	1	2	3	-1	-2
		Listened to the radio	0	1	2	3	-1	-2
		Read a newspaper	0	1	2	3	-1	-2
		Read a magazine	0	1	2	3	-1	-2
		Went to the movies	0	1	2	3	-1	-2
		Accessed the internet (If never, skip to J20).	0	1	2	3	-1	-2
J8-18	In the last 12 months, how often would you say that you entered the following websites:		Never	Less than once a week	At least once a week	Every day	Don't know	No answer
		¿Yahoraqué? / And now what?	0	1	2	3	-1	-2
		Clubenconexión	0	1	2	3	-1	-2
		MiZona H	0	1	2	3	-1	-2
		Red Segura	0	1	2	3	-1	-2
		Facebook	0	1	2	3	-1	-2
		Tweeter	0	1	2	3	-1	-2
		Manhunt	0	1	2	3	-1	-2
		Gayguatemala(solo Guatemala)	0	1	2	3	-1	-2
		MundoAnuncio	0	1	2	3	-1	-2
		Elchat.com	0	1	2	3	-1	-2
		Badoo	0	1	2	3	-1	-2

PASMO SURVEY ON HIV RISK BEHAVIORS

J20	INTERVIEWER: Show “Hombres deVerdad” card 19. Do you remember seeing this campaign before?	No (jump to J30)	0
		Yes	1
		Don't know (Jump to J30)	97
		Don't answer (Jump to J30)	97
J21	INTERVIEWER: Do not read options aloud. For coding purposes only. What is this campaign inviting the people to do?	Nothing	0
		Protect myself and/or my partner from AIDS	1
		Use a condom	2
		Get tested for HIV	3
		Not discriminate persons with HIV/AIDS	4
		Be faithful to my partner	5
		Carry a condom with me	6
		Other (specify)	7
		Don't know	-1
		Don't answer	-2
J22	Is this campaign directed to you?	No	0
		Yes	1
		Don't know	-1
		Don't answer	-2
J30	INTERVIEWER: Show “TienesPidelo” card 20. Do you remember seeing this campaign before?	No (jump to J40)	0
		Yes	1
		Don't know (Jump to J40)	97
		Don't answer (Jump to J40)	97
J31	INTERVIEWER: Do not read options aloud. For coding purposes only. What is this campaign inviting the people to do?	Nothing	0
		Protect myself and/or my partner from AIDS	1
		Use a condom	2
		Get tested for HIV	3
		Not discriminate persons with HIV/AIDS	4
		Be faithful to my partner	5
		Carry a condom with me	6
		Other (specify)	7
		Don't know	-1
		Don't answer	-2
J32	Is this campaign directed to you?	No	0
		Yes	1
		Don't know	-1
		Don't answer	-2

PASMO (CENTRAL AMERICA)

J40	INTERVIEWER: Show "Impresiónala" card 21. Do you remember seeing this campaign before?	No (jump to J50) Yes Don't know (Jump to J50) Don't answer (Jump to J50)	0 1 97 97
J41	INTERVIEWER: Do not read options aloud. For coding purposes only. What is this campaign inviting the people to do?	Nothing Protect myself and/or my partner from AIDS Use a condom Get tested for HIV Not discriminate persons with HIV/AIDS Be faithful to my partner Carry a condom with me Other (specify) Don't know Don't answer	0 1 2 3 4 5 6 7 -1 -2
J42	Is this campaign directed to you?	No Yes Don't know Don't answer	0 1 -1 -2
K1	In the last 12 months, have you been given condoms for free (through an outreach service, sexual health clinic or other)?	No Yes Don't Know Don't Answer	0 1 -1 -2
K2	In the past year, how many times have you been approached either in group or privately by an educator from any institution and discussed ways to prevent HIV or AIDS one on one?	_____ times	If "none" write 0 and jump to K20
K3	INTERVIEWER: Show voucher card 31. How many of these times did a voucher like this was either provided to you or required from you?	_____ times	If "none" write "0". Number should be lower than K2 .

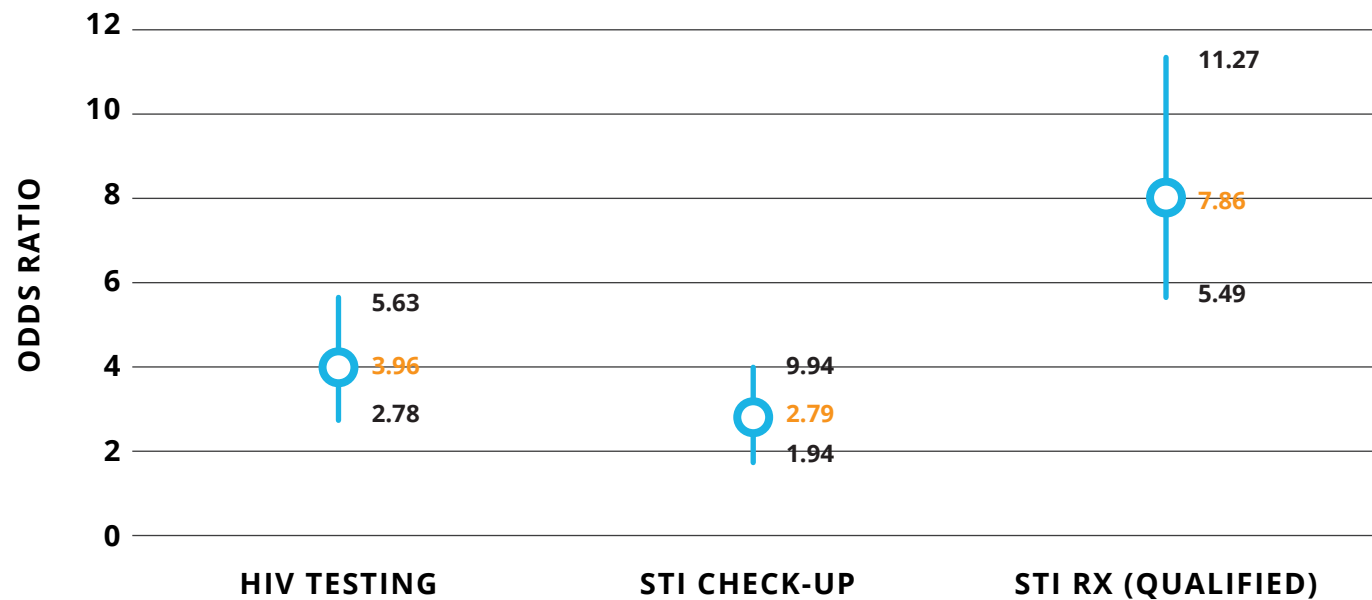
PASMO (CENTRAL AMERICA)

K10-18	INTERVIEWER: Show cards as listed I will show you some images of games that you could have played with a PASMO educator. Please tell me how many times you played with each in the last year?		If none, write 0.
	Card 22: Elreto(BZE: Pataki)	_____ times	
	Card 23: XY	_____ times	
	Card 24: 123saludable(123 safe)	_____ times	
	Card 25 (not in BZE):Decisiones	_____ times	
	Card 26 (BZE only): Decisions	_____ times	
	Card 27 (BZE only): Condoms and ladders	_____ times	
	Card 28 (BZE only) Matchmaker	_____ times	
	Card 29 (BZE only): Large cards	_____ times	
	Card 30 (BZE only): Sex busters	_____ times	
K20	In the past year, how many times have you talked in a chat room about the ways of preventing HIV or AIDS ?	_____ times	If none, write "0" and jump to K30 .
K21	INTERVIEWER: Show Card 32 with Voucher. How many of these times did a voucher like this was either provided to you or required from you?	_____ times	Number should be lower than K20.
K30	Do you have any friends in organizations that work with HIV, AIDS or Sex workers?	No Yes Don't know No answer	0 1 -1 -2
L1	INTERVIEWER: Show card 33 with voucher. In the last 12 months, how many times have you been provided with a voucher like this, same figure and color, for you to access free services?	_____ times	If "none" write 0 and then jump to next section.
L2	And how many of these did you use?	_____ times	If "none" write 0 and then jump to next section.
L13-17	And in the last year how many times have you used one of these vouchers to access...		Write the exact amount, If "none" write 0.
	Free counseling about alcohol and substance abuse (AA, NA, etc.)	_____ times	
	Free counseling about discrimination and stigma issues	_____ times	
	Free counseling about violence issues	_____ times	
	Free counseling about legal matters (immigration, documents for personal identification, etc.)	_____ times	
	Other	_____ times	

APPENDIX 5. MODEL TABLES

PSI-MADAGASCAR PROGRAM EFFECTS ON INFANT AND YOUNG CHILD FEEDING PRACTICES			
INDICATORS	DID NOT RECEIVE A HOME VISIT BY A CHW ON NUTRITION AND DID NOT PARTICIPATE IN A CHW LED NUTRITION GROUP ACTIVITY/TALK AND DID NOT HEAR RADIO MESSAGE	RECEIVED A HOME VISIT BY A CHW ON NUTRITION OR PARTICIPATED IN A CHW LED NUTRITION GROUP ACTIVITY/TALK BUT DID NOT HEAR RADIO MESSAGE	SIG.
BEHAVIOR	(N=219)	(N=138)	
	%	%	
Children 6-23 months of age who received foods from 4 or more food groups during the past 24 hours	43.0	55.7	*
Children 6-23 months of age who received the minimum dietary diversity and the minimum meal frequency during the past 24 hours	36.5	47.2	*
Caregivers who consistently use Zacatomady (about at least three sachets) in the last week	14.2	26.7	**
Caregivers who only breastfed children 0-5 months	86.4	93.3	ns
Caregivers who continued breastfeeding beyond six months of age	82.4	81.5	ns
INDICATORS	DID NOT HEAR RADIO MESSAGES ON NUTRITION AND DID NOT RECEIVE OR PARTICIPATE IN IPC ACTIVITIES	HEARD RADIO MESSAGES ON NUTRITION BUT DID NOT RECEIVE OR PARTICIPATE IN IPC ACTIVITIES	SIG.
BEHAVIOR	(N=75)	(N=340)	
	%	%	
Children 6-23 months of age who received foods from 4 or more food groups during the past 24 hours	29.8	52.2	**
Children 6-23 months of age who received the minimum dietary diversity and the minimum meal frequency during the past 24 hours	19.6	43.6	***
Caregivers who consistently use Zacatomady (about at least three sachets) in the last week	10.3	15.7	ns
Caregivers who only breastfed children 0-6 months	89.1	88.4	ns
Caregivers who continued breastfeeding beyond six months of age	76.3	87.3	*

PSI-MYANMAR PROGRAM EFFECTS ON HIV RISK BEHAVIORS AMONG FEMALE SEX WORKERS



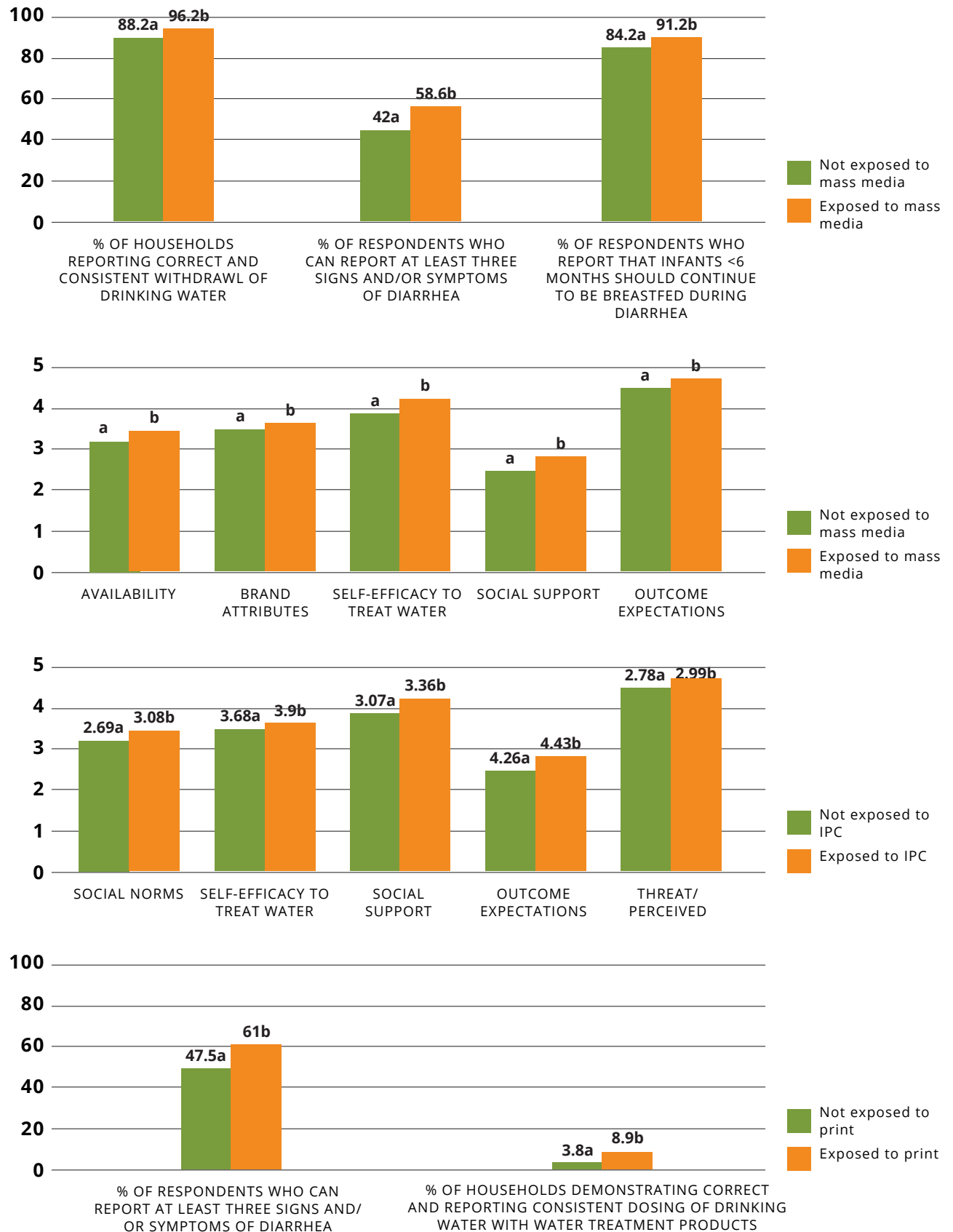
PSI-CHINA PROGRAM EFFECTS ON SAFE INJECTING DRUG USE

	Never shared needles or syringes in past 3 months				Always kept a new needle on hand in past 3 months			
Predictors	Pooled sample: (n=823)		Matched sample: (n=762)		Pooled sample: (n=823)		Matched sample: (n=762)	
	Bivariate	Multivariate	Bivariate	Multivariate	Bivariate	Multivariate	Bivariate	Multivariate
	OR, 95% CI	Adj. OR, 95% CI	OR, 95% CI	Adj. OR, 95% CI	OR, 95% CI	Adj. OR, 95% CI	OR, 95% CI	Adj. OR, 95% CI
Program exposure								
Participated in DiC-based intervention in past 12 months	1.03 (0.55-1.94)	1.03 (0.53-2.01)	0.76 (0.39-1.48)	0.76 (0.37-1.56)	1.35 (0.99-1.85)	1.24 (0.89-1.74)	1.59** (1.17-2.17)	1.47* (1.05-2.06)
Participated in outreach in past 12 months	0.95 (0.49-1.87)	0.93 (0.46-1.90)	0.91 (0.45-1.85)	1.04 (0.49-2.24)	1.42* (1.03-1.98)	1.32 (0.93-1.87)	1.48** (1.07-2.05)	1.28 (0.90-1.83)
Ethnicity (Han)	---	0.74 (0.28-2.02)	---	0.39 (0.09-1.70)	---	1.49 (0.93-2.40)	---	1.43 (0.86-2.38)
Education (High school education or above)	---	1.59 (0.76-3.36)	---	2.79* (1.18-6.61)	---	1.03 (0.74-1.44)	---	1.13 (0.82-1.56)

Note: *significant at $p < 0.05$, ** significant at $p < 0.01$; *** significant at $p < 0.001$. Age, gender, respondents' working hours, resident city, and ever use of MMT were matched for the CEM mode

PSI-ZIMBABWE PROGRAM EFFECTS ON WATER TREATMENT

Evaluation Graphs: PSI interventions and water treatment behaviors and related correlates among adults 15 to 59 years in Zimbabwe, 2013



PASMO PROGRAM EFFECTS ON CONDOM USE AND HTC AMONG MSM (ALL COUNTRIES)

	Full sample (n=3531) *		Matched Sample (n=2922) **	
Condom use at last sex with any partner				
	OR (95%CI)	AdjOR (95%CI)	OR(95%CI)	AdjOR (95%CI)
Any	1.49 (0.81, 2.74)	1.50 (0.81, 2.79)	1.44 (0.74, 2.82)	1.44 (0.73, 2.84)
Behavioral	1.56 (0.72, 3.35)	1.55 (0.72, 3.35)	1.57 (0.67, 3.66)	1.56 (0.66, 3.69)
Biomed	1.64 (0.65, 4.13)	1.67 (0.65, 4.26)	1.73 (0.63, 4.71)	1.73 (0.62,4.81)
Complementary	1.24 (0.50, 3.05)	1.26 (0.50, 3.13)	1.11 (0.44, 2.84)	1.10 (0.42, 2.84)
Behavioral+Biomed	2.32 (0.467, 11.46)	2.16 (0.43, 10.78)	2.48 (0.42, 14.64)	2.39 (0.40, 14.32)
Behavioral+Compl	1.38 (0.37, 5.20)	1.33 (0.35, 5.05)	1.29 (0.33, 5.09)	1.21 (0.30, 4.88)
Combination	2.35 (0.24, 22.80)	2.26 (0.23-22.09)	2.50 (0.20, 30.82)	2.48 (0.20, 31.26)
Condom and lubricant use at last sex with any partner				
	OR (95%CI)	AdjOR (95%CI)	OR (95%CI)	AdjOR (95%CI)
Any	1.35 (0.90, 2.02)	1.32 (0.87, 1.99)	1.40 (0.90, 2.17)	1.31 (0.84, 2.05)
Behavioral	1.73 (1.06, 2.83)	1.66 (1.01, 2.74)	1.84 (1.08, 3.14)	1.70 (0.98, 2.95)
Biomed	1.54 (0.88, 2.71)	1.53 (0.86, 2.71)	1.71 (0.94, 3.12)	1.57 (0.85, 2.90)
Complementary	0.86 (0.47, 1.58)	0.86 (0.47, 1.59)	0.86 (0.46, 1.61)	0.83 (0.44, 1.58)
Behavioral+Biomed	3.14 (1.22 ,8.13)	2.88 (1.10, 7.51)	3.51 (1.26, 9.80)	3.05 (1.08, 8.64)
Behavioral+Compl	1.50 (0.64, 3.54)	1.44 (0.61, 3.42)	1.56 (0.64, 3.80)	1.43 (0.58, 3.53)
Combination	3.54 (0.91, 13.75)	3.33 (0.85, 13.04)	4.12 (0.94, 18.15)	3.76 (0.84, 16.88)
Consistent condom use in the last 30 days with all partner types				
	OR (95%CI)	AdjOR (95%CI)	OR (95%CI)	AdjOR (95%CI)
Any	1.17 (0.66, 2.06)	1.21 (0.68, 2.16)	1.38 (0.75, 2.53)	1.37 (0.70, 2.50)
Behavioral	1.01 (0.52, 1.97)	1.05 (0.53-2.07)	1.13(0.55, 2.30)	1.17 (0.56, 2.44)
Biomed	1.47 (0.63, 3.45)	1.45 (0.61, 3.46)	1.89 (0.75, 4.77)	1.82 (0.71, 4.65)
Complementary	0.82 (0.38, 1.80)	0.82 (0.37,1.81)	0.93 (0.41, 2.09)	0.90 (0.39, 2.04)
Behavioral+Biomed	1.12 (0.34, 3.67)	1.09 (0.33, 3.60)	1.28 (0.37, 4.45)	1.23 (0.35, 4.37)
Behavioral+Compl	0.71 (0.25, 2.02)	0.69 (0.24, 1.99)	0.84 (0.28, 2.48)	1.09 (0.30, 3.98)
Combination	0.72 (0.17, 3.10)	0.70 (0.16, 3.06)	0.90 (0.19, 4.22)	0.85(0.18, 4.17)
Consistent condom use in the last 30 days with regular partners				
	OR (95%CI)	AdjOR (95%CI)	OR (95%CI)	AdjOR (95%CI)
Any	1.72 (1.15, 2.57)	1.71 (1.13, 2.57)	1.69 (1.09, 2.62)	1.51 (0.88, 2.57)
Behavioral	1.83 (1.12, 3.00)	1.78 (1.08, 2.93)	1.88 (1.10, 3.21)	1.88 (1.09, 3.25)
Biomed	1.18 (0.68, 2.05)	1.17 (0.67, 2.05)	1.19 (0.66, 2.14)	1.19 (0.65, 2.17)
Complementary	1.64 (0.90, 3.00)	1.63 (.089, 3.01)	1.50 (0.79, 2.82)	1.53 (0.80, 2.90)
Behavioral+Biomed	1.66 (0.71, 3.88)	1.56 (0.66, 3.67)	1.74 (0.71, 4.28)	1.63 (0.67, 4.19)
Behavioral+Compl	1.81 (0.76, 4.30)	1.72 (0.72, 4.10)	1.69 (0.69, 4.16)	1.67 (0.67, 4.14)
Combination	1.41 (0.43, 4.54)	1.33 (0.41,4.32)	1.34 (0.39, 4.57)	1.31 (0.38, 4.56)

Consistent condom use in the last 30 days with commercial partners				
	OR (95%CI)	AdjOR (95%CI)	OR (95%CI)	AdjOR (95%CI)
Any	1.47 (0.91, 2.37)	1.58 (0.97, 2.59)	1.39 (0.83, 2.33)	1.51 (0.88, 2.57)
Behavioral	1.44 (0.82, 2.53)	1.58 (0.89, 2.83)	1.36 (0.74, 2.50)	1.5 (0.79, 2.78)
Biomed	0.99 (0.51, 1.93)	1.03 (0.52, 2.03)	0.86 (0.428, 1.74)	0.9 (0.45, 1.90)
Complementary	1.87 (0.97, 3.62)	1.95 (1.00, 3.82)	1.69 (0.84, 3.38)	1.8 (0.87, 3.61)
Behavioral+Biomed	0.88 (0.30, 2.59)	0.94 (0.32, 2.80)	0.78 (0.25, 2.44)	0.8 (0.26, 2.68)
Behavioral+Compl	2.34 (0.96, 5.70)	2.47 (1.00, 6.09)	2.11 (0.84, 5.32)	2.2 (0.85, 5.69)
Combination	1.44 (0.38, 5.41)	1.52 (0.40, 5.82)	1.28 (0.312, 5.18)	1.4 (0.32, 5.66)
Had an HIV test and received results in the last 12 months				
	OR (95%CI)	AdjOR (95%CI)	OR (95%CI)	AdjOR (95%CI)
Any	3.98 (2.54, 6.23)	4.06 (2.57, 6.42)	3.04 (1.88, 4.91)	2.98 (1.82, 4.87)
Behavioral	2.49 (1.49, 4.17)	2.48 (1.47, 4.18)	1.86 (1.07, 3.22)	1.76 (1.01, 3.10)
Biomed				
Complementary	2.54 (1.28, 5.08)	2.55 (1.27, 5.12)	1.97 (1.0, 4.05)	1.95 (0.94, 4.03)
Behavioral+Biomed				
Behavioral+Compl	3.52 (1.23, 10.13)	3.45 (1.19, 9.97)	2.77 (0.93, 8.30)	2.63 (0.87, 7.93)
Combination	39.58 (0.80, 1978.03)	37.78 (0.75, 1894.03)	43.22 (0.37, 5070.70)	40.03 (0.34, 4716.04)

* Results weighted for network size, recruitment chain, and city population size

** Results weighted for network size, recruitment chain, city population size, and matching stratum

Note: Country-specific data from Belize, Costa Rica, El Salvador, Guatemala, Nicaragua, and Panama are also available upon request.

APPENDIX 6. REFERENCES

- Albouy, D. (n.d.). Program Evaluation and the Difference in Difference Estimator [course notes]. Berkeley. Accessed 2 Feb 2015. Available online at: http://eml.berkeley.edu/~webfac/saez/e131_s04/diff.pdf
- Amon, J., T. Brown, J. Hogle, J. MacNeil, et al. (2000). Behavioral Surveillance Surveys: Guidelines for Repeated Behavioral Surveys in Populations at Risk of HIV. Durham, NC: Family Health International.
- Bennett, G. & N. Jessani (eds.) (2011). The Knowledge Translation Toolkit. Thousand Oaks, CA: Sage and International Development Research Centre.
- Berry, S., M.C. Escobar & H. Pitorak (2012). "I'm Proud of My Courage to Test": Improving HIV Testing and Counseling among Transgender People in Pattaya, Thailand [web publication]. AIDSTAR-One. Accessed 3 December 2014. Available online at: http://www.aidstar-one.com/sites/default/files/imagecache/AIDSTAR-One_CaseStudy_HTC_Sisters_Thailand.pdf.
- Blackwell, M., S. Iacus, G. King & G. Porro (2009). cem: Coarsened Exact Matching in Stata. The Stata Journal, 9.4, 524-546.
- California State University, Long Beach. Quasi-Experimental Research Designs [web page]. Accessed 10 Dec 2014. Available online at: <http://www.csulb.edu/~msaintg/ppa696/696quasi.htm>
- Center for Theory of Change (2013). Theory of Change [web page]. Accessed 29 Jan 2015. Available online at: <http://theoryofchange.org>.
- Firestone, R. (2012). How to Plan a Research Study [web page]. Population Services International. Accessed 4 Feb 2015. Available on KIX.
- Gertler, P.J., S. Martinez, P. Premand, L.B. Rawlings & C.M.J. Vermeersch (2011). Impact Evaluation in Practice. Washington, DC: World Bank.
- Harrell, F.E. (2002). Regression Modeling Strategies. New York, NY: Springer.
- Heckathorn, D. (2012). Respondent Driven Sampling [web page]. Cornell University. Accessed 10 Dec 2014. Available online at: <http://www.respondentdrivensampling.org>.
- Heckman, J.J., R.J. LaLonde & J.A. Smith (1999) The Economics and Econometrics of Active Labor Market Programs. In O. Ashenfelter & D. Card (eds.), Handbook of Labor Economics (Volume 3), 1865-2097. Elsevier. Available online at: doi:10.1016/S1573-4463(99)03012-6.
- Iacus, S.M., G. King & G. Porro (2011). Causal Inference without Balance Checking: Coarsened Exact Matching [web publication]. Oxford University Press. Available online at: http://gking.harvard.edu/files/gking/files/political_analysis-2011-iacus-pan_mpr013.pdf.
- Institute for Digital Research and Education. Stata Topics: Logistic (and Categorical) Regression [web page]. Accessed 3 Dec 2014. Available online at: http://www.ats.ucla.edu/stat/stata/topics/logistic_regression.htm
- Institute for Digital Research and Education. Stata Topics: Regression [web page]. Accessed 3 Dec 2014. Available online at: <http://www.ats.ucla.edu/stat/stata/topics/regression.htm>
- Institute for Digital Research and Education (2014). Statistical Computing Seminars: Visualizing Main Effects and Interactions for Binary Logit Models in Stata [web page]. Accessed 10 Dec 2014. Available online at: http://www.ats.ucla.edu/stat/stata/seminars/stata_vibl/
- Mitchell, M.N. and X. Chen (2005). Visualizing main effects and interactions for binary logit models. The Stata Journal, 5.1, 64-82.
- Morris, D.S., M.P. Rooney, R.J. Wray & M.W. Kreuter (2009). Measuring Exposure to Health Messages in Community-Based Intervention Studies: A Systematic Review of Current Practices. Health Education Behavior, 36, 979-998.

- Patton, M.Q. (2008). *Utilization-Focused Evaluation*. Thousand Oaks, CA: Sage.
- Pawa, D., G. Mundy, Y. Jittakoat & T. Nakpor (2013). How can HIV prevention programs reduce HIV risk behavior for transgender women in Thailand? [conference abstract]. Washington, D.C.: Population Services International. Available on KIX.
- Pawa, D., G. Mundy, Y. Jittakoat & S. Ratchasi (2013). Assessing the Impact of a HIV Prevention Program Among Transgender People in Thailand Using Coarsened Exact Matching [conference presentation]. Washington, D.C.: Population Services International. Available on KIX.
- Pawa, D., R. Firestone, S. Ratchasi, O. Dowling, Y. Jittakoat, A. Duke & G. Mundy (2013). Reducing HIV Risk among Transgender Women in Thailand: A Quasi-Experimental Evaluation of the Sisters Program. *PLoS ONE* 8(10), e77113.
- Piotrow, P.T., D.L. Kincaid, J.G. Rimon II, W. Rinehart (1997). *Health Communication: Lessons from Family Planning and Reproductive Health*. Westport, CT: Preager.
- Population Services International (2009). Log Frames for Proposals [web page]. Accessed 4 Feb 2015. Available on KIX.
- Population Services International (2013). M&E Indicators Database [web page]. Accessed 29 Jan 2015. Available on KIX.
- Population Services International (2013). Standard Operating Procedures [web page]. Accessed 4 Feb 2015. Available on KIX.
- Privitera, G.J. (2012). Introduction to Hypothesis Testing. In *Statistics for the Behavioral Sciences*, 225-260. Thousand Oaks, CA: Sage.
- Reynolds, Arthur J. (1998). Confirmatory Program Evaluation: A Method for Strengthening Causal Inference. *American Journal of Evaluation*, 19.2, 203-221.
- Schiavo, R. (2007). *Health Communication: From Theory to Practice*. San Francisco, CA: Jossey-Bass.
- Schlesselman, J.J. (1982). *Case Control Studies: Design, Conduct, Analysis*. (Volume 2). Ann Arbor, MI: Oxford University Press.
- Shadish, W.R., T.D. Cook & D.T. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Cengage Learning.
- Stanley, D. (2011) How to Plus or Minus: Understand and Calculate the Margin of Error [web page]. Research Access. Accessed 2 Feb 2015. Available online at: <http://researchaccess.com/2011/11/how-to-plus-or-minus-understand-and-calculate-the-margin-of-error/>
- Stuart, E.A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25.1, 1-21.
- Trochim, W.M.K. (2006). The Nonequivalent Groups Design [web page]. Research Methods Knowledge Base. Accessed 10 Dec 2014. Available online at: <http://www.socialresearchmethods.net/kb/quasnegd.php>
- Trochim, W.M.K. (2006). The Regression-Discontinuity Design [web page]. Research Methods Knowledge Base. Accessed 2 Feb 2015. Available online at: <http://www.socialresearchmethods.net/kb/quasird.php>
- United Nations (2008). *Designing Household Survey Samples: Practical Guidelines*. New York, NY: United Nations.
- Victora, C., G. Black & J. Bryce (2007). Learning from new initiatives in maternal and child health. *The Lancet* 370.9593, 1113-4.
- Warlick, J.L. (1981). Participation as a measure of program success. Institute for Research on Poverty Focus. Available online at: <http://www.irp.wisc.edu/publications/focus/pdfs/foc51d.pdf>
- Weiss, C.H. (1998). Purposes of Evaluation. In *Evaluation* (2nd edition), 20-45. Upper Saddle River, NJ: Prentice Hall.
- Windsor, R., T. Baranowski, N. Clark & G. Cutter (1994). *Evaluation of Health Promotion, Health Education, and Disease*

Prevention Programs (2nd edition). Mountain View, CA: Mayfield.

Wooldridge, J.M. (2009). Introductory Econometrics: A Modern Approach (4th edition). Mason, OH: Cengage Learning.

World Bank Group (2011). Evaluation Designs [web page]. Accessed 29 Jan 2015. Available online at www.web.worldbank.org.

Yanovitsky, I., E. Zanutto & R. Hornik (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning*, 28, 209-220.



POPULATION SERVICES INTERNATIONAL
1120 19TH ST., NW
WASHINGTON, DC 20036