WORKING PAPER

Impact Evaluations of Large-Scale Public Health Interventions

Experiences from the Field

Martha Priedeman Skiles
Aiko Hattori
Siân L. Curtis

August 2014

WP-14-157

USAID FROM THE AMERICAN PEOPLE
PEPFAR
MEASURE Evaluation

WORKING PAPER

Impact Evaluations of Large-Scale Public Health Interventions

Experiences from the Field

Martha Priedeman Skiles
Aiko Hattori
Siân L. Curtis

August 2014

MEASURE Evaluation
University of North Carolina at Chapel Hill
400 Meadowmont Village Circle, 3rd Floor
Chapel Hill, NC 27517 USA
Phone: +1 919-445-9350
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. Wp-14-157

USAID FROM THE AMERICAN PEOPLE
PEPFAR
MEASURE Evaluation

Executive Summary

Agencies are articulating the need to causally attribute health outcomes to investments in an era of shrinking resources and increasingly complex development environments. The opportunity to uncover essential information for program planning and resource allocation is a strong motivation for impact evaluations. Findings from large-scale IEs can be instrumental for decision-making, yet they are not without challenges and costs.

In this paper we share field experiences from a number of evaluation studies undertaken during MEASURE Evaluation Phases II and III. A series of case studies highlight design and implementation challenges that required creative solutions to move forward; plus analysis across studies revealed common reoccurring themes and valuable lessons. Examples of these cross-cutting themes presented include: challenges with identification and selection of program beneficiaries, random assignment in complex environments, identification of a robust comparison or control group for estimating the counterfactual, heterogeneity of program impacts, timing of baseline data collection, and absence of baseline data and a counterfactual.

Field experiences from MEASURE Evaluation Project demonstrate the need for transparency and collaboration among the key partners, the inevitable balancing of technical requirements with programmatic priorities, and the flexibility required to adapt designs in order to answer the most valuable evaluation questions. Interest in accountability of funding of public health interventions continues to grow, promising continued interest in IEs. Evaluators, implementers and funders can share in these learnings as we move forward with expanding our understanding of the costs and benefits for rigorous evaluations.

Introduction

The demand for impact evaluations by the development community has increased dramatically over the past decade, particularly following the 2006 call to action by the Evaluation Gap Working Group, "*When will we ever learn?: Improving lives through impact evaluation.*"(1) This call to action challenged policy makers, funders, implementers and researchers alike to create an environment that values evaluation as a public good, to work collaboratively to improve the quality and utility of evaluations, and to place a premium on impact evaluations. While the methods and use of impact evaluations are well documented, the Working Group brought to the forefront the value gained by rigorous evaluation to answer pressing questions of program impact. Following the establishment of the *International Initiative for Impact Evaluations* (3ie) which was created to share knowledge and resources on impact evaluation, the commitment to rigorous evaluations is now a key feature in many proposal requests and program designs.

Agencies are articulating the need to causally attribute health outcomes to investments in an era of shrinking resources and increasingly complex development environments.(2-4) Well-designed and executed impact evaluations (IEs) provide valuable information from a program and policy perspective. While performance monitoring helps identify whether or not a change occurred and performance evaluation may answer questions of targeting and implementation, impact evaluations "measure the change in an outcome that is attributable to a defined intervention by comparing actual impact to what would have happened in the absence of the intervention (*the counterfactual scenario*)."(5, p2) This quantification of a change with causal attribution is fundamental to decision-making regarding programming and resource allocation. However, the requirements for a rigorous IE are high, setting IEs apart from other monitoring and evaluation activities on a number of fronts.

A robust estimation of a counterfactual is required to support the evaluation hypothesis that a change in outcome among participants is due to the program and not to other factors, observed and unobserved, that might account for the observed change.(4) While there are a number of recognized approaches for estimation of the counterfactual, the chosen estimation method will have implications for the study design, data needs, and resource requirements.

Complementing the counterfactual is the requirement for independent baseline data and a window of time over which change occurred. Not all estimation strategies strictly require baseline data; however, baseline data may strengthen the claim of attribution.

Theoretical knowledge along with guidance exist to steer researchers and decision-makers along the path of evaluation (6-10), and new methods and designs have been proposed to address the shortcomings of narrowly focused evaluations (11), yet few have shared the challenges or proposed solutions for some of the common barriers encountered when implementing large-scale impact evaluations.(12-15) Furthermore, the evaluation of HIV/AIDS-related efforts pose additional challenges due to the sensitivities and scope of the HIV/AIDS epidemic and the populations affected. HIV/AIDS programs provide complex behavioral and biomedical interventions among vulnerable and sometimes criminalized populations, often with the motivation for rapid national implementation due to the medical and political implications of delaying services.(3) The objective of this paper is to transform tacit knowledge gained through field experience into explicit knowledge to aid future efforts for impact evaluations in general as well as considerations for HIV/AIDS–related IEs that impose additional constraints.

This working paper will draw on practical experiences from the MEASURE Evaluation Project to illustrate real world challenges for large-scale impact evaluations, showcasing some of the solutions employed to design and conduct IEs for public health in general and HIV/AIDS in particular. Following a brief summary of the case study methodology employed, a number of field situations will be presented alongside examples from MEASURE Evaluation studies that detail challenges and highlight solutions for strengthening the design and implementation of rigorous impact evaluations.

Methods

Over the course of MEASURE Evaluation Phase II (2003-2008) and Phase III (2009-2014), a number of evaluations have been requested, designed, and executed. In an effort to understand the lessons learned from these experiences, ten evaluations were selected for review. Criteria for the selection of evaluations included impact-oriented evaluations, a range of technical areas, and a range of evaluation designs. For each evaluation, the authors interviewed the MEASURE Evaluation Project study personnel regarding the evaluation objectives, design and outcome. Interview questions reflected our interest in challenges and successes encountered during the planning and implementation of these evaluations (see Interview Guide, Appendix A). Review of study reports and papers in peer-reviewed and grey literature provided additional context and details for each case study. The ten case studies are listed in Table 1 with abstracts for each found in Appendix B. Some of these evaluations have been completed

while others are works in-progress; all provide useful information for those interested in impact evaluations.

Table 1. MEASURE Evaluation Case Studies

Evaluation	Country	Status
Impact Evaluation of the NGO Health Service Delivery Project (NHSDP)	Bangladesh	On-going
Impact Evaluation of the Bangladesh Smiling Sun Franchise Program (BSSFP)	Bangladesh	Completed 2012
Early Marriage Evaluation Study (EMES)	Ethiopia	Completed 2009
Evaluation Plan for the Ghana National Strategy for Key Populations	Ghana	On-going
Impact Evaluation of the Western Highlands Integrated Program (WHIP)	Guatemala	On-going
Impact Evaluation of the Kingston Priorities for Local AIDS Control Efforts Intervention	Jamaica	Completed 2009
Evaluation of the Community Care for Vulnerable Children in an Integrated Vulnerable children and Home-Based Care Program	Mozambique	Completed 2014
Impact Evaluation of the SUAAHARA-GPM Nepal Program	Nepal	On-going
Impact Evaluation of Malaria Control Interventions on Mortality in Children in Mainland Tanzania	Tanzania	Completed 2012
Impact Evaluation of the Strengthening Tuberculosis Control in Ukraine (STbCU) Project	Ukraine	On-going

Findings

Each case study identified study-specific design and implementation challenges that required creative solutions to move forward; and analysis across studies revealed common reoccurring themes that provide valuable lessons. The remainder of the paper will focus on these cross-cutting themes, including challenges with identification and selection of program beneficiaries, random assignment in complex environments, identification of a robust comparison or control group for estimating the counterfactual, heterogeneity of program impacts, timing of baseline data collection, and absence of baseline data and

a counterfactual. Examples from the case studies in Appendix B are presented to illustrate these themes.

Identification and Selection of Program Beneficiaries

Identifying program beneficiaries is required to measure impact among the exposed; however it is not always straightforward. One challenge may be that the sampling frame is incomplete due to the uncertainties in the future scale-up of the program. With a newly awarded program that has not yet finalized geographic or demographic targets, or has not yet enrolled beneficiaries, the identification of beneficiaries is still emerging. A different challenge is faced when a static target population is defined, yet the program may not have records suitable for the development of a sample frame for an evaluation. The following MEASURE Evaluation Project evaluations illustrate some of the hurdles faced when creating the beneficiary sampling frame.

In Bangladesh, the IE of the NGO Health Service Delivery Project (NHSDP) was tasked with measuring the impact of the project on the use of selected maternal and child health services in areas that were identified by the government as having inadequate public health service delivery systems. NHSDP operates in all districts of the country, but the project catchment areas do not cover entire districts; other parts of the districts typically receive services from government fieldworkers and clinics. At issue was how best to translate the catchment areas of the clinics served by NHSDP into defined geographic areas with associated target populations to serve as the sampling frame for the evaluation. Fortunately, each fixed and satellite clinic in Bangladesh has NGO-administrative data that tracks the number of eligible couples by catchment area. These data were mapped to defined geographic areas for the sampling frame. The catchment areas for the program-supported clinics were considered to be the total target population from which a random sample was selected for the evaluation. This catchment area data for the program-supported clinics was made available from the implementing NGOs and served as the sampling frame for the study. The sampling frame was assumed to be essentially constant, or close to constant, over time but during the household listing exercise it was found that this assumption did not completely hold.

The facility catchment areas change over time as new sites are added and old ones close down once deemed no longer needed or viable. Adjustments to clinic lists were particularly relevant for the satellite clinics, reflecting the evolving needs of the populations. Essentially the intervention population was not static but for the purposes of the evaluation, a frame at a given point in time had to be finalized.

Many of the planned expansion sites were not included in the frame because they were not identified at the beginning of the project when baseline surveying was completed; however, about 80% or more of the sites were consistent over time. To manage the evolving program clinic list, protocols were in place for review and documentation of resolution for each case as it was encountered in the field. Most of the types of cases were encountered in the first 2-3 weeks of the listing operation, allowing the team to apply consistent, documented rules when similar cases came up later in the listing operation. For the analysis, interpretation and reporting for NHSDP, findings must clearly state that the evaluation covers only sites in place at the time the frame was constructed and exclude sites added later. Some loss to follow-up may occur in the endline survey if some of the satellite clinics originally sampled close or move by the end of the project.

A similar issue in a different context was encountered when conducting the baseline survey for the Western Highlands Integrated Program (WHIP) Evaluation in Guatemala. Among other activities, WHIP integrates agriculture with health and nutrition initiatives designed to decrease poverty and malnutrition in the Western Highlands. For this IE, construction of the sampling frame required matching agricultural producers' association members, who were intended beneficiaries of the agriculture intervention, with the census enumeration areas where they lived. However, the survey coincided by design with the early stages of implementing the program's agricultural component, and during study planning many intended beneficiaries had yet to be identified. Unlike the situation in Bangladesh, the farmers' associations tended to maintain different information about their beneficiaries than that required for the survey sample. As a consequence, study teams spent considerable time working with the program's implementing partners to list and map beneficiaries in participating communities. The extra work required on the part of both the study team and the program implementers added several months on to the design phase, with associated budget implications.

Mozambique provides another example of program records that were not suitable for sampling purposes. The Community Care Program is a USAID/Mozambique bilateral project focused on community-based response to HIV/AIDS. Recently the program changed their care model by cross-training health workers to offer dual services, that is services to both orphans and vulnerable children (OVC) and to clients of home-based care (HBC) services, rather than having separate workers address each population. USAID/Mozambique was interested in the effect of this change to dual programming, specifically on services provided to OVC. To evaluate the program it was necessary to identify those households receiving OVC and/or HBC services, yet incomplete family registration forms required

manual linking of forms with household registers; a long error-prone process that may have missed some program participants.

In theory, an evaluation implicitly assumes that the intervention population is well-defined and static but in practice the targeting of beneficiaries may be evolving. Translating the dynamic nature of the program target population into a frame that can be used for sampling requires problem-solving specific to individual situations.

Random Assignment in Complex Environments

A well designed randomized control trial (RCT) may alleviate the challenge of identifying the intervention and control populations if individuals or communities are randomly assigned ex ante to intervention and control groups. Furthermore randomized assignment will theoretically render selection bias and confounding inconsequential in the final impact assessment, leading many to recommend the RCT as a design that provides the strongest evidence for causal associations.(4) However, conducting an RCT in real-world conditions is often infeasible due to issues of scale and competing priorities.

USAID/Nepal was very interested in an impact evaluation to measure the effect of an array of capacity building interventions with Health Facility Operations and Management Committees (HFOMCs) under the SUAAHARA-GPM Program. Specifically they were interested in the effect of the program on the equitable use of maternal and child health services as well as the quality of care provided to women and disadvantaged groups seeking maternal and child nutrition and health services. Random assignment of communities to three arms – control arm, standard HFOMC capacity building activities, and HFOMCs receiving additional gender and social equality integration interventions – was suggested to control for selection bias due to program targeting. Subsequent negotiations over the study design raised several concerns and illustrated the often inherent tension between the objectives of a program and the objectives of an evaluation. First the project prime wanted to target the new interventions to a strategically located region already identified as having the staffing capacity to implement the program; programmatically this was the best option but meant that randomization of districts was not possible nor would it reflect how such a program would be implemented in practice. The next option considered was randomization at a smaller geographic unit; however the Mission and Government of Nepal were interested in understanding the scalability of the program to entire districts. Hence results based on implementation at the sub-district were not considered helpful. Lastly, the implementers were

concerned that randomization at the smaller geographic scale would increase the burden of implementation beyond their staffing capacity; the program technical approach would vary from community to community and this would require greater organization and tracking than a blanket approach across a district. The project's overriding goal was to meet performance targets in a large district-wide effort while the evaluators were focused on selecting comparable intervention and control communities. In this scenario, the RCT was not compatible with the program intentions and therefore rejected in favor of an evaluation of district-level implementation using a difference-in-differences estimation strategy.

The impact evaluation of the Priorities for Local AIDS Control Efforts (PLACE) intervention in Jamaica highlights some of the favorable conditions for implementing an RCT in real-world situations. Findings from a PLACE survey conducted in 2001 identified over 400 venues in Kingston where new sexual partners are met. The Ministry of Health, led by the director of the HIV/AIDS program, designed the PLACE intervention strategy to provide prevention activities at these venues. This same director of the HIV/AIDS program had a keen interest in understanding the causal effect of the PLACE intervention on safe sex behavior and had an appreciation for the advantages of an RCT. As the leader for the development of the evaluation plan as well as the intervention program, the director made it possible to build the intervention around the evaluation design without prioritizing performance targets over evaluation needs. Under his leadership the committee agreed to a randomized assignment of the potential venues to intervention and control groups. Following randomization, pooled cross-sections of patrons were surveyed pre- and post-intervention to compare the proportion of patrons in intervention and control venues who reported new or concurrent partnerships and recent inconsistent condom use. The intention was to produce balanced groups that controlled for selection bias and other confounding factors.

Identifying a Robust Comparison/Control group for Estimating the Counterfactual

A robust comparison/control group for estimating a counterfactual lends credibility to the estimate of program impact by defending a study from threats to internal validity. The identification of a valid comparison/control group will be influenced by perceived threats to internal validity and potential options for controlling these threats.

Selection bias is a common challenge to internal validity. Typically there are two selection processes – the program selection or targeting to certain populations, and the self-selection by an individual to

participate in a program. If one or both of these selection processes are in effect then participants will very likely be different than non-participants. Moreover, if factors that influence both the outcome and the program participation are not controlled for then the estimation of program effect may be biased.

The Strengthening Tuberculosis Control in Ukraine (STbCU) project goal is to decrease the TB burden in Ukraine, leading to a reduction of TB morbidity and mortality. One of the IE evaluation questions seeks to measure the impact of the patient social support program on improving TB treatment adherence and the subsequent treatment outcome. The study hypothesis is that patients at high-risk for defaulting on TB treatment who receive social support services will improve treatment adherence. Concerns about selection bias motivated the study design and identification of the comparison group.

Over the past decade, USAID provided support for TB prevention and control activities in 10 regions of Ukraine. The selection criteria for the target areas included high TB and HIV disease burden, inadequate TB treatment services, geographic location clustered in the eastern and southern regions, concentration of vulnerable populations, distribution of other NGO operations, and desire of local government officials to participate. Essentially, prolonged TB aid was prioritized to regions with the highest disease burden and some of the poorest services. This program targeting at a regional level made it very difficult to claim comparability with other regions which differ on observed and unobserved factors from the intervention area. Selecting a prospective comparison group within the regions was considered; however the operating assumption was that the program would be offered region-wide to all eligible high-risk patients with very low refusal rates.

The second selection process was the physician referral for services. The social support program relies on physician adherence to a documented referral protocol for patients needing support. Compliance with the protocol is left to each facility to enforce. Controlling for provider- or clinic-specific patient selection added another challenge to the estimation of a counterfactual.

The history of the social support program in Ukraine offered the best option. The social support program was developed and piloted in 2010, a break in services occurred in 2011 for all sites, then activities resumed in 2012. Given this information, a quasi-experimental design was chosen, using retrospective medical records and prospective interviews in the same clinics over time. Data from three time periods will be collected: baseline pre-program data from 2011, data following the re-introduction of services in 2012, and prospective data collection of new cases is proposed in 2015-16. The advantage of returning to the same facilities is the ability to control for unobserved facility and regional

characteristics that may be correlated with the selection of the site and the outcome. Within each facility, the referral protocol is used to guide the sampling of different patient populations including high-risk patients exposed and unexposed at baseline and midline as well as low-risk patients unexposed at both periods. Sampling the different clinic populations and collecting their risk-profile data will facilitate the control of provider targeting based on risk factors. Self-selection by the patient upon referral was considered negligible because of the high acceptance rates tracked by the program.

*Spillover effect*s occur when the intervention has an impact on individuals not in the intervention group with the potential to bias estimates of program impact. Unlike selection bias, spillovers cannot necessarily be controlled for through random assignment. In the Jamaica PLACE intervention evaluation, social venues were enrolled as intervention or control sites and repeat cross-sectional surveys of patrons conducted. Patrons interviewed at baseline were free to travel and visit different sites. It was anticipated from the outset that there would be mixing between patrons visiting the intervention sites with patrons of the control sites; the level of exposure to the intervention varied by venues patrons visited. The evaluation team preferred a site-based cohort design with prospective assignment to intervention and control venue, however it was ruled out due to difficulties anticipated in identifying, recruiting, enrolling and tracking site-based cohorts. Efforts were made to separate sites geographically in order to limit mixing and this mixing of populations was monitored; however it was a noted study limitation.

In a small area, distancing intervention and control sites from each other to avoid spillovers may not be feasible; however, even in larger geographic areas, balancing the risk of spillover and the risk of population differences potentially confounding results is difficult. It is not always possible or preferable to select comparison sites at great distance from the intervention sites. For example, in Guatemala the selection of comparison sites was limited to the same departamentos in the Western Highlands as the intervention sites in order to improve the comparability of agricultural environment among the groups. Likewise in Nepal, the three target districts were all selected from the same region to assure comparable ecological environments, socio-economic characteristics, and existing level of SUAAHARA-GPM program activities in all three areas.

While spillover effects are often considered a spatial issue, there may also be spillovers due to program implementation. In the Ukraine social support evaluation, fidelity to program eligibility posed a greater risk to spillover than geographic selection. Patient exposure to the social support program was

dependent on physician referral according to established criteria. During data collection it was determined that some clinics did not follow the prescribed protocol, in fact referrals were based on patient compliance, with those patients demonstrating the best compliance being rewarded with referral to the social support program. This referral process led to spillover of social support to more compliant patients or to those at lower risk for defaulting on treatment, potentially diluting the measure of impact among the high-risk patients. To better understand the referral process and estimate the impact, multiple groups were sampled to allow comparisons between high-risk intervention and comparison patients; between high-risk intervention and low-risk comparison patients; and between low-risk patients across time. Collecting information on risk criteria for all subjects will allow us to adjust for the risk profile in our final analysis.

Contamination is another threat that may bias evaluation findings if members of the comparison group are exposed to another intervention which affects the same outcome. In most countries the development landscape is complex with multiple donors and programs targeting the same or similar health outcomes in overlapping geographic spheres.(16) This may lead researchers to try and balance the effect of these other programs across the intervention and comparison groups, to identify and collect data on all potential confounding programs in an attempt to control for them in analysis, or to step back from specific program attribution and take a broader look at the entire scope of efforts underway that may influence the chosen health outcomes. An example from Ghana provides one solution to managing potential contamination.

Development of the Ghana HIV/AIDS evaluation plan was a national evaluation effort undertaken in Ghana by the Ghana AIDS Commission (GAC). The goal of the activity was to develop a national HIV/AIDS evaluation plan following the UNAIDS-MERG guidance.(3) The evaluation objectives focused explicitly on measuring change in outcomes among men who have sex with men (MSM) and female sex workers (FSW). Recognizing that donors offer complex interventions targeting different activities to different populations, that HIV/AIDS prevention programs are not offered in isolation (i.e., other safety net programs may be at work), and that often multiple projects are working on HIV/AIDS prevention with the same or overlapping populations, it was agreed that attribution to specific donors or programs would be untenable. This did not eliminate the need to identify inputs and program exposure, but it did help the GAC collectively agree on definitions such as “program reach” to assure that programs were reporting activity outcomes using the same frame of reference. This activity reporting coupled with existing baseline behavioral survey data and planned endline survey data collection, and a performance

evaluation of various programs, will provide useful information that is both affordable and feasible to implement, but stops short of quantifying impact of any specific program.

Heterogeneity of Program Impacts

Identification and selection of intervention and comparison groups may be further influenced by interest in outcomes among specific subpopulations. If the conceptual framework suggests that program effects are likely to vary across sub-populations and differences in effects are of interest from a program or policy perspective, then measuring impact for key populations of interest is imperative. This will typically require larger sample sizes for an IE which may require balancing cost considerations with the value of sub-population estimates. In the case of the NHSDP evaluation, an interest in equity of effects prompted the request to measure program impact across different wealth or socio-economic groups. However, the required sample size to measure effect across sub-groups was prohibitively expensive. Instead the sample size was powered on differences for the whole population, with some sample size inflation to improve precision of the outcome estimates within subgroups.

Heterogeneity of impacts may also be anticipated based on the implementation vehicle of a program. For example, in Ukraine one of the key objectives is to improve the integration of TB and HIV/AIDS services to improve identification and treatment for co-infected patients. Historically, the management, provision, and monitoring of these services have been vertically implemented. With this new effort towards integration, it would be interesting to differentiate progress made by each health service. Powering by clinic population (TB and HIV/AIDS) was cost-prohibitive however. Instead, sampling was stratified by service point-of-entry to enable estimates by service type. Moreover, a descriptive analysis depicting the cascade of services from each point-of-entry was designed to provide some information from a clinic-specific service perspective.

Timing of Baseline Data Collection

Almost all of the estimation strategies for IEs require or will benefit greatly from clean baseline data with an adequate window of time during which an intervention can produce a measurable impact. The best practice recommendation is to identify impact evaluation opportunities before program start-up and integrate the evaluation design into the program planning. (1,4) Early initiation provides the best opportunity for coordinated planning yet even with forethought, integration and rapid baseline data collection cannot be assured .

In the case of the Ukraine IE, the evaluation team was brought into the process soon after the program award was made. Despite the full cooperation of the implementing agency, it was clear after the first in-country visit that the program was still developing the monitoring plan, finalizing indicators, and planning activities. This left significant gaps in the evaluation design. For example, targets for the outcome indicators were not set and relevant details of the activities in the program workplan remained unclear. The study design phase extended over one year, at which time the opportunity for a real-time clean baseline data collection was lost. Fortunately because this was an intervention targeting a well-defined clinical population, the availability of patient record data allowed us to create a retrospective baseline during a period when the social support program was not active.

Despite early involvement and excellent collaboration between implementers and evaluators, baseline data collection can be delayed by changes in program design in response to emerging evidence. In Nepal, the evaluation team and stakeholders planned to evaluate a three-way comparison among different levels of capacity building and training with the HFOMCs (described earlier). After the initial IE design phase and initiation of the evaluation work, the Government of Nepal (GoN) revised its basic capacity building curriculum for HFOMCs to incorporate the Health Sector Gender Equality and Social Inclusion (GESI) strategy. GESI was designed to provide a framework for integrating gender equality and social inclusion into the health sector to assure quality health service access for all. Lessons learned from recent GESI work prompted the GoN to implement GESI improvements immediately, notwithstanding the HFOMC evaluation design. This effectively eliminated the the basic intervention arm of the study and required changing the comparison groups and the evaluation questions to fit the new model. Programmatically this change allowed for incorporating and acting on new, emergent findings for improved services; however, this study design change was one factor in a series of delays that effectively reduced the window of change from two years to 1-1.5 years from baseline to endline. Despite this setback, communication and collaboration among the evaluators, implementers and funders allowed for a design change to improve program implementation and balance the needs of the key players.

Sometimes the interest in an IE surfaces after the opportunity for a clean baseline has passed. In the case of Mozambique, the request for an IE came after the program was initiated. Given the project was already underway, there was no practical means of collecting clean baseline data because newly trained workers were already deployed in all of the program provinces. In reviewing the program status and the evaluation interests of USAID, a concern about heterogeneity in program execution was raised due to

implementation of the new dual service approach for the Community Care Program by multiple NGOs. Additionally there was concern that integration might result in less attention paid to OVC services if workers shifted their efforts to sick HBC patients. An evaluation of the program implementation would provide useful information to answer these questions at a lower cost and in a timely fashion. This serves as an example of resetting evaluation expectations to focus an appropriate evaluation to answer the primary questions of interest. In this case understanding the differences in performance across varying implementation vehicles was more important than measuring impact. A performance evaluation was better suited to providing useful information.

When a Comparison/Control and Baseline Data are Unavailable

We covered a number of IE examples that did not follow the textbook path of evaluation yet solutions were found to adapt the evaluation design in response to development environments. But what if a valid comparison/control group for estimating a counterfactual is just not available and baseline data collection is impossible? One option is to review the interests of the evaluation solicitors and determine if there is a different question that would be relevant and useful given the evaluation conditions.

The Early Marriage Evaluation Study (EMES, 2007) illustrates how to adjust design features when the feasibility of a rigorous IE is compromised by field realities. In 2006, USAID/Washington requested MEASURE Evaluation to evaluate the effects of child marriage prevention activities on reproductive health, unintended pregnancy, physical and sexual violence and mental health. Ethiopia was identified as a country of particular interest in this regard; however there was no formal child marriage (CM) policy or CM-specific agenda in Ethiopia. The existing Mission-funded reproductive health and education projects were not originally conceptualized as CM prevention projects. Rather CM-related programming was added to these projects during their second year once CM was identified as part of the causal pathway leading towards improved health and education outcomes. While there was keen interest from USAID/Washington and USAID/Ethiopia to support an IE of these CM-related activities, this misalignment between the projects' key outcomes and the IE objectives was a concern.

A second alignment challenge was that between the timing of the activities and the proposed evaluation. The interest in evaluation these intervention came three years after CM activities were added to the existing projects, essentially eliminating the possibility of clean baseline data. Moreover, the identification of a robust comparison group to control for selection bias, contamination and potential spillovers was not possible.

Engaging stakeholders in the evaluation design was the key to moving forward on this evaluation. The evaluation team met virtually and in-person with USAID/Washington, USAID/Ethiopia, local implementing partners, and in-country government and NGO stakeholders to collectively articulate the primary research questions. This work was followed by field visits with implementing partners to gain an understanding of the reproductive health and education projects and how they fit together with CM-related activities. Additionally perceptions in the field from community members (e.g., parents, fistula patients, priests, clinic staff, project coordinators, justice department representatives, government officials, etc.) deepened the team's appreciation for the child marriage situation in Ethiopia. Using this information, the evaluation team along with the collaborators mapped the CM activities to create a logic model. This exercise was instrumental in identifying parents as the key decision-makers to target for the evaluation.

With a better understanding of the projects and the interests of the stakeholders in mind, the evaluation team proposed a study to document the CM context in Ethiopia, the prevalence, determinants and consequences of CM; determine level of exposure to CM prevention activities and their association with knowledge, attitudes, and skills needed to forestall marriage; document the process of early marriage cancellation; and identify factors that shaped health and social outcomes for girls whose marriages were cancelled. Additionally the team was able to identify approximately 20% of the districts as free from the intervention which allowed the team to compare outcome measures in program and non-program areas. However, the study stopped short of quantifying and attributing program impacts due to the lack of baseline data and known selection effects that limited the ability of the non-program area to serve as a strong counterfactual.

The final study was well received by professional stakeholders as well as grassroots organizations and community members with an interest in understanding the role of CM on health and education outcomes. Useful knowledge about CM and the project activities was gained and shared to assist programs in their project designs and to aid funders in program planning.

A multi-agency evaluation in Mainland Tanzania, commissioned by the Roll Back Malaria (RBM) Partnership and supported by the President's Malaria Initiative (PMI), provides another example of adapting the evaluation design to accommodate the absence of baseline data for a program national in scope.

A rapid increase in targeted funding for malaria control in Mainland Tanzania was associated with a significant increase in uptake of recommended interventions from 2000 to 2010. The The National Malaria Control Program (NMCP) scaled-up insecticide treated bednets (ITNs) reaching national coverage by 2006. Additional efforts, including case-management with artemisinin combination therapies (ACTs) and intermittent preventive treatment in pregnancy (IPTp), reached national coverage and responded to changes in WHO recommendations along the way. Because these efforts were national in scope, the identification of a suitable comparison group was impossible. Moreover, the request for a large-scale evaluation came well after efforts were implemented, making primary baseline data collection infeasible. Instead The study adopted a plausibility evaluation design based on the underlying program impact pathway,(17) sometimes referred to as a theoretical design.(11) They used a before-versus-after model (i.e. single-difference model) to assess the secular changes in the outcomes of interest and describe the programmatic efforts that plausibly influenced these changes..

The evaluators relied on the Rollback Malaria's Monitoring and Evaluation Reference Group (MERG) causal framework and key indicators to track along that framework. Secondary data sources collected over several years and from multiple sources were used to analyze trends in intervention coverage and outcomes. In addition, published studies were reviewed to understand the estimated impact of various malaria interventions to support plausibility argument.

This secondary analysis was not intended to critique program implementation or effectiveness, nor was it expected to attribute outcomes to select interventions or specific funders, rather it was designed to look at the malaria control activities as a whole and assess plausible impact on malaria morbidity (malaria parasitemia and severe anemia) and all-cause mortality for children under five years of age. Analyses examined the spatial, temporal and dose-response relationships between interventions and expected outcomes. Contextual factors such as education, health services, and environment, added depth to the analysis and understanding of the potential mechanisms for change. Many other interventions target health among the under-five, these other proximate determinants to health, including immunizations, water and sanitation improvements, and nutrition to name a few, were reviewed as well to understand the role each may have played in reducing all-cause mortality. Using the Lives Saved Tool (LIST), the evaluators modeled the expected contributions of various health interventions (including but not limited to malaria control) to estimate changes in mortality of children between 1999 and 2010. In total, an estimated 45% drop in under-five mortality plus a 50% drop in anemia during these same years when ITN use spiked dramatically, all lend credibility to the claim that

malaria control interventions reduced malaria-related morbidity and mortality in Tanzania from 1999 to 2010.

Discussion

A growing body of literature on impact evaluations covers theoretical and methodological considerations as well as findings from IEs undertaken in a variety of disciplines and settings. This literature is a great resource for understanding the value of an IE and the methods available for handling various technical issues; however, when applying this theory in the context of real programs operating at some degree of scale, IEs are fraught with practical design challenges not often covered in the literature.

On the technical side, the featured case studies noted challenges with identifying the intervention and comparison populations, as well as availability of clean baseline data. Creation of a sampling frame for intervention or program beneficiary populations was a notable difficulty for programs targeting participants spatially, such as identifying the NHSDP clinic-based catchment populations in Bangladesh, and programmatically such as identifying farmers participating in the WHIP value chain program in Guatemala. In the case of NHSDP, administrative data on eligible catchment populations was available but the catchment area boundaries were continually adjusting for programmatic reasons. This required the IE to select one set of boundaries at one point in time, and use those defined populations for the sampling frame. In Guatemala, the administrative beneficiary data had not been updated and the information tracked for program implementation purposes were not adequate for evaluation purposes, making it problematic to identify the VC beneficiaries.

Identification of a suitable control or comparison group with which to estimate the counterfactual presented its own set of challenges. Randomized assignment alleviated concerns of confounding and selection bias in Jamaica but did not eliminate spillover effects because random assignment was at the venue-level rather than the individual-level and mixing of individuals was unavoidable. Using a historical comparison group enabled controlling for selection bias by region and facility in Ukraine; however, potential spillover due to degree of referral fidelity may prove more difficult to eliminate in this social support study. Contamination is a threat with many public health impact evaluations. Not only are there often multiple players addressing the same issues, such as found in the Ghana evaluation of

HIV/AIDS prevention efforts, but there are also different initiatives that target the same health outcomes (e.g., child mortality) but through different channels, as noted in the Tanzania-PMI evaluation. In each of the case studies, the design choices made reflect a balancing of evaluation objectives, field limitations, and threats to internal validity, requiring accommodations to best match designs to objectives.

When the request for an IE comes after a project has begun different evaluation options may need to be explored, as seen in Mozambique and Ethiopia. Some IE requests come in parallel with a project award as recommended by evaluation specialists,(1) although delays in baseline data collection may still occur depending on the time required to develop the details of the intervention which are needed for the evaluation design, as demonstrated in Nepal and Ukraine. However, delays in implementing the baseline data collection associated with time taken for planning at the project level also delay actual project implementation, potentially offsetting concerns that the delayed baseline will be contaminated by early project effects. Even so, the reduced window of opportunity for the implementation to have an effect may have negative consequences for detecting change or necessitate increasing sample sizes to detect smaller changes. Collecting independent baseline data is particularly challenging when programs are continuations of long-term interventions that were initiated well before interest in an impact evaluation gained support. Study design decisions are often dictated by the availability of baseline data in this context.

Operationalizing these technical considerations may be facilitated by the alignment or hindered by the misalignment between the objectives of funders, implementers and evaluators. Inherent tensions often exist between program objectives and evaluation objectives. Programmatically it may be better for program implementers to scale up to multiple sites purposively selected to maximize performance targets. Yet this targeting strategy typically increases risk of selection bias for the evaluation. Interventions with limited geographic or population reach may provide for strong comparison groups in an evaluation, but may not be entertained politically due to interests in understanding a program at scale, such as the case in Nepal. Evaluators want to minimize selection bias and identify robust comparison groups, thereby bolstering claims of attribution and unbiased estimations.

Aligning the program and evaluation objectives may require some level of compromise or creative design to accommodate these various valid priorities. One solution often sought is a phased implementation to allow for a comparison group receiving the intervention at a later date. Some

projects may prefer a phased roll-out yet may be unwilling to abdicate the right to adjust their phase-in schedule as new recommendations are made or additional resources are available to address the health outcomes pursued. Evaluators, however, must maintain some control over research designs that are fundamentally tied to preservation of a robust comparison group.

A second pragmatic consideration is the decision to undertake an IE or opt instead for a different type of evaluation or study depending on the environment, the interests and the resources. As noted by Farley and colleagues, "Impact evaluations are an essential tool for learning and for accountability but are not the right tool for every project."(14) IEs are expensive, labor intensive, and in most cases require a time commitment of several years before final results are available. Choosing an IE should be motivated by the type of evaluation question posed and the intended use of the findings, plus the degree of attribution required to satisfactorily answer the question and make decisions. Considerations of the feasibility of an IE are also important; for example is it possible to identify a suitable comparison group and is there an opportunity for clean baseline data collection. If these two criteria are not met, then an alternate evaluation plan may be required. Lastly the potential for valuable findings must be balanced with the cost of an IE.

Recommendations

1. Alignment between funders, implementers and evaluators is critical to a well-designed evaluation that complements the program.

Many of the challenges detailed above can be overcome or at least improved upon with open and frank communication between funders, implementers and evaluators. Discussions regarding the key players' interest and commitment to the evaluation goals, accommodations required for the evaluation and for the program, and realistic assessment of the feasibility of an IE are topics that must be explored fully as equal partners in the process. We must avoid equating an independent evaluation with limits to negotiation and cooperation among parties. Without full participation of the funder, the implementer, and the evaluator, development of evaluation designs may fall short of expectations or create unnecessary delays or complications in moving forward in a timely and productive fashion. In the Jamaica PLACE evaluation, the government requestor of the evaluation was also heavily invested in the design of the program. Evaluation needs were clearly articulated and the decision for an RCT was

informed both by the program and the evaluation needs. In both Nepal and Ethiopia, challenges to the evaluation design were assessed and reconciled through extensive discussions between all the key stakeholders with an eye to creating solutions that would be useful and practical.

Clear expectations about cooperation and transparency between the program and evaluation need to be established by the funders, yet requirements need to be tempered by recognition that in some cases availability of information critical to the evaluation planning may be beyond the control of the program. One recommendation from the WHIP evaluation was to improve the alignment between contractual specifications for the implementing partner and the external evaluators. Had the contractual language required collaboration and sharing of data between the implementers and evaluators, then data system capacity may have been evaluated early on, providing more time to reconcile data needs with data availability.(18) There has been a noted shift in project procurements that specify the intention to award an independent evaluation award. Yet often these documents lack specificity requiring joint planning. Contractual language that clearly indicates the expectations of collaboration and prioritizes transparent dialogue to facilitate meeting both program and evaluation objects may facilitate cooperation moving forward.

2. Creative adaptation of the evaluation design to fit operations is the norm not the outlier.

Designing evaluations requires flexibility, creativity and judgment. Rarely do IEs in real world situations follow the textbook design; rather identifying operational challenges and options for overcoming these challenges is critical to the success of IEs in these circumstances. Evaluating the TB social support program in Ukraine required retrospective data collection to find clean baseline data and pilot testing revealed potential spillovers requiring diversification of the groups sampled to assure that sufficient data was collected to control for variable fidelity to the referral protocols. In Ghana the national scope of the program and the multiple interventions and implementing partners meant that attribution to program-specific interventions was not feasible. Instead, working with partners, the evaluation team designed a plausibility study to draw on numerous data sources to construct a conceptual framework to map inputs with outcomes. Crafting these evaluation solutions takes time and clear communication with parties involved to assure that the solutions are the best available given resource constraints and evaluation objectives.

3. Flexibility to entertain other evaluation solutions, finding the right tool for the task at hand.

Impact evaluations can provide valuable information not produced by other evaluation designs; however they are not without cost. Balancing the information needs with the feasibility and costs of evaluations requires judicious selection of IEs. Judgment on the part of the evaluators coupled with candid discussions of potential limitations and alternatives may reveal that compromises necessary to accommodate the real world challenges undermine the utility of results. Sometimes a wise move forward is a reconsideration of the best tool to produce actionable results. As seen in Mozambique, the ability to collect baseline data from a suitable intervention and comparison group were compromised. Yet discussions regarding the information sought provided an opening to move forward on a performance evaluation to determine whether the new dual-purposed staffing plan would meet the service expectations of the population at risk. Understanding this initial piece of the puzzle will lay the groundwork for a future IE if and when the stakeholders determine the need for one.

In the case of the malaria impact evaluation in Mainland Tanzania, the evaluation goals were well articulated but the feasibility of a traditional impact evaluation was untenable given the national scope of the already well-established program interventions. Since the primary interest was in attribution of changes in malaria morbidity and mortality to national strategies rather than individual implementers, a plausibility design was proposed. This design, while not without limitations, mapped out clear causal pathways from inputs to outcomes and drew on multiple data sources and previous intervention-specific evaluations to build a body of evidence to support or refute the proposed causal attribution of increased inputs to improved health outcomes.

Conclusions

The opportunity to uncover essential information for program planning and resource allocation is a strong motivation for impact evaluations in public health. Findings from these large-scale IEs can be instrumental in key policy and program decisions, yet they are not without costs. Field experiences from MEASURE Evaluation Project demonstrate the need for transparency and collaboration among the key partners, the inevitable balancing of technical requirements with programmatic priorities, and the flexibility required to adapt designs in order to answer the most valuable evaluation questions. Interest in accountability of funding of public health interventions continues to grow, promising continued interest in IEs. Evaluators, implementers and funders can share in these learnings as we move forward with expanding our understanding of the costs and benefits for rigorous evaluations.

References

(1) Savedoff WD, Levine R, Birdsall N. When will we ever learn? Improving lives through impact evaluation. 2006.

(2) PEPFAR. PEPFAR blueprint: creating an AIDS-free generation. 2012.

(3) UNAIDS. Strategic guidance for evaluating HIV prevention programmes. 2010.

(4) USAID. USAID evaluation policy. 2011.

(5) PEPFAR. PEPFAR evaluation standards of practice. 2014.

(6) Lance PM, Guilkey DK, Hattori A, Angeles G. How do we know if a program made a difference? A guide to statistical methods for program impact evaluation. Chapel Hill, NC: MEASURE Evaluation; 2014.

(7) Gertler PJ, Martinez S, Premand P, Rawlings LB, Vermeersch CM. Impact evaluation in practice. : World Bank Publications; 2011.

(8) Rogers PJ, RMIT University (Australia), Better Evaluation. Introduction to impact evaluation. 2012;No. 1.

(9) Perrin B. Linking monitoring and evaluation to impact evaluation. 2012;No. 2.

(10) Bamberger M. Introduction to mixed methods in impact evaluation. 2012;No. 3.

(11) Stern E, Stame N, Mayne J, Forss K, Davies R, Befani B. Broadening the range of designs and methods for impact evaluations. 2012;Working Paper 38:.

(12) Bryce J, Victora CG, MCE-IMCI Technical Advisors. Ten methodological lessons from the multi-country evaluation of integrated Management of Childhood Illness. Health Policy Plan 2005 Dec;20 Suppl 1:i94-i105.

(13) Clemens MA, Demombynes G. When does rigorous impact evaluation make a difference? The case of the Millennium Villages. 2010;Working Paper 225.

(14) Farley K, Lucas S, Molyneaux J, Penn K. Principles into practice: impact evaluations of agriculture projects. 2012.

(15) Bernard T, Delarue J, Naudet J. Impact evaluations: a tool for accountability? Lessons from experience at Agence Française de Développement. Journal of Development Effectiveness 2012;4(2):314-327.

(16) Victora CG, Black RE, Boerma J, Bryce J. Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations. The Lancet 2011;377(9759):85-95.

(17) Habicht JP, Victora CG, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. Int J Epidemiol 1999 Feb;28(1):10-18.

(18) Taylor TM. The Western Highlands Integrated program (WHIP) evaluation baseline survey in Guatemala: a case study in evaluation practice. 2014.

APPENDIX A
IE Lessons Learned: Interview Tool
May 2014

Purpose: To transform your tacit knowledge gained through field experience into explicit knowledge that others can act on.

Introduction: Each evaluation must grapple with cross-cutting challenges that influence study design, implementation, and utilization of results; some challenges are study-specific requiring tailored solutions. The intent of this interview is to learn about the challenges you faced across the life cycle of your evaluation. While we are interested in this work specifically, if a challenge or issue is raised that you found to be more relevant for other MEASURE Evaluation work you have done, then please feel free to draw on these examples as well.

Areas of Interest

Concept & Utility

1. Genesis of Study
    a. What policy or programmatic needs prompted the evaluation
    b. How were research questions determined
    c. Was the scope realistic given the timeline, resources, and project implementation (including time required for the program to influence the outcome)
    d. Did you conduct the study as an external evaluator
    e. Was the evaluation incorporated in the program design from the beginning or did the request for the evaluation come after the program was designed and/or launched

2. Dissemination / Use
    a. How broadly were findings disseminated?
    b. How were the findings used? By whom?
    c. Were expectations met?

Structural/System/Logistics

3. Competing Priorities – How well did the evaluation goals match the program priorities and/or other competing interests such as desire for technical capacity building yet expectations for timely, cost-efficient evaluation was still expected? If so, how did you manage these expectations?

4. Alignment between project and evaluation – How well were the evaluation expectations communicated to the program and/or included in the original RFA and contract/agreement? What might have worked better in your situation vis-à-vis engagement and cooperation from the implementing program?
5. Stakeholders – whose role/responsibility was it to engage stakeholders (implementing partner, government, other donors, etc). Did this help or hinder the process?

<u>Technical</u>

6. Level of Certainty
    a. How important was it to attribute effects to the specific program or donor?
    b. Was a measurable level of certainty necessary or was a plausibility argument sufficient?
    c. What options were considered?
        - Research design
            - Quantitative (e.g. experimental, quasi-experimental)
            - Qualitative
            - Mixed method (both quantitative and qualitative)
        - Evaluation of the program as a whole or evaluation of each component
        - Study site
            - The entire program population/area vs. a subset of it
        - Data collection
            - Frequencies (baseline, endline)
            - (Pooled) cross-sectional vs. longitudinal
        - Any required changes to the study design

7. Identification of the Comparison Group
    d. Describe the method used to identify the comparison group, including data limitations, challenges, accommodations, etc

8. Sampling
    e. How did you balance sample size, level of certainty, and costs?
    f. Did you use any innovative sampling methods?

9. Conceptual framework and empirical model
    g. How was the conceptual framework developed
    h. Was the empirical model built based on the conceptual framework?

APPENDIX B

Evaluation Case Studies

Evaluation of the Bangladesh Smiling Sun Franchise Program

Background

The Bangladesh Smiling Sun Franchise Program (BSSFP), a continuation of the NGO Service Delivery Program, aimed to increase the use of an essential health services including family planning, maternal and child health, and other basic services among poor and underserved populations in Bangladesh. The program adopted a social franchise model and provided health services at static clinics and satellite sites in the network. The program was conducted between 2007 and 2011 in areas where the Government of Bangladesh (GoB) identified to have inadequate delivery system of public health services and sought assistance from BSSFP partners.

Outcomes of Interest

The impact evaluation of the BSSFP examined the impact of the project on use of selected maternal and child health services. Primary outcomes of interest included contraceptive prevalence among ever-married women of reproductive age (CPR) and use of antenatal care for the most recent birth in the last 3 years (ANC).

Research Design

The evaluation focused on quantitative analysis and adopted a difference-in-differences approach in which data were collected through cross-sectional baseline and endline surveys from project and non-project areas conducted in 2008 and 2011, respectively. The project areas were defined as catchment areas of BSSFP static clinics or satellite posts, and the non-project areas were selected from adjacent areas where public health clinics were operating. Approximately 15,000 urban households and 30,000 rural households were interviewed through the two surveys.

Evaluation Status

The evaluation study was completed in 2012. For further reading, see:

Lance P., Angeles G, Kamal N. (2012). *Smiling Sun Franchise Program (BSSFP) Impact Evaluation Report.* Chapel Hill, North Carolina: MEASURE Evaluation.

Evaluation of the Bangladesh NGO Health Service Delivery Project (NHSDP)

Background

The NGO Health Service Delivery Project (NHSDP) is a continuation of the Bangladesh Smiling Franchise Program (BSSFP) that was conducted between 2007 and 2011. The NHSDP was awarded as a four-year long project in 2012 and aims to increase access to and use of maternal and child health services among poor and underserved populations in Bangladesh by supporting delivery of an essential package of health services through a network of NGOs. Similar to BSSFP, the project operates in areas that have been identified by the government of Bangladesh (GoB) to have inadequate public health service delivery systems where the GoB sought assistance from partners to fill the service gap.

Outcomes of Interest

The impact evaluation of the NHSDP seeks to determine the impact of the project on use of selected maternal and child health services. Outcomes of interest include contraceptive prevalence among married women of reproductive age (CPR) and use of antenatal care for the most recent birth in the last 2 years (ANC).

Research Design

Similar to the evaluation design of the BSSFP, the evaluation of the NHSDP relies on quantitative analysis of data collected through cross-sectional baseline and endline surveys from project and non-project areas. The current plan is to analyze the data using a difference-in-differences approach to estimate the impact on the outcomes of interest. The project areas are defined as catchment areas of NHSDP static clinics or satellite sites, and the non-project areas are selected from geographic areas outside of the catchment areas but adjacent to the project areas. Approximately 20,000 urban households and 14,000 rural households will be interviewed at baseline.

Evaluation Status

The evaluation is on-going as of August 2014. Data collection for the baseline survey was initiated in April 2014 and is due to end in late August 2014. Endline data collection will take place in 2017.

Early Marriage Evaluation Study (EMES)

Background

Due to the negative consequences of early childhood marriage, many countries have enacted laws to reduce or prevent marriage for youth under 18 years of age. In Ethiopia, US-funded health and education projects identified early marriage as detrimental to the reproductive health and education outcomes that were being promoted for women. Early marriage prevention activities were incorporated into existing reproductive health and basic education projects in an effort to reduce child marriage. The Early Marriage Evaluation Study (EMES) was an effort to evaluate the effects of these prevention activities on child marriage.

Study Objectives

The primary objective of the EMES was to document the scope and assess effects of early marriage prevention efforts in the Amhara Region of Ethiopia. This was inclusive of documenting early child marriage prevalence, determinants, consequences; measuring exposure to child marriage prevention activities, and assessing the relationship between select health outcomes such as unintended pregnancy and physical or sexual abuse, and exposure to the implemented prevention activities.

Research Design

The study adopted a post-test-only evaluation design to examine outcome measures in program and non-program areas. Cross-sectional population-based survey data were collected with analysis focused on identification of risk factors, and associations between program interventions and early marriage practices. Attribution of differences between communities to specific programs was not feasible due to the lack of a robust estimation of counterfactual.

Evaluation Status

The study was completed in 2007. For further reading, see:

Gage, AJ (Ed.). Coverage and effects of child marriage prevention activities in Amhara Region, Ethiopia: findings from a 2007 study. USAID. MEASURE Evaluation. Addis Continental Institute of Public Health. 2009.

Evaluation plan for the Ghana National Strategy for Key Populations

Background

Ghana faces a mixed HIV/AIDS epidemic with concentration among female sex workers (FSW) and men who have sex with men (MSM). The Government of Ghana (GoG) developed a national strategic plan for 2011-2015 and had a strong interest in developing a targeted evaluation plan as part of the operationalization of the national plan. The Ghana AIDS Commission (GAC) collaborated with the MEASURE Evaluation Project to facilitate an evaluation planning process that followed the UNAIDS-MERG *Strategic Guidance for Evaluating HIV Prevention Programs*.

Research Question

The study identified three research questions to evaluate the HIV prevention programs targeting FSW and MSM. The primary question of interest was whether any change in prevalence or incidence of HIV/AIDS or sexually transmitted infections (STIs) in the FSW and MSM key populations could be plausibly attributed to the implementation of prevention programs. Secondary questions of interest included whether any change in prevalence or incidence of HIV/AIDS or STIs were measured and a performance question evaluating the scope and quality of the FSW and MSM activities planned and implemented.

Research Design

The study adopted a plausibility evaluation design because it would provide the GAC with needed information that was affordable, feasible and maximized existing data. Quantitative data using a before-versus-after model (i.e. single-difference model) drew on existing baseline data from the Integrated Biological-Behavioral Surveillance Survey (IBBSS) in 2011 with a follow-up IBBSS planned in 2015 to assess the secular changes in prevalence and incidence. A performance evaluation of various HIV-related activities and programs is under development and a suggested media scan was planned to understand any contextual events that may confound trends.

Evaluation Status

The evaluation plan was completed in 2013. For further reading, see:

Ghana AIDS Commission. Evaluation Plan for the Ghana National Strategy for Key Populations. 2013. Available at http://www.cpc.unc.edu/measure/publications/sr-13-75

Impact Evaluation of the Western Highlands Integrated Program (WHIP)

Background

In Guatemala, the Western Highlands Integrated Program (WHIP) has been underway since 2012 as one of the largest activities supported by the USAID Mission under an integrated strategy addressing multiple development objectives. The program combines technical support for smallholder farmers with health and nutrition initiatives, and is designed to decrease poverty and malnutrition in priority municipalities of the Western Highlands.

Research Question

The evaluation seeks to examine the program's performance and impact on the prevalence of poverty, chronic malnutrition among children under age of five, and other key indicators.

Research Design

The study is comprised of two evaluations, an impact evaluation and a performance evaluation, and will use data collected through a baseline and two follow-up surveys. The impact evaluation adopts a difference-in-differences approach to examine the impact of the program on the prevalence of poverty and chronic malnutrition through a comparison of intervention and comparison groups. There are three intervention groups defined by level of exposure to the program's agricultural component: people in households enrolled in an agricultural-support program, those exposed only indirectly to the agriculture program through residence in an area where recipients live, and those living outside of areas where any agricultural support recipient resides. Everyone in these three groups is part of the beneficiary population for health and nutrition interventions under the integrated program. Additionally, two comparison groups are included in the study: residents of census tracts matched to the first and second intervention groups, and residents of census tracts matched to the third intervention group. The performance evaluation focuses on tracking changes over time in the intervention groups on a range of health and economic indicators.

Evaluation Status

The evaluation is on-going. The baseline survey was completed in 2013, and the post-intervention surveys are scheduled to be conducted in 2015 and 2017. For further reading see:

Angeles G, Hidalgo E, Molina-Cruz R, Taylor T, Urquieta-Salomón J, Calderón C, Fernández JC, Hidalgo M, Brugh K, Romero M. Monitoring and evaluation Survey for the Western Highlands Integrated Program, Baseline 2013. USAID, MEASURE Evaluation Project, Chapel Hill, NC. Available at: http://www.cpc.unc.edu/measure/publications/tr-14-100

Taylor T. The Western Highlands Integrated Program (WHIP) Evaluation Baseline Survey in Guatemala: A Case Study in Evaluation Practice. USAID, MEASURE Evaluation Project, 2014. Available at: http://www.cpc.unc.edu/measure/publications/sr-14-106.

Impact Evaluation of the Kingston Priorities for Local AIDS Control Efforts (PLACE) Intervention

Background

In response to an increase in the number of persons infected with HIV and a plateau in condom use despite national efforts to reduce HIV transmission in Jamaica, the Jamaica Ministry of Health developed a new prevention initiative in 2005. The strategy of the Kingston Priorities for Local AIDS Control Efforts (PLACE) was shaped by findings from targeted PLACE surveys, surveillance and national survey data, and experience from other programs, and adopted site-based prevention programs focused on promoting safe sex behavior among people with new and concurrent sexual partnerships. It targeted public sites where people socialize and meet new sexual partners in Kingston where the case-rate of AIDS is high. The intervention adopted a multi-level strategy targeting the public sites, groups of individuals, and individuals at these sites.

Research Questions

The primary evaluation outcomes of the Kingston PLACE intervention were the proportions of self-reported new or multiple partnerships and self-reported inconsistent use of condom among patrons of the sites. The IE also examined the proportions of patrons who self-reported having an HIV test in the 12 months prior to the interview.

Research Design

A total of 147 public sites were included in the study and grouped into 50 clusters based on the geographic locations. The geographic clusters of public sites were then randomized into intervention and control groups. The intervention was conducted between January and June 2006. The data were collected through cross-sectional baseline and endline surveys at the sites; approximately 3,000 patrons at these sites were surveyed in 2005 and 2007, respectively. An intent-to-treat analysis was applied to examine the impact.

Evaluation Status

The evaluation was completed in 2009. For further reading see:

Figueroa, J. P., Weir, S. S., Byfield, L., Hall, A., Cummings, S. M., & Suchindran, C. M. (2010). The challenge of promoting safe sex at sites where persons meet new sex partners in Jamaica: results of the Kingston PLACE randomized controlled trial. *Tropical Medicine & International Health*, *15*(8), 945-954.

Weir, S. S., Figueroa, J. P., Byfield, L., Hall, A., Cummings, S., & Suchindran, C. (2008). Randomized controlled trial to investigate impact of site-based safer sex programmes in Kingston, Jamaica: trial design, methods and baseline findings. *Tropical Medicine & International Health*, *13*(6), 801-813.

Evaluation of the Community Care for Vulnerable Children in an Integrated Vulnerable children and Home-Based Care Program

Background

The Community Care Program is a five-year program in Mozambique that was initiated in 2010 and is scheduled to be completed in 2015. The program aims to increase the community-based response to HIV/AIDS and to enhance the health status and quality of life of target populations, including people living with HIV (PHLIV) in need of home based care (HBC) and orphans and vulnerable children (OVC). The program supports community-based organizations that work with cadres of community workers who conduct home visits and/or offer services and referrals to HBC and/or OVC. Traditionally, there was a cadre of OVC community workers and another for HBC workers. Under the Community Care Program, combined services are offered through one cadre of dual-purposed community workers. This integration of service support involved selecting the cadre of workers and training them to serve both HBC and OVC.

Research Question

The study objective was to understand the implications of the program integration on service provision for vulnerable children. Outcomes of specific interest included determining whether services for vulnerable children varied by presence of HBC clients in the household and status of the HBC client; understanding *community worker* perspectives about their work within an integrated project including the benefits/challenges of integration; and understanding the utility of an integrated approach to beneficiary groups and stakeholders.

Research Design

The study is a performance evaluation and adopts a mix-method design based on descriptive cross-sectional analysis of both qualitative and quantitative data to describe the outcomes of interest. Quantitative data was collected through a self-administered survey of community workers, and questionnaires with caregivers at the household level. Qualitative data was collected through interviews with stakeholders at the national and sub-national levels and focus group discussion with community workers.

Evaluation Status

The evaluation is completed. For further reading, see:

Cannon M, do Nascimento N, Chariyeva Z, Foreit K. Mozambique Program Assessment: Community Care for Vulnerable Children in a n Integrated Vulnerable Children and Home-Based Care Program. USAID, President's Emergency Plan for AIDS Relief, MEASURE Evaluation Project, Chapel Hill, NC 2014. Available at: http://www.cpc.unc.edu/measure/publications/sr-14-100

Impact Evaluation of the SUAAHARA-GPM Nepal Program

Background

The SUAAHARA-GPM (Gender, Policy, and Measurement) program aims to address inequalities in access to health services among vulnerable populations, including women and marginalized populations in Nepal. The program is implemented in selected Western Hill sub-region districts, and seeks to increase women's and marginalized group's use of health services through strengthening capacity within health facility operations and management committees (HFOMCs) to address gender equity and social inclusion for quality health services. The project seeks to build individual-level knowledge and skills; strengthen the organizational-level processes that make committees more responsive to the needs of women and other marginalized groups; and translate needs into actions that strengthen systems necessary for improving the responsiveness, oversight, and accountability of health facilities.

Research Question

The evaluation seeks to examine the impact of integrating gender and social inclusion (GESI) approaches into capacity strengthening initiatives with HFOMCs on use and quality of maternal and child health and nutrition services.

Research Design

The evaluation of the SUAAHARA-GPM proposes a three arm model: a comparison group, one intervention group receiving the new GESI-integrated HFOMC training, and the other intervention group receiving the new GESI-integrated HFOMC training along with a GESI community engagement approach. The evaluation adopts a mixed-method design comprised of both quantitative and qualitative analyses. For the quantitative component, a difference-in-differences approach will be used to analyze data from baseline and endline surveys collecting information on communities, households and women with children under age two. The qualitative methods are comprised of exit interviews with MNCH/FP patients, key informant interviews with health facility staff and district-level stakeholders, and in-depth interviews and focus group discussions with HFOMC members; observations of HFOMC meetings and of health facility waiting rooms; and focus group discussions with men and women with children under 1,000 days of age.

Evaluation Status

The evaluation is on-going as of August 2014. Baseline data collection is in progress and due to end in October 2014. The endline surveys are scheduled for 2016.

Impact Evaluation of Malaria Control Interventions on Mortality in Children in Mainland Tanzania

Background

Malaria control efforts in Tanzania have been rapidly scaled up in the last decade from a significant increase in available funding, including the Global Fund to Fight AIDS, Tuberculosis, and Malaria Grant and President's Malaria Initiative (PMI). This has resulted in demands and interests among stakeholders, including policy-makers, program implementers, and funding agencies, to examine the impact of the malaria control interventions on the malaria epidemic and to assess progress toward national and internationally agreed goals, including Millennium Development Goals. Main interventions include: insecticide treated nets (ITN), malaria case-management with artemisinin combination therapies (ACTs), intermittent preventive treatment in pregnancy (IPTp), and indoor residual spraying (IRS).

Research Question

The study aimed to examine whether the interventions had impacts on the burden of malaria. Primary outcomes of interest include: all-cause mortality in children aged under five, malaria parasitemia, and prevalence of severe anemia among children 6-59 months.

Research Design

The study adopted a plausibility evaluation design and used a before-versus-after model (i.e. single-difference model) to assess the secular changes in the outcomes of interest. Multiple sources of secondary data were used for the analysis, including national household surveys data, programmatic data, and case-study data. The study also included a review of relevant published literature to assess support the plausibility argument.

Evaluation Status

The study was completed in 2012. For further reading, see:

Tanzania Malaria Impact Evaluation Research Group. Evaluation of the Impact of Malaria Interventions on Mortality in Children in Mainland Tanzania. USAID, President's Malaria Initiative, Department of Health and Human Services, CDC, Department of State United States of America. 2012.

Impact Evaluation of the Strengthening Tuberculosis Control in Ukraine (STbCU) Project

Background

The Strengthening Tuberculosis Control in Ukraine (STbCU) project goal is to reduce TB morbidity and mortality in Ukraine. Broadly speaking, the project seeks to improve the quality and availability of DOTS-based services, build capacity for programmatic management of drug-resistant TB, improve access to TB/HIV co-infection services, and improve infection control practices, providing a safer medical environment for workers. The project began in March 2012 and builds on more than 10 years of TB assistance in 10 geographic priority areas in Ukraine.

Research Questions

Two research questions guide the impact evaluation study. First is whether a patient social support program to improve TB treatment adherence impacts the subsequent treatment outcome. The study hypothesis is that patients at high-risk for defaulting on TB treatment who receive social support services will improve treatment adherence and outcomes. The second study question is whether an integrated TB-HIV referral and service delivery system with cross-trained health workers will result in early initiation of treatment for the co-infected and reduce all-cause mortality.

Research Design

For each study question above, a difference-in-differences approach comparing intervention and comparison patient populations receiving TB treatment is proposed. Sampling for the social support study will compare patient populations at high-risk and low-risk for defaulting on TB treatment, selected at three points in time, in 2011 pre-program, 2012 following program initiation, and 2015 prospectively with program available. For the TB-HIV integration study, intervention and comparison populations will be selected from matched regions with and without the STbCU program. Review of patient TB treatment records will capture different treatment outcomes or exit events with varying duration times from entry to exit; hence the data lends itself to survival analysis. Using data from complete and censored cases, survival curves will be generated to estimate the time to exit event for different intervention and comparison groups, with log-rank statistical tests to test differences in the survival functions across groups.

Evaluation Status

The evaluation is on-going, with baseline data collection in progress and due to end in September 2014. The endline data collection is prospective and planed for 2015-2016.

WORKING PAPER

MEASURE Evaluation

University of North Carolina at Chapel Hill

400 Meadowmont Village Circle, 3rd Floor

Chapel Hill, North Carolina 27517

Phone: +1-919-445-9359 • measure@unc.edu

www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. WP-14-157