Design Considerations for Evaluating the Impact of PEPFAR: Workshop Summary

ISBN: 0-309-11673-2, 0 pages, , ()

DESIGN CONSIDERATIONS FOR EVALUATING THE IMPACT OF PEPFAR

WORKSHOP SUMMARY

Clara Cohen, Michele Orza, and Deepali Patel, *Rapporteurs*

Board on Global Health

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

This study was supported by Contract No. SAQMPD05D1147 (STAT-7394) between the National Academy of Sciences and the Department of State. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the organizations or agencies that provided support for this project.

International Standard Book Number-13: 978-0-309-11672-5
International Standard Book Number-10: 0-309-11672-4

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

For more information about the Institute of Medicine, visit the IOM home page at: **www.iom.edu.**

Printed in the United States of America

The serpent has been a symbol of long life, healing, and knowledge among almost all cultures and religions since the beginning of recorded history. The serpent adopted as a logotype by the Institute of Medicine is a relief carving from ancient Greece, now held by the Staatliche Museen in Berlin.

Suggested citation: IOM (Institute of Medicine). 2008. *Design considerations for evaluating the impact of PEPFAR: Workshop summary.* Washington, DC: The National Academies Press.

"Knowing is not enough; we must apply.
Willing is not enough; we must do."
—Goethe

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Advising the Nation. Improving Health.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

Reviewers

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

Ties Boerma, Measurement and Health Information System, World Health Organization
Fred Carden, International Development Research Centre
Helen Gayle, CARE
Ruth Levine, Center for Global Development
Phillip Nieburg, HIV/AIDS Task Force, Center for Strategic and International Studies
Nancy Padian, Research Triangle Institute

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the final draft of the report before its release. The review of this report was overseen by

Dr. Frederick A. Murphy, Department of Pathology, the University of Texas Medical Branch at Galveston. Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authors and the institution.

Contents

List of Tables, Figures, and Boxes

TABLES

FIGURES

BOXES

Preface

Human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) has become one of the largest epidemics in history, with more than 33 million people living with the disease, over 2 million deaths, and more than 2 million new infections estimated last year (UNAIDS and WHO, 2007). Developing countries—where the epidemic has caused not only loss of life, but also major social and economic dislocations—have borne a disproportionate share of the HIV/AIDS disease burden.

The United States has become a major player in the global response to HIV/AIDS through the President's Emergency Plan for AIDS Relief (PEPFAR). Legislated in 2003 through the United States Leadership Against HIV/AIDS, Tuberculosis, and Malaria Act of 2003 (The Leadership Act), PEPFAR is a 5-year effort that seeks to prevent 7 million new AIDS infections, treat 2 million people with AIDS, and care for 10 million orphans and other vulnerable groups, with a focus on 15 target countries.

ABOUT THE WORKSHOP

Also included in The Leadership Act legislation was a mandate to the Institute of Medicine (IOM) to appoint an expert committee to conduct an evaluation of the implementation of PEPFAR. The IOM committee began work on the evaluation very early in PEPFAR's implementation because the evaluation was mandated to be delivered 3 years after the legislation was passed. It was possible to evaluate only the first phase of the implementa-

tion, and at the close of the committee's evaluation, PEPFAR had been supporting programs in the focus countries for less than 2 years.

As the final element of its project on PEPFAR, the IOM convened a workshop, "Design Considerations for Evaluating the Impact of PEPFAR," on April 30 and May 1, 2007. The workshop focused on developing methodological, policy, and practical design considerations for a future evaluation of PEPFAR's impact at a point when the program is sufficiently mature to fairly judge its impact. The workshop underscored what the evaluation committee and workshop participants would have liked to have evaluated in the long term and sought to outline more meaningful questions about the true impact of the program. Largely because of the mandated timing, the implementation evaluation that the IOM committee provided in its recently published report, *PEPFAR Implementation: Progress and Promise* (IOM, 2007), could not answer the questions that deeply interest the U.S. Congress and others about the impact of the program. Although the PEPFAR report was limited in looking at early indicators—inputs, processes, and a few outputs—the workshop was able to address what longer term outcomes and impacts could be evaluated in the future.

Three main perspectives on accountability were sought at the workshop: "upward" accountability to the U.S. Congress, "horizontal" accountability to global partners, and "downward" accountability to country partners and intended beneficiaries. The workshop was widely consultative and brought together a range of interested parties—including staff of the U.S. Congress; PEPFAR officials and implementers; major multilateral organizations such as The Global Fund to Fight AIDS, Tuberculosis, and Malaria (The Global Fund), the Joint United Nations Programme on HIV/AIDS (UNAIDS), and the World Bank; evaluation experts experienced with similar types of evaluations; and representatives of partner countries, particularly the PEPFAR focus countries.

The first day of the workshop was devoted to outlining broader issues and hearing from relevant perspectives; the second day focused on technical and methodological issues in evaluating the impact of PEPFAR. Is PEPFAR helping partner countries to succeed, and how could it do better? What are the right specific questions to ask, and what are the best ways to get the answers? How can PEPFAR coordinate and harmonize to get the most from its evaluation resources? These are among the basic questions that the workshop discussions addressed.

ORGANIZATION OF THE WORKSHOP SUMMARY

This workshop summary is divided into four chapters, preceded by an Overview. The Overview puts forth core messages that arose from the workshop presentations and discussions that may be of greatest interest to

decision makers. Chapter 1 introduces the PEPFAR program and includes workshop discussions on the definition of impact evaluation and previous evaluation efforts, including those internal to the PEPFAR program and that by the IOM committee. In Chapter 2, workshop participants provided their vision for the questions of interest that could be addressed in a future impact evaluation of PEPFAR. These questions include those about both the process and the results of program implementation. Chapter 3 describes the benefits, costs, and opportunities for conducting the impact evaluation in a way that incorporates guiding principles of coordination, harmonization, and capacity building. Finally, Chapter 4 describes workshop participants' discussions of the methodologies and approaches that can be used in impact evaluation, lessons learned from previous evaluations of HIV/AIDS programs, and specific methodological challenges and opportunities. The meeting agenda, list of acronyms, list of participants, and bibliographic references are included in the report's appendixes.

The authors prepared this summary on the basis of attendance at the workshop, associated materials, and a transcript, webcast, and audio-recordings of the meeting, presentations, and discussions that took place during the workshop. Chapters have been edited and organized around major themes to provide a more readable summary and to eliminate duplication of topics. The material presented reflects only the views and opinions of those participating in the workshop and not the consensus view of a formally constituted study committee. The summary reflects only what was covered at the workshop and is not intended to be a comprehensive examination of the subject matter.

ACKNOWLEDGMENTS

We are grateful to the many people who contributed to making the workshop a success. Many thanks to Ruth Levine for superbly moderating the meeting and to Phil Nieburg for his able assistance in moderating. We appreciate the continued service of the IOM Committee for the Evaluation of PEPFAR Implementation in serving as the steering committee for the workshop, especially those members who also participated in the meeting: Drs. Stefano Bertozzi, Geoff Garnett, Bill Holzemer, Carl Latkin, and Jim Sherry. All of the speakers and discussants were admirably generous with their considerable expertise but limited time, particularly Jonathan Mwiindi and Dr. Agnes Binagwaho, whose travel from Africa required an investment of several days in addition to the workshop. We also appreciate the continued hard work of the staff—Dr. Michele Orza, Kimberly Scott, and Angela Mensah—for whom the preparations for this workshop followed immediately upon the release of the committee's report, leaving them no time for even a brief respite. They could not have accomplished it without

the kind and expert assistance of Hellen Gelband. Deepali Patel's editorial support in revising the draft workshop summary in response to the comments of the external reviewers is also greatly appreciated. Thank you to the PEPFAR evaluation team, led by Drs. Kathy Marconi and Paul Bouey, for their support in planning the meeting and participation throughout. We are grateful to the Kaiser Family Foundation for its webcast of the workshop proceedings, allowing people who were not able to join us to see and hear the proceedings. Last, but most definitely not least, special thanks to our lead author, Dr. Clara Cohen, who somehow managed to condense 2 full days of detailed presentations and involved discussions into this coherent and useful summary.

THE ROAD AHEAD

The workshop and this summary are intended to be helpful to the U.S. Congress in developing expectations for the evaluation of PEPFAR's impact as well as to those involved in implementing and evaluating the PEPFAR program. Because the law authorizing PEPFAR will expire in September 2008, I hope the report will contribute to developing a compelling, informed, and expanded vision for building on PEPFAR's initial success.

In her remarks to conclude the workshop, moderator Ruth Levine used a colorful and creative analogy comparing PEPFAR to a car with a full tank of gas—several billions of dollars worth—with instructions to go as fast as possible. She described back-seat drivers in the car who admonish the driver about speed, direction, and number of passengers and a road that is also moving at the same time. Where is the car relative to where it wants to be? Where is the car relative to where it was? Is the car moving in the most direct way to where it should be? Should we stop driving or continue driving in the same direction without looking at the signals along the way? This workshop has demonstrated that the best option is to ask key questions, look for signs that can help orient us all, and keep moving.

Jaime Sepúlveda, *Chair*
IOM Committee for the Evaluation of PEPFAR Implementation

Overview

A 2-day workshop on methodological, policy, and practical design considerations for a future evaluation of human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) interventions carried out under the President's Emergency Plan for AIDS Relief (PEPFAR) was convened by the Institute of Medicine (IOM) on April 30 and May 1, 2007. Participants at the workshop included staff of the U.S. Congress; PEPFAR officials and implementers; major multilateral organizations such as The Global Fund to Fight AIDS, Malaria, and Tuberculosis (The Global Fund), the Joint United Nations Programme on HIV/AIDS (UNAIDS), and the World Bank; evaluation experts experienced with similar kinds of evaluations; and representatives of partner countries, particularly the PEPFAR focus countries. The workshop represented a final element of the work of the congressionally mandated IOM Committee for the Evaluation of PEPFAR Implementation, which published a report of its findings in 2007 (IOM, 2007) evaluating the first 2 years of implementation, but could not address longer term impact evaluation questions.

This overview describes core messages from the workshop's presentations and discussions. First, background is provided on the definition and uses of impact evaluation, the process of internal evaluation within the PEPFAR program, and the recommendations concerning the design of

---

The workshop summary has been prepared by the workshop rapporteurs as a factual summary of what occurred at the workshop.

future impact evaluations by the IOM committee that evaluated PEPFAR implementation. Next, a vision is described for the types of questions workshop participants would like to see addressed in future impact evaluations, along with suggestions for how the process of impact evaluation would ideally be carried out. The final section addresses methodological issues that were raised by participants as being important to consider in the design of future impact evaluations.

DEFINING AND USING IMPACT EVALUATION

Meeting participants proposed a working definition of impact evaluation as a measurement of net change in outcomes attributable to a specific program using a methodology that is robust, available, feasible, and appropriate both to the question under investigation and to the specific context. Workshop participants noted that impact evaluation is not only about outcomes, but also the process that leads to outcomes; that is, it includes both means and ends. Participants argued for a definition of impact evaluation that is longer term, more broadly defined, and less linear. Although a more traditional definition of infectious disease impact evaluation (that is, one that is limited to metrics such as prevalence, incidence, infections averted, morbidity, and mortality) is important, the broader and deeper impact evaluation envisioned would also include measurement of changes in health status, systems capacity, quality of services, economic development, and social, economic, and political outcomes.

Two major uses of evaluation were described: (1) use of evaluation for judging the performance of the program for purposes of accountability (summative evaluation) and (2) use of evaluation for informing the improved decision making within a program (formative, or utilization-focused, evaluation). Formative evaluation of PEPFAR is important to inform both congressional decision making and programmatic decision making in partner countries, although decision makers at different levels may have different evaluation needs, participants said. For each evaluation question, participants noted, it is important to clarify who needs the information, what information is needed, and when.

INTERNAL EVALUATION OF PEPFAR

According to PEPFAR officials, a key aspect of PEPFAR's monitoring process and infrastructure is to track progress toward the program's goals in prevention, treatment, and care. PEPFAR internal monitoring supports the principles of local leadership and ownership of the HIV/AIDS response by building local capacity, using local infrastructure, implementing the program according to national guidelines, monitoring using locally developed

indicators, and funding programs based on results, PEPFAR officials noted. Over time, PEPFAR's evaluation activities have expanded field operations relative to central operations and have expanded reporting infrastructure and trained personnel in-country.

Initially, reported PEPFAR officials, the conceptual framework for PEPFAR impact evaluation was much more narrowly defined, and country capacity for monitoring was limited, with a lack of consistent targets across countries and a lack of hard data. Evaluation efforts and reporting requirements were loosely coordinated among U.S. implementing agencies. Over the first 5 years of PEPFAR, evaluation planners developed a monitoring and evaluation system that relied on survey data and periodic targeted evaluations on selected topics. Although this approach provided results on the deployment and use of funds, the development and delivery of services, and the beneficiaries of program services, as the complexity of PEPFAR program strategies grows and as the program undergoes a transition from an emergency response to a sustained response, the definition of impact and approaches used to measure impact will need to be broadened. PEPFAR has recently developed a new, centrally managed public health evaluation (PHE) structure that can be used to aggregate results across multiple countries, multiple time points, and multiple settings. The PHE approach is designed to support evaluation by helping to set priorities, provide technical assistance, establish common protocols, and coordinate projects across countries.

FUTURE EVALUATION DESIGN: PERSPECTIVE FROM THE IOM COMMITTEE

Speaker Jaime Sepúlveda, chair of the IOM committee that authored the report, *PEPFAR Implementation: Progress and Promise* (IOM, 2007), reviewed the recommendations of the committee on the design of impact measures—both AIDS-specific and more general indicators—for future evaluation of PEPFAR. Future evaluation should include measurement of what Sepúlveda referred to as the "three generations" of HIV/AIDS surveillance: prevalence and incidence of HIV infection, measurement of behavioral change, and measurement of stigma and discrimination. Other important indicators include measures of survival, quality of life, development of drug resistance, and the overall physical, mental, and social well-being of those people affected by HIV/AIDS. PEPFAR evaluation should also develop more general indicators, such as the empowerment of women and girls, overall health status, capacity of community-based organizations to respond, and public health infrastructure and capacity.

USE OF EVALUATION TO FOCUS ON MEANINGFUL ENDPOINTS

There was substantive focus during the workshop on what participants wanted to accomplish through PEPFAR. Although the workshop drew participants from wide-ranging and diverse perspectives—representatives of the U.S. government, evaluation methodological experts, representatives of partner countries, and global partners—most participants converged on the same kinds of endpoints. Participants emphasized a need to move beyond counting numbers of people who are "touched" by the program and to instead focus on a set of more meaningful and more strategic questions about what interventions succeed. Participants cautioned that an emphasis on counting can be misleading, can drive inflated reporting, and can jeopardize sustainability of a program.

Most participants called for a broader interpretation of impact that includes measurement of not only AIDS-specific impacts, but also more general impacts; measurement of not only results of implementation, but also the process of implementation; and measurement not only of the overall benefits, but also the distribution of benefits. Workshop participants identified questions for evaluating impact that can be clustered into the following nine broad categories: cost-effectiveness, logic of conceptual approach, health impacts, impacts beyond health, capacity building and health systems strengthening, coordination and harmonization, sustainability, equity and fairness, and unintended impacts. These questions are summarized in Box O-1.

USE OF EVALUATION TO FOCUS ON COLLECTIVE OUTCOMES

The meeting set a tone for the ideal conduct of evaluation—emphasizing collaboration, consultation, harmonization with the host countries, and coordination among global partners. Several workshop participants observed that the principles of coordination and harmonization need to be reflected in both the evaluation effort and in the overall implementation of PEPFAR. Workshop participants noted that many different actors—PEPFAR, The Global Fund, UNAIDS, the World Bank and others—are all working in HIV/AIDS response in many of the same countries and can learn from one another. Given the costs of evaluation, setting up evaluation in a collaborative way can take advantage of this synergy and save resources. Many participants observed that although exclusive attribution of program successes to specific funders may not be realistic or constructive, coordinated evaluation may be able to better illuminate what types of interventions are the most effective. Workshop participants acknowledged that there have been earnest efforts to improve coordination and harmonization over the life of PEPFAR, and there is increased agreement to focus evaluation on collective outcomes.

BOX O-1
Summary of Impact Evaluation Questions as Identified by Workshop Participants

Cost-effectiveness

Approaches, strategies, and interventions, such as prevention services and treatment options

Conceptual Approach

Countries targeted
Populations targeted
Budget allocations for types of interventions
Management and financing

Health Impacts

HIV/AIDS-specific health impacts

- Prevalence, incidence, morbidity, mortality, longevity
- Prevention of HIV transmission
- Quality of life
- Behavioral change
- Stigma and discrimination

Other health impacts, disaggregated by population

- Overall mortality, lives saved, survival
- Child mortality
- Fertility, unintended and intended pregnancy

Impacts Beyond Health

Gender equality

- Effectiveness of PEPFAR in addressing underlying causes of women's vulnerability
- Effectiveness in building men's and women's analytical skills and competencies
- Effectiveness of messages for behavioral change
- Effectiveness of interventions to reduce the spread of HIV infection to women and girls
- Effectiveness of the "packaging" of gender interventions

Child welfare

- Effectiveness in improving parenting skills
- Health, nutritional, and educational status of orphans and vulnerable children

Continued

BOX O-1 Continued

Security, development, productivity, and poverty alleviation

- National peace and security in a nation
- Poverty alleviation and economic growth
- National development

Institutional and societal changes

- Policy changes such as property rights, inheritance laws, and human rights; political will; community ownership; food and water security; engagement of vulnerable populations; destigmatization; and discrimination measures
- National priorities and political views

Impacts on Capacity Building and Health Systems Strengthening

Health care workforce

- Effects of PEPFAR on workforce shifts
- Impact of PEPFAR's training approaches
- Impact of PEPFAR's programs to support health care workers
- Effectiveness in sustaining local workforce development systems

Effectiveness of institution-building efforts

- Institution building of community-based organizations
- Capacity building of nontraditional institutions

Infrastructure

- Effectiveness of strengthening supply chain management and drug delivery systems

Quality of care and service delivery

- Prevention, care, treatment, support, and mitigation; antiretroviral (ARV) drug retention rates; levels of client satisfaction; appropriateness of referrals; community attitudes toward people living with HIV/AIDS; and rational prescription behavior
- Development of a knowledge base of what interventions work
- National-level health agenda
- Integration with other health issues; change in national-level health agendas

Coordination and Harmonization

Coordination among U.S. government implementing agencies

- Consistency of targets among implementing agencies
- Positive and negative impacts of complementary interventions, or "wrap-around" programs

BOX O-1 Continued

Harmonization and alignment with partner countries

- Openness and accessibility of plans and projects at the community level
- Presence of instruments and structures for joint decision making
- Degree of information sharing and joint implementation among partners
- Existence of mechanisms to make coordination more flexible

Coordination among program implementers

- Development and effectiveness of a variety of coordination tools or mechanisms for fostering the exchange of learning
- Level of harmonization of drug procurement systems

Sustainability Impacts

Degree to which additional resources have been leveraged from other donors
Extent to which long-term learning and research have been promoted
Measures of capacity building and sustained contributions to institutions and systems
Degree to which local implementation, ownership, and coordination have been promoted

Equity and Fairness Impacts

Existence and effectiveness of processes for goal setting and implementation
Fairness impacts, disaggregated by group, of program integration within the health system
Existence and effectiveness of compensatory mechanisms to improve fairness

Positive and Negative Unintended Impacts

Impacts of earmarking on program integration
Diversion of resources from neglected health care areas and the broader health care system
Impact of PEPFAR on corruption
Impact of PEPFAR on access to services
Impact of treatment on adverse and high-risk behavior
Impact of nutritional programs
Impact of PEPFAR programs on reproductive health and family planning

- Impact of counseling and testing on pregnancy care
- Impact of ARV treatment on fertility and orphanhood

Coordination and harmonization of the evaluation process can provide the benefits of mutually influencing others' work, minimizing transaction costs, ensuring more efficient use of funds, and bringing to bear the strengths and perspectives of key stakeholders. Partner countries, implementing partners, and beneficiaries are among the perspectives that are critical to evaluation design, workshop participants said. Partner countries add value because they are accountable to their citizens and have experience in dealing with the challenges of service delivery. Implementing partners offer a familiarity with program data and lessons, along with a keen understanding of the challenges of delivering services. Local people add value to the evaluation process because of their deep contextual knowledge and expertise about program impact at the community level.

Although the process of coordination can be very time-consuming, there is great potential for coordination and harmonization to improve the sharing of data, approaches, and evaluation research among partners. Workshop participants highlighted key opportunities to share the outcomes of PEPFAR's evaluation efforts with the 5-year evaluation of The Global Fund, a planned evaluation by UNAIDS, and evaluative efforts by the World Bank, the Organisation for Economic Co-operation and Development, and the World Health Organization. Participants also noted that opportunities to harmonize the design, conduct, and interpretation of evaluation results with country-level partners include the conduct of joint field evaluations by collaborating partners; the development of centralized funding, knowledge management, and data aggregation systems; and discussions leading to consensus among partners on monitoring approaches and overall program objectives.

USE OF EVALUATION TO BUILD LOCAL CAPACITY

The evaluation effort is an element of the larger effort to build capacity in partner countries so that they have, going forward, a lasting capacity to respond to their own epidemics, workshop participants said. Many participants reflected on the fact that evaluators often rely on their own data collection and evaluative capacities instead of helping countries to develop their own. Impact evaluation is most constructive when it is done in a way that builds countries' capacities to collect, analyze, and use program information, thereby strengthening the sustainability of programs.

Participants noted, however, that multiple constraints stand in the way of building local evaluative capacity, including weak or absent country-level systems for gathering data, funding mechanisms that limit the prioritization of monitoring and evaluation, competition for resources between evaluation and implementation, poor engagement of stakeholders in defining the

monitoring agenda, and a tendency of researchers to drive a more narrow and less useful evaluation agenda.

Opportunities for strengthening local capacity to conduct impact evaluation include the dissemination of methodologies, provision of technical assistance, recruitment and training of personnel in evaluation methods, and promotion of country-driven priority-setting processes for evaluation.

DESIGNING AN EVALUATION BASED ON ROBUST METHODOLOGIES

Workshop participants stressed the importance of designing impact evaluation on the basis of a logical conceptual approach and robust methodologies. Case studies of evaluations of HIV/AIDS interventions were presented, and challenges and opportunities in evaluating impact were discussed.

A number of general principles or observations by participants emerged from the discussions:

- **Prioritization** is needed to narrow down what needs to be measured. For long-term evaluations, for example, only those issues common to all projects might be selected. For a large portfolio of activities, a more narrowly defined set of indicators might be selected. Information about who needs the evaluative information and when can inform the prioritization process.
- **Formative or "learning" evaluation** is a common component of many of the evaluations discussed, that is, ongoing evaluation to improve programming and to inform decision making, as opposed to evaluation at the end of a program to judge the success or failure. Negative evaluation results also offer value for learning and thus should be as widely disseminated as positive results.
- **Multiple methodologies** were used in many of the evaluations discussed and can provide richer results than just one or two methods. Many evaluations used both qualitative and quantitative methodologies. These included case studies, working papers, interviews, models, literature reviews, surveys, field work, participatory approaches, and theory. Multiple methodologies can be more valuable if selected strategically to complement each other.
- **Randomization** is a powerful technique that can add value and credibility to evaluation. Although perceptions persist that randomization is difficult to implement and impractical at the country level, new methods are available to more easily incorporate randomization into a study.
- **Consultation and communication** are an important part of the evaluative process. Changes resulting from the evaluation may have more

to do with the communications and consultations used during the process than with the actual results of the evaluation. Consultation is an important part of understanding what decision makers want from evaluation.

- **Limitations of data and models** need to be well understood. The data collection systems in many partner countries are weak, and empirical data are often inadequate or inaccurate. Similarly, models have limitations that need to be understood. Models need to be validated with empirical data and made more accurate through the addition of variables.
- **Early design of the evaluation** is critical to ensuring that the design is appropriate and that impacts can be detected early. Early design is especially needed to facilitate the use of randomized approaches.
- **Comparison across contexts** is an important attribute of evaluations, but change may be highly contextual. Workshop participants noted that success in one context may not necessarily be transferable to another. Factors independent of program interventions may have a significant influence on change.

Methodological Challenges and Opportunities in Evaluation

Challenges and Opportunities in Measuring HIV/AIDS-Specific and General Impacts

Workshop participants explored the limitations of commonly used methods in evaluation, as well as new prospects, in measuring both HIV/AIDS-specific and more general impacts. These are summarized in Table O-1.

Challenges and Opportunities in Attributing Impact

Attribution—relating the impact of a particular investment to a particular donor—is one of the greatest methodological challenges in impact evaluation, workshop participants said. It is extremely difficult to tease out the exclusive impacts of efforts of any one donor from those of others given the number and diversity of programs and funders working in the area of HIV/AIDS. Evaluating donor-specific attribution, however, may not be constructive. Several participants said that it is perhaps more useful to determine, using evidence-based approaches, what interventions are most effective and then to judge donors by whether they invest in those approaches.

Challenges and Opportunities in Aggregating Evaluation Results

The statistical synthesis or aggregation of the results of multiple studies is a methodological frontier, workshop participants said. Meta-analysis,

which combines results from multiple studies as if they were a single large study, is a tool that is currently underdeveloped for application to impact evaluation. Multiple analyses also have value in the independent validation of results.

TABLE O-1 Challenges and Opportunities in Measuring HIV/AIDS-Specific and General Impacts

HIV/AIDS-Specific Impacts	
Metric	**Definition**
HIV prevalence	The proportion of individuals within a population infected by HIV. Prevalence is a function of both the death rate of those infected and the rate at which new infections occur.
HIV incidence	The number of new cases of HIV within a population at risk over a given period of time.
Infections averted	The difference between expected and actual annual incidence.

Challenges	Opportunities
• Measurement through testing of pregnant women at antenatal clinics (ANCs) tends to overestimate prevalence because ANCs are urban.	• New tools that overcome some of the limitations of HIV prevalence measurement through ANC surveillance have been developed. • A second population-based survey of HIV testing will soon become available, allowing further analysis of prevalence. • Respondent-driven sampling methods are being developed for examining prevalence in high-risk groups.
• Longitudinal cohort studies may not reflect the true incidence in the population, and many participants in cohort studies may be lost to follow-up. • Laboratory assays used to distinguish recent infections from long-term infections tend to overestimate the proportion of most recent infections. • Modeling tools are limited in their ability to accurately measure population-level risk. • Empirical data on incidence by age and sex are lacking.	• An adjustment formula for the HIV incidence laboratory assay has been developed. • A more specific laboratory assay will improve the ability to distinguish long-term and recent infections. • A new population-based survey will provide important age-, sex-, and geography-specific incidence information. • Modeling from prevalence data and accounting for survival of infected individuals can be used to calculate incidence.
• Infections averted are a "nonevent," and their measurement requires multiple assumptions. • Population projection modeling has limitations because of the gaps in data available in developing countries. • Models may not account for epidemiological contextual factors.	• A new model called Spectrum takes epidemiological contextual factors into account. • A Futures Group model can be used to attribute infections averted to specific interventions. • Serial HIV population surveys can help to deduce changing incidence and infections over time. • Cross-country comparative analyses of HIV dynamics and intervention uptake can be used to measure the relative effectiveness of interventions in averting infections.

Continued

TABLE O-1 Continued

Metric	Definition
Survival and mortality rates	Mortality rate: the ratio of deaths in an area compared to the population of that area per unit of time. Survival rate: the percentage of people in a study or treatment group who are alive for a given period of time after diagnosis or treatment.
Behavioral change	Modification of sexual, injection, and drug-adherence practices.

Challenges	Opportunities
• Measuring overall mortality change in response to treatment is challenging because increased survival of HIV-infected individuals and increased opportunities for viral transmission to others decrease and increase mortality, respectively. • Mortality data are poor because of weak and inaccurate mortality surveillance systems and loss of patients to follow-up. Cause-specific and cohort-specific survival data are lacking. • Models frequently use only vital registration, population-level data.	• Mortality rate data quality can be improved by aggressively pursuing information on those patients lost to follow-up and by standardizing methods for collecting information on deaths from hospital records. • Use of "verbal autopsies" can be used to follow up on deaths in households and determine the cause of death. • Corporate-sector surveillance systems can provide early indicators of the impact of treatment programs on mortality. • Age-specific and population-based mortality data can be gathered for improved measurement of mortality impact.
• Data and surveillance gaps exist, particularly in behavioral surveillance with biomarkers. • Surveys and simulation models do not illuminate why specific populations are affected differently. • Current methods might not be able to determine the extent and coverage of behavioral change. Incomplete behavioral change can have worse consequences than no behavioral change. • Factors independent of the program intervention may influence behavioral change.	• Change at the social and institutional levels to build and sustain infrastructure for risk reduction can be tracked. • Opinions of leaders, impediments to behavioral change, and unintended negative consequences of behavioral change can be measured. • Behavior surveys, particularly those targeting younger people, who are an early indicator of prevalence changes, can be useful for attributing changes in HIV incidence to specific changes in risk. • Models combining trends in prevalence and incidence with studies of risk behavior can be a useful tool for retrospectively understanding how interventions might have worked to maximize declines in HIV prevalence.

Continued

TABLE O-1 Continued

Metric	Definition
Stigma and discrimination	Negative attitudes, beliefs, and actions toward people who are perceived to have HIV/AIDS and those associated with them.
Orphanhood prevention	Prevention of the death of one or usually both parents of a child.

Challenges	Opportunities
• Rigorous research and data collection approaches are absent. Most of the literature is based on anecdotal evidence, testimonials, and a few qualitative studies. • Absence of scales to measure stigma and its effects and tools to measure the effectiveness of strategies for mitigating stigma.	• An International Planned Parenthood Federation stigma index is now available. Another instrument, reflecting 33 factors measuring people's perceptions, has also proved reliable for measuring stigma. • New sources of data from focus groups have been useful in assessing stigma.
• Measurements do not distinguish between children who have lost one parent (single orphans) and children who have lost both parents (double orphans) to HIV. • Treatment has an unclear impact on orphanhood. Treatment of HIV-positive orphans extends years of orphanhood, and while treating HIV-positive parents can reduce orphanhood years of existing children by prolonging parents' lives, it can also generate years of orphanhood among children who are born to HIV-positive parents during treatment. • Methods do not exist for conducting cost-effectiveness analysis on interventions to prevent orphanhood.	• New models are under development to better quantify the impact of treatment and prevention in preventing the orphaning of children. • Useful indicators, such as "years of orphanhood averted" and "number of children who reach age 18 before the death of a parent whose life is extended by antiretroviral therapy (ART)," have been developed.

Continued

TABLE O-1 Continued

Metric	Definition
Development of drug resistance	The evolved capability of HIV to withstand a drug to which it was previously sensitive.
General Impacts	
Metric	**Definition**
Health systems strengthening	Improvement of a broad range of factors related to health care service delivery, including accessibility, quality, efficiency, and equity of services; management; procurement and distribution systems; human resource use; policy environment; and infrastructure.
Health care workforce strengthening	Improvement of a range of capacities related to health care personnel, including training, supervision, and job satisfaction.

Challenges	Opportunities
	• The threshold survey can be used to assess transmitted HIV infection using blood tested at ANC sentinel surveillance sites. Blood sampled from young women (age 25 or younger) in their first pregnancies who are likely not to be in ARV treatment can be used to track the transmission of drug-resistant HIV strains. • Therapy monitoring can be used to measure drug resistance by sampling and monitoring patients in ARV treatment from the initiation of therapy over a 1-year time period. Indicators of drug resistance such as outcome, viral load, and drug adherence can be monitored.

Challenges	Opportunities
• Effects on the health system can be positive or negative, intended or unintended. • Health systems represent a diverse set of institutions that may or may not be easily compared. • Health systems include a diverse range of elements. • Empirical estimates of impacts are lacking. • The small sample size and short time interval over which change is often evaluated limit many studies. • Impact attribution is difficult in this type of analysis.	• Facility surveys, provider surveys, and qualitative interviews can be used to measure a variety of attributes of the health system. • The quantity of non-HIV health services delivered before and after the introduction of basic HIV care can be compared, using regression analysis to control for independent effects.
	• A range of new indicators could be developed to measure impacts of interventions on capacity development, training and supervising effectiveness, gender equality, competencies, etc. • A system could be created to track health care workers over time, from registration to retirement. • Methods could be developed to evaluate the degree to which interventions strengthen institutions that regulate the workforce (that is, accrediting associations).

Continued

TABLE O-1 Continued

Metric	Definition
Effectiveness of complementary interventions	Effectiveness of programs complementary to more narrowly focused HIV services, including interventions in areas such as malaria, tuberculosis, nutrition education, food security, social security, education, child survival, family planning, reproductive health, medical training, health systems, and potable water.
Effectiveness of gender-focused activities	Improvements in gender equality and women's empowerment.
Effectiveness of coordination and harmonization	Increased alignment of HIV/AIDS interventions with country-level plans and coordination of efforts among other implementing partners.
Effectiveness of community- or population-level service delivery	Improved service delivery for specific populations, that is, children, families, communities, HIV-infected groups, high-risk groups, etc.

Challenges	Opportunities
	• Prospective randomized evaluation can be used to compare later program enrollees to earlier program enrollees by monitoring a range of indicators (that is, education and health indicators).
• Given the multidimensional, open, complex, nonlinear, and adaptive nature of gender, it is difficult to define what constitutes success. • Few outcome evaluations and few tools have been developed on how gender-focused activities affect HIV risk, and few good indicators exist that are useful in understanding social dynamics. • Evaluations of gender activities tend to underrepresent the perspectives of local people.	• The Gender Equitable Men's (GEM) scale can be used to look at gender norm attitudes and how they change over time. The scale is an index of 24 items, including home and child care, sexual relationships, health and disease prevention, violence, homophobia, and relations with other men.
	• The Country Harmonization and Alignment Tool (CHAT) can be applied to the standardization of approaches for alignment of interventions with country-level plans and coordination of efforts among partners.
	• Community-level program information reporting systems (CLPIR) have been developed to examine community-level service delivery and help answer questions such as when, how, and where people want testing and treatment.

1

Introduction to Impact Evaluation for PEPFAR

This chapter summarizes discussions at the workshop about the meaning and uses of impact evaluation. Uses of impact evaluation to judge performance—summative evaluation—and to inform decision making for program improvement—formative evaluation—are described. Next, the chapter reviews the approach for evaluating human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) interventions carried out through the President's Emergency Plan for AIDS Relief (PEPFAR) and how that approach has evolved over time. Finally, the chapter considers the major findings of the Institute of Medicine's (IOM's) preliminary evaluation of PEPFAR, *PEPFAR Implementation: Progress and Promise* (IOM, 2007), including recommendations to inform the design of future impact evaluations.

MEANING AND USES OF IMPACT EVALUATION

Defining Impact Evaluation

Workshop moderator Ruth Levine of the Center for Global Development proposed a definition of impact evaluation as a measurement of net change in outcomes attributable to a specific program using a methodology that is robust, available, feasible, and appropriate, both to the question under investigation and to the specific context. In the context of PEPFAR, she noted, impact evaluation can provide insights about the outcomes from specific interventions, types of approaches, or different methodologies.

Impact concerns not only outcomes, but also the change that leads to outcomes, noted speaker Julia Compton of the UK Department for International Development. Impact evaluation, therefore, is a combination of both means—or processes—and ends, observed speaker Mary Lyn Field-Nguer of John Snow, Inc.

Speakers Compton and Sara Pacqué-Margolis of the Elizabeth Glaser Pediatric AIDS Foundation emphasized the long-term nature of impact evaluation. Compton observed the importance of considering longer term results, such as unintended effects and sustainability, in impact evaluation. She noted that the Organisation for Economic Co-operation and Development's Development Assistance Committee includes such concepts in its definition of impact.

In the context of evaluating the impact of PEPFAR, several speakers called for a shift to a broader definition of impact evaluation and to a more nonlinear concept of causation. Speaker Paul De Lay of the Joint United Nations Programme on HIV/AIDS (UNAIDS) observed that although the traditional definition for infectious disease impact evaluation of prevalence, incidence, infections averted, morbidity, and mortality is important, an assessment about what PEPFAR has actually accomplished should include broader concepts such as the intensity, quality, targeting, and equity of services. Key aspects of economic development and social change should also be built into the definition, he noted. Pacqué-Margolis added that a broader definition of impact evaluation would include measuring changes in health status, systems capacity, and other social, economic, and political outcomes. All impact evaluations must be based on a conceptual model of causation and intervention, observed speaker Nils Daulaire of the Global Health Council, but cultural-, political-, and location-specific factors and shifting influences make causality in the world of HIV/AIDS highly nonlinear. Innovative thinking may therefore be required when designing an impact evaluation for such a system.

Uses of Impact Evaluation

How is impact evaluation used? Workshop participants discussed uses of impact evaluation both to assess whether a project met its goals and to inform improvement in a project or indicate the need for midcourse corrections.

Summative Evaluation for Accountability and Advocacy

Impact evaluation has an important role in judging the performance of a program in order to account to specific constituents. In the case of PEPFAR, noted speakers De Lay and Agnes Binagwaho of the Rwanda Na-

tional AIDS Control Commission, impact evaluation is important to inform congressional decision makers, and the U.S. taxpayers they represent, about the success or failure of interventions. Evaluative information in turn has an important advocacy function, noted Pacqué-Margolis. To ensure sustained funding for a program, effort must be invested in interpreting and using research findings.

Speaker Daulaire and moderator Levine described the tensions that exist between attributing impact to particular funders and building knowledge—regardless of attribution to a particular funder—about what programs or interventions work. Even among congressional staff at the workshop, there was a difference of opinion about what type of attribution—to funders or to programs—is most useful. While speaker Savannah Lengsfelder from the Senate Committee on Foreign Relations emphasized the importance of focusing less on money and the return on investment per U.S. taxpayer dollar than on the success or failure of programs, speaker Christos Tsentas from Representative Barbara Lee's office acknowledged that data on what proportion of a program the United States supports are very helpful to promote a particular program. Attribution is further discussed in Chapter 4, in the section titled "Methodological Challenges and Opportunities in Evaluating Impact."

Formative Evaluation for Improved Decision Making

Workshop participants agreed that a formative, or utilization-focused, emphasis on evaluation—to improve decision making and course correction related to a particular program—is important for better and longer lasting results. "Qualitative and operations research are a critical part of the learning and doing process," said speaker Daulaire.

It is important to know who is taking the decisions informed by evaluation, observed speaker Compton: Who needs information at the top level, what information do they need, and when do they need it? What information is needed by the people who are taking decisions at the ground level? Some consideration for how the impact evaluation's results will be packaged, disseminated, and used at these multiple levels is important in evaluation design, added Pacqué-Margolis.

Speaker Jonathan Mwiindi of the Kijabe HIV/AIDS Relief Program in Kenya commented that decision makers at different levels may have different evaluation needs; sometimes the results and indicators of donors will benefit the donor but will be of little use to local operations. Speaker Binagwaho appealed to workshop participants to remember that local decision makers have decision-making needs that are just as important as those of top-level decision makers. "Impact evaluation should not just be for Congress," she asserted.

Decision making in Congress. Workshop participants discussed what Congress wants and needs to learn from the impact evaluation. Speaker Allen Moore of the Center for Strategic and International Studies emphasized that a primary role for Congress now is to design the reauthorization of PEPFAR. He noted that data from impact evaluation will help speed the reauthorization process, which could prevent the undesirable extension of the program on a yearly basis, via the Foreign Assistance Act, through an appropriations bill. More rapid progress on PEPFAR reauthorization would have added benefits, Moore noted, in sending signals to recipient countries to increase internal investments in health, to donors to shoulder their contribution to HIV/AIDS, and to implementing partners who have geared up, built up, and hired staff and need assurance that the program will continue to support their efforts.

Discussant Jim Sherry of George Washington University further reinforced the value of progress on reauthorization in leveraging investments by the United States, international partners, and local partners. Although we are still early in the epidemic, Sherry noted, since phase 1 of PEPFAR, there has been a 30-fold increase in funding from other U.S. sources, funding from other countries, and an increase in spending by national governments to about half of the total resources. Sherry also remarked that a change in political context—a new president and a Democrat-dominated Congress—at the time of PEPFAR reauthorization may have implications for the level of support for investments in global health.

Many workshop participants emphasized the importance of communicating impacts of key provisions of the PEPFAR legislation—such as the focus country model and the earmarks for investments across prevention, treatment, and care interventions—to congressional decision makers. Speaker Daulaire stressed that implementers need to convey to both advocates and policy makers what works and what does not, so that what is being pushed has relevance on the ground.

Programmatic decision making in partner countries. Impact evaluation also has value for decision makers at the level of partner countries. Speaker Binagwaho outlined the following benefits of evaluation results at the country level:

- To improve performance
- To continue and expand good initiatives
- To improve planning, monitoring, and evaluation
- To provide examples of implementation practices where available opportunities and resources have been optimally used

Speaker Jody Kusek of the World Bank emphasized the importance of relevant, high-quality, and widely disseminated evaluation information for optimizing the usefulness of evaluation results in changing practices and policies on the ground. In Swaziland, she noted, wide dissemination of the results of trials in Kenya and Uganda showing that surgically appropriate male circumcision provides significant protection against HIV infection has stimulated the crafting of a new policy to create better access to the intervention.

Speaker Field-Nguer stressed that the data collected by a program should be usable for program improvements and accessible and understandable to providers, managers, and policy makers. A closed communication loop between program managers and service providers about service data collected and reported is critical to program improvements.

PEPFAR'S EVALUATIVE APPROACH

PEPFAR's Approach to Strategic Information

Speaker Tom Kenyon, principal deputy U.S. Global AIDS Coordinator and chief medical officer of the Office of the U.S. Global AIDS Coordinator (OGAC), provided an overview of PEPFAR's strategic information—or monitoring and evaluation (M&E)—approach. Kenyon noted that the approach emphasizes the importance of sharing information not only vertically to the executive and legislative branches and to the taxpayer, but also horizontally to international partners and to country-level partners. PEPFAR's monitoring approach is based on "breaking out of the donor–recipient paradigm" to a model of true partnership in which multiple U.S. agencies are coordinating with host-country agencies. One aim of PEPFAR's monitoring approach is to stimulate a culture of accountability in partner countries through the establishment of monitoring systems, such as national health interview surveys.

A key aspect of PEPFAR's monitoring process and infrastructure is to track progress toward the program's goals in prevention, treatment, and care (see Box 1-1 for an overview of PEPFAR). Other M&E priorities include enhanced surveillance of behavior incidence and prevalence, HIV incidence and prevalence, drug-resistant HIV, HIV within the tuberculosis (TB) population, and drug-resistant TB. Mortality, morbidity, orphans averted, and social and economic change are other PEPFAR impact measures mentioned in a later presentation by Theresa Diaz of the U.S. Centers for Disease Control and Prevention.

According to Kenyon, both the overall budget for M&E and the proportion of the budget devoted to field versus central operations have increased over the period 2004–2007 (Figure 1-1). Budgets for strategic

BOX 1-1
Introduction to PEPFAR

In May 2003, the U.S. Congress passed the United States Leadership Against HIV/AIDS, Tuberculosis, and Malaria Act of 2003 (The Leadership Act) and established the U.S. Global AIDS Initiative. The legislation required the executive branch to establish a comprehensive 5-year strategy to combat HIV/AIDS, the President's Emergency Plan for AIDS Relief (PEPFAR). The legislation also established the position of the Global AIDS Coordinator within the U.S. Department of State to oversee and coordinate all U.S. international activities conducted by numerous U.S. government agencies, including the U.S. Agency for International Development; the Centers for Disease Control and Prevention, the Food and Drug Administration, the Health Resources and Services Administration, and the Substance Abuse and Mental Health Services Administration of the Department of Health and Human Services; the Department of Defense; the U.S. Peace Corps; the U.S. Census Bureau; and the Department of Labor.

The first 5-year phase of PEPFAR funding, from 2004 to 2008, is funded at $15 billion.

The U.S. Global AIDS Initiative focuses on 15 partner countries selected on the basis of their ability to scale up prevention, treatment, and care response by 2009.

PEPFAR's 5-year performance targets for the 15 focus countries include

- Prevention of 7 million new HIV infections
- Treatment of 2 million HIV-infected people
- Care for 10 million people infected with and affected by HIV/AIDS, including orphans and vulnerable children

These targets—generated using limited HIV incidence and prevalence data available in 2003—represented at the time of PEPFAR authorization about half of those eligible for treatment, half of those in need of care, and half of the new infections, said speaker Tom Kenyon of the Office of the U.S. Global AIDS Coordinator. There were assumptions that some HIV-positive individuals would never seek services and others would obtain services from the private sector.

SOURCE: Kenyon, 2007.

information are about 5 percent to 7 percent of overall country budgets in fiscal years (FYs) 2004–2007.

PEPFAR is expanding each country's reporting infrastructure and increasing the number of personnel who are trained in the field of strategic information. In country, the PEPFAR approach is based on the following principles:

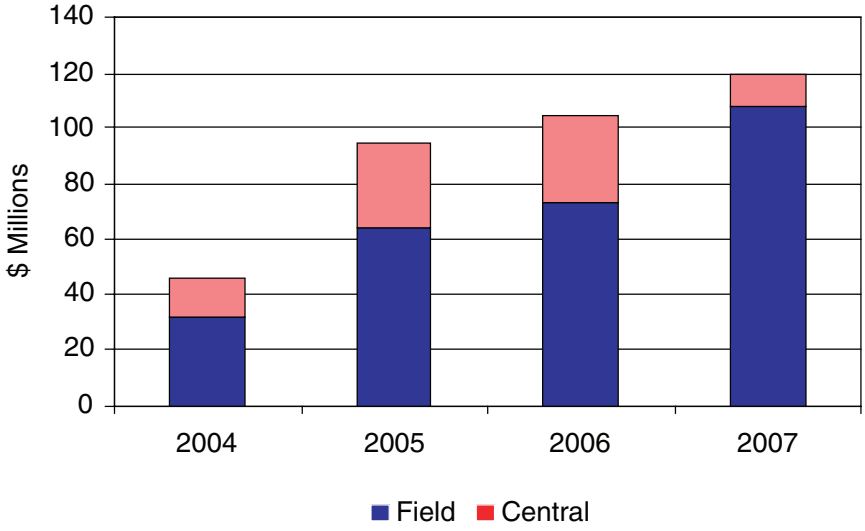

FIGURE 1-1 PEPFAR strategic information budget, 2004–2007.
SOURCE: Kenyon, 2007.

- Supporting local leadership and ownership of HIV/AIDS response by encouraging countries to develop one national plan, one coordinating mechanism, and one monitoring and evaluation plan[1]
- Building capacity of indigenous partners, including infrastructure and human capacity, through support of an annual implementers meetings
- Using local health infrastructure and community structures
- Implementing the program according to national guidelines
- Monitoring using internationally agreed-upon indicators
- Funding programs based on results

Kenyon presented PEPFAR's FY 2006 achievements, stating that 83 percent of partners were local organizations supporting 15,000 project sites:

- Antiretroviral treatment for 822,000 people; treatment programs projected to save 3.4 million life years by 2009

[1]One national plan, one coordinating mechanism, and one monitoring and evaluation plan are known as "the three ones."

- Prevention of mother-to-child transmission of HIV (PMTCT) services for women during more than 6 million pregnancies, averting an estimated 101,500 infant infections
- 18.7 million counseling and testing sessions for men, women, and children
- Care for nearly 4.5 million people, including more than 2 million orphans and vulnerable children
- Aversion of approximately 229,000 orphans through 2006 using treatment programs for parents; projected aversion of approximately 864,000 orphans through 2008

Evolution of Evaluation in PEPFAR

Speaker Kathy Marconi, director of monitoring, evaluation, and strategic information from OGAC, described how M&E have evolved over the life of the PEPFAR program and how they may change in the future.

Evaluation at the Inception of PEPFAR

PEPFAR's initial definition of impact—infections averted and lives saved—was narrow, linear, and basic. Capacity for monitoring and availability of data were similarly limited at that time. Within the U.S. government, each agency had its own reporting system, and there was a lack of consistent targets across countries. Program results were not yet based on robust data; for example, epidemic modeling was based primarily on urban antenatal clinic (ANC) sentinel sites. Although Marconi noted a general lack of country capacity to collect strategic information, some resources were available, such as in-country M&E leadership, a global commitment to M&E harmonization, international indicators of UNAIDS, reference groups working on modeling and M&E systems, and U.S. government technical resources for surveillance, surveys, information, and communication. PEPFAR's strategic information reporting was fed by an annual planning and reporting cycle for target setting, program implementation, funding tracking, and results.

Development of an Evaluation Framework

At the initiation of PEPFAR, evaluation planners established an "ideal" national strategic information system that would be informed by different surveillance approaches. Input on scale-up and coverage would be collected from facility surveys; estimates of behavioral change would be gathered from population-based surveys; and information on prevalence of HIV infection would be gathered from serum surveys at ANCs. Periodic targeted

evaluations would be conducted on selected topics. The evaluation framework developed using this simple logic model provided the results needed initially concerning, for example, the deployment and use of funds, the development and delivery of services, and the beneficiaries of program services. For example, in the context of treatment, the evaluation framework enabled PEPFAR to measure PMTCT, care, counseling, testing, scale-up over time, and people reached. Trends such as gender, children reached, infections averted, and years of life added through antiretroviral therapy can be measured using the framework.

Future Design of PEPFAR Monitoring and Evaluation

As the complexity of PEPFAR program strategies grows in the future, the definition of impact will need to be broadened to include factors such as long-term sustainability, health care workforce systems development, and other aspects of development such as nutrition, clean water, education, and gender equity. These factors will need to be reflected in the next evaluation framework.

Future M&E frameworks for PEPFAR may be informed by the development and strengthening of a number of new tools, including a UNAIDS methodology for helping countries define their own evaluation targets, health provider reporting systems, surveillance studies, population surveys, HIV testing tools, and supply chain management tools. A Global Fund impact study also will be available soon to provide information on disease rates, mortality, morbidity, and health systems within countries.

In the discussion following the presentation, speaker Marconi and workshop participants discussed some of the measurement challenges that will need to be addressed by planners of future PEPFAR evaluations. Several participants expressed concern that evaluation of progress on macro-level indicators needs to move beyond counting numbers of people touched by the program. For example, numbers of patients receiving drugs may not capture whether they are staying alive longer. Similarly, evaluation of orphan and vulnerable children programs may not capture impacts on improved nutrition, better education, and strengthened capability to be productive adults. Workshop participants also noted that predictive tools and evidence for assessing the effectiveness of prevention interventions at the country level are lacking. Marconi acknowledged that development of impact measures is still needed and in progress across areas such as quality of care, successful service treatment, treatment within different care settings, and effectiveness of prevention interventions. She also noted that there is still debate and lack of consensus on what appropriate impact measures should be in some areas. For example, she noted, years of life saved

is accepted now as a treatment indicator, but no consensus yet exists on whether mortality should be used as a treatment indicator.

Public Health Evaluation by the U.S. Global AIDS Coordinator

Although evaluations have been conducted routinely over the life of PEPFAR, as the program undergoes a transition from an emergency response to a sustainable strategy—and a transition from a country-level program to a global program—an expanded and broadened evaluation approach is now required. Shannon Hader, senior scientific advisor of OGAC, outlined PEPFAR efforts to develop public health evaluation (PHE), an approach that can be used to aggregate results across multiple countries, multiple time points, and multiple settings. In contrast to the targeted evaluations (TEs) used in the first phase of PEPFAR, which focused on immediate results for rapid project intervention and for individuals receiving services, PHE aims to improve services for communities and populations as a whole. The development of PHE will help to determine the effectiveness of interventions at the community and population levels, scaled-up services at the national level, and expansion of services and coverage to different types of populations, including difficult-to-reach populations. Ultimately, PHE is designed to support evaluation in order to strengthen scientifically sound and cost-effective methods of programming.

Structures and Administration of Public Health Evaluation

A new structure in OGAC has been established to support the quality, consistency, and coordination of PHEs and use of the results, methodologies, and tools generated. This structure includes a formalized, annual PHE priority-setting process to identify the most important questions for advancing PEPFAR impact. With oversight by a PHE subcommittee, PHE teams—drawn from PEPFAR headquarters, field offices, and partners—provide technical assistance for developing projects, establish common protocols (that is, guidelines for studies involving human subjects), and coordinate projects across countries. Figure 1-2 shows the PHE organizational structure.

PHE identifies priority issues for study not addressed by current evaluation projects; for example, the limited number of studies on behavioral outcomes led PHE to prioritize sexual transmission as one of the first evaluation projects. Other pilot PHE teams have been established to investigate care, treatment, and mother-to-child transmission of HIV (MTCT), and future teams are planned to focus on food and nutrition, orphans and vulnerable children, human capacity development, and counseling and testing. PHE teams emphasize an approach based on robust methodologies

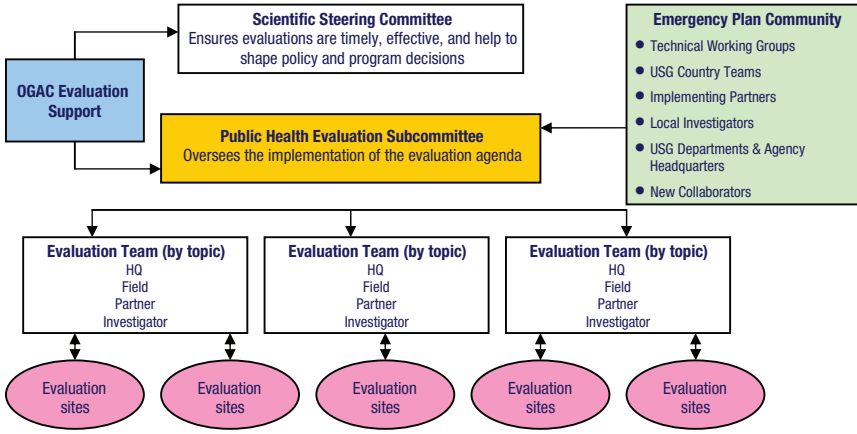

FIGURE 1-2 Structure of PHE.
SOURCE: Hader, 2007.

reviewed by scientific experts, the participation and capacity building of local investigators at multiple levels, and a clear plan of analysis. Involving implementers in evaluation design, implementation, and sharing of results is intended to contribute to ensuring independence and transparency of the evaluation.

Products of PHE

Although PHE does not extend to basic or investigational clinical research activities, it will result in the production of studies of program activities, characteristics, outcomes, and impact, which can in turn be used to determine program effectiveness, compare program models, and answer operational questions for implementation. In FY 2007, PHE conducted a combination of country-funded studies (110) and centrally funded (18) studies. The studies were characterized by many common areas of evaluation, including treatment, TB/HIV co-implementation, and MTCT. A key strength of PHE is its ability to aggregate data and therefore maximize investments across countries, in contrast to TEs, many of which have been limited in terms of expert technical assistance; partner experience in methodology; comfort with sampling, statistical, and analytic plans; and access to tools. PHE can be useful in providing technical support and in connecting groups doing similar studies (that is, on drug adherence or infant-feeding interventions) so that outcomes can be measured in similar ways for purposes of cross-country comparison.

PHE is also developing tools to ensure quality and consistency of data, enhance the capability to aggregate results across different countries and different settings, and set priorities using a more open and systematic process.

EVALUATIVE APPROACH AND MAJOR FINDINGS OF THE IOM PEPFAR EVALUATION COMMITTEE

Speaker Jaime Sepúlveda of the University of California–San Francisco provided background on the work of the IOM expert committee appointed by Congress to conduct an evaluation of PEPFAR implementation. The committee, which Sepúlveda chaired, began work on the project early in its implementation because the evaluation of PEPFAR was mandated to be delivered 3 years after the legislation was passed. Because of the time frame, it was only possible to evaluate the first phase of the implementation. Sepúlveda summarized the major conclusions and recommendations from the committee's report, *PEPFAR Implementation: Progress and Promise* (IOM, 2007) (see Box 1-2). Sepúlveda also offered the committee's perspective on the design of impact measures for future evaluation of PEPFAR.

BOX 1-2
Main Recommendations from IOM Evaluation of PEPFAR

Address long-term factors
- Emphasize prevention
- Empower women
- Build workforce capacity
- Expand knowledge base

Improve harmonization
- Improve coordination
- Support the World Health Organization prequalification process
- Remove budget allocations

Expand, improve, integrate services
- Data-driven prevention
- Adequate medications for treatment
- Community-based, family-centered care
- Target for orphans and vulnerable children
- Attention to marginalized populations

Transition to Sustainability

An overarching recommendation of the IOM committee was a needed shift in PEPFAR from an emergency relief mode to a greater emphasis on capacity building for sustainability. The need for continuity, improvement, and flexibility of programming are common themes of the IOM report. The committee underscored the importance of long-term factors, such as expanding and improving prevention interventions, empowering women and girls, strengthening the capacity of the workforce, and expanding the knowledge base through conduct and sharing of research. The committee also made specific recommendations to increase the program's flexibility and harmonization with other actors by supporting the World Health Organization (WHO) drug prequalification process[2] and removing specific budget allocations for prevention, treatment, and care.

Expansion, Improvement, and Integration of Services

The committee recommended the expansion, improvement, and integration of prevention, treatment, and care services, with emphasis on evidence-driven prevention interventions, an adequate supply of medications for treatment, and care based on a family-centered, community-based model. The committee emphasized the importance of evidence-based programming and robust and ongoing program evaluation. The committee highlighted the importance of addressing the needs of marginalized populations.

Design Considerations for Measuring Impact

Sepúlveda next described the committee's perspective on how PEPFAR's impact should be measured in the future. The committee urged PEPFAR to participate in joint attribution of outputs, outcomes, and impact with other actors in HIV/AIDS. The committee also recommended the development of both AIDS-specific and more general indicators.

[2]The Prequalification Project, set up in 2001, is a service provided by WHO to facilitate access to medicines that meet unified standards of quality, safety, and efficacy for HIV/AIDS, malaria, and tuberculosis. Any manufacturer wishing its medicines to be included in the prequalified products list is invited to apply. Each manufacturer must present extensive information on the product (or products) submitted to allow qualified assessment teams to evaluate the product's quality, safety, and efficacy. (See http://www.who.int/mediacentre/factsheets/fs278/en/index.html.) PEPFAR funding can be used only to purchase Food and Drug Administration–approved drugs.

AIDS-Specific Impact Indicators

AIDS-specific indicators should be developed that track change of the following "three generations" of HIV/AIDS surveillance: measurement of prevalence and incidence of HIV infection using specific state-of-the-art methods; measurement of behavioral change, including risky behaviors and risk-reducing behaviors; and assessment of stigma and discrimination. Other indicators to be developed include measures of survival, quality of life, development of drug resistance, and the overall physical, mental, and social well-being of people affected by HIV/AIDS.

General Impact Indicators

PEPFAR evaluation should also develop more general indicators, such as the empowerment of women and girls, general health (that is, infant mortality and overall mortality), capacity of community-based organizations to respond, and public health infrastructure and capacity (that is, supply chain and health care workforce). Among those indicators developed to track change in public health capacity, measures for monitoring the health care workforce are particularly important in light of the depletion of that workforce by the disease itself, the flux of health care workers to developed countries, and the sequestration of the health care workforce to vertical health programs. In addition, PEPFAR should also develop measures of the degree to which HIV/AIDS interventions are used to drive both desired improvements in the health system and incremental incorporation of other health priorities into national agendas. PEPFAR should contribute to the development of the knowledge base for how best to implement, scale up, and sustain prevention, care, and treatment services.

2

Envisioning a Meaningful Impact Evaluation for PEPFAR: Moving Beyond Counting

This chapter distills the results of a visioning exercise conducted by workshop participants to identify, frame, and prioritize the questions that matter for the evaluation of HIV/AIDS interventions by the President's Emergency Plan for AIDS Relief (PEPFAR). Workshop participants as diverse as congressional representatives, PEPFAR representatives, global partners, implementers, and country partners converged on some of the same impact evaluation questions.

Workshop participants stressed the importance of identifying meaningful and appropriate impact indicators and moving beyond quantitative outputs. Workshop speakers Jonathan Mwiindi of the Kijabe HIV/AIDS Relief Program, Kenya, and Mary Lyn Field-Nguer of John Snow, Inc., observed that the use of inappropriate indicators for evaluating program impact can sometimes tell an incomplete story, mislead, or mask other problems. For example, large numbers of individuals receiving first-line HIV treatment, Mwiindi noted, may mean that clinicians are not well trained to recognize when patients are in need of second-line drugs, not that the need for second-line drugs does not exist. Speakers Mwiindi and Kent Glenzer of the Cooperative for Assistance and Relief Everywhere, Inc. (CARE), described some of the drawbacks of and an overemphasis on quantitative measures. Mwiindi noted that an emphasis on counting sometimes drives inflated reporting as programs compete for limited funds. Glenzer observed that the

rush to produce quantitative outputs for goals like 2–7–10[1] can sometimes jeopardize sustainability. He called for a reduction in the number of indicators and a focus on a more strategic set of results. Such prioritization could be accomplished using a criteria-based approach, noted one participant. The generation of improved indicators, noted discussant Caroline Ryan of the Office of the U.S. Global AIDS Coordinator (OGAC), will ultimately provide better information for making programmatic course corrections.

Many participants argued in favor of a broader interpretation of impact as opposed to a narrower, or more pure, notion of impact. Many participants also believed that priorities and opportunities for evaluation include not only those describing the *results* of program implementation, but also the *process* of implementation—for example, coordination and capacity building—because the means of implementation are inseparable from the ends. Speaker Julia Compton of the UK Department for International Development (DFID) noted that defining the boundaries of what to measure—around what is AIDS and what is AIDS spending—is a challenge, and that defining the boundaries too widely or too narrowly involves risks. Defining boundaries includes answering the question of whether to measure direct AIDS impacts very narrowly or to measure broader impacts such as gender empowerment or land rights or health systems.

Workshop participants identified questions for evaluating impact that can be clustered into the following nine broad categories: cost-effectiveness, conceptual approach, health impacts, impacts beyond health, capacity building and health systems strengthening, coordination and harmonization, sustainability, equity and fairness, and unintended impacts.

COST-EFFECTIVENESS

Workshop participants described the importance of developing indicators that track the cost-effectiveness of PEPFAR. "Are we getting the biggest bang for our buck?" asked workshop speaker Christos Tsentas from the office of Representative Barbara Lee. Discussant Mead Over of the Center for Global Development emphasized that cost-effectiveness of interventions—and whether costs can be afforded in particular countries once donor funding is discontinued—has important implications for program sustainability. Cost-effectiveness measures can be used to evaluate different approaches, strategies, and interventions. Speaker Stefano Bertozzi of the National Institute of Public Health, Mexico, urged that cost-effectiveness measures be designed to assess types of prevention services delivered. He described a

[1]PEPFAR's 5-year goals—known as 2–7–10—are to support treatment for 2 million people, prevention of 7 million new infections, and care for 10 million people, including orphans and vulnerable children.

study of prevention service delivery in five countries in which a difference of three orders of magnitude in cost per service delivered was found: there were voluntary counseling and testing (VCT) programs that cost $7, $70, and $700 per client. Cost-effectiveness measures can help to ensure that VCT clinics not be established in contexts where there are not going to be any clients for them.

Discussant Over recommended that cost-effectiveness measures also be used to test AIDS treatment interventions. Such analysis can determine the overall balance of adverse biological effects of treatment (that is, the spread of resistance or the increased opportunity of infection of longer living HIV-positive individuals), adverse behavioral effects of treatment (that is, increased risk behaviors), treatment effects on nontarget populations, and treatment effects on orphans. He added that cost-effectiveness evaluation could be used to test the assumption that PEPFAR should focus support on government-financed, free, high-quality therapy as compared to high-quality private-sector therapy. He noted that the support of public health care may be a problem in countries where government is weak and where private health care is abundant, but of poor quality.

Workshop participants debated whether cost-effectiveness should be expressed in terms of results per donor investment versus results per approach investment (also see Chapter 1, "Meaning and Uses of Impact Evaluation," and Chapter 4, "Methodological Challenges and Opportunities in Evaluating Impact"). Even if attribution cannot be shown, observed workshop moderator Ruth Levine of the Center for Global Development, it is useful to show what programs or approaches have the greatest impact.

CONCEPTUAL APPROACH

A number of proposed impact evaluation questions centered on the conceptual approach of PEPFAR and the assumptions that had been made in the program's design: the countries and populations targeted, the budget allocations for different types of interventions, and the management and financing of the program. Such impact evaluation questions can be helpful, noted workshop speaker Compton, in thinking through the conceptual model of how PEPFAR inputs lead to outputs and outcomes.

Countries Targeted

Workshop speakers Allen Moore of the Center for Strategic and International Studies and Compton urged PEPFAR decision makers to evaluate the impact and usefulness of working with the 15 "focus countries." Is PEPFAR targeting focus countries appropriately? Compton added that impacts of regional factors should be evaluated; for example, focus-country

programs should track the impact of migration from a nonfocus country that is a major source of infection.

Populations Targeted

Several speakers urged the development of impact indicators that track whether the appropriate populations are being targeted. Speaker Compton cited work by David Wilson of the World Bank showing that although 76 percent of adult male infections in Ghana are linked to commercial sex workers, only 1 percent of World Bank prevention funding supports prevention programs for sex workers (MOH and ORC Macro, 2006). She and speaker Theresa Diaz of the U.S. Centers for Disease Control and Prevention emphasized the importance of evaluating the targeting of high-risk and marginalized groups—such as commercial sex workers and injection drug users—who are the main drivers of epidemics in many countries. Speaker Bertozzi further observed that because patterns of risk behavior may vary by country, PEPFAR may want to evaluate how effectively it targets not only infected individuals and orphans, but also populations where new infections are expected to occur.

Budget Allocations for Types of Interventions

Considerable discussion focused on the value of assessing the impact of the budget allocations for prevention, treatment, and care on PEPFAR's programming. Many participants cited the Institute of Medicine (IOM) study recommendations urging the removal of constraining congressional budget allocations (IOM, 2007). Observing that the current allocations came about as a result of highly effective advocacy by HIV treatment interests, speaker Field-Nguer called for an evidence-based assessment of the effectiveness of these allocations. Speaker Moore added that the usefulness of budget allocations and the appropriateness of funding priorities should be evaluated. Discussant Ambassador Jimmy Kolker, OGAC, suggested that evaluative information would be useful in informing future targets for scaling up prevention, care, and treatment interventions. Citing the imbalance in enrolling 90,000 people in treatment programs while 135,000 people were newly infected each year over the first phase of PEPFAR, speaker Compton urged the development of impact indicators to assess the appropriate balance among prevention, treatment, and care interventions. She also suggested an evaluation of the balance among interventions within prevention, noting that a significant proportion of the prevention funding is required to support abstinence and faithfulness (AB) programs.

Speaker David Gootnick of the U.S. Government Accountability Office suggested that the assessment of prevention interventions provides an

opportunity for a deeper level of analysis that could inform the dialogue about earmarks. Citing the Futures Group GOALS model (Stover and Bollinger, 2006), he suggested that specific interventions (that is, partner reduction, male circumcision, opt-out testing, abstinence programs, and sex worker peer education programs) can be evaluated for their effectiveness in averting infections. Another workshop participant suggested that impact evaluation be used to assess the degree to which earmarks and pipeline programming may have translated into a lack of integration of programs on the ground.

Several workshop participants stressed the importance of distinguishing between the evaluation of implementation of a particular policy on the ground and the evaluation of a policy as it is written. In many countries, noted speaker Jessica Price of Family Health International, there is still ambiguity in interpreting PEPFAR policies and how money can be used. For example, those receiving AB money interpreted PEPFAR policies to mean they could not educate youth about sexually transmitted infections. Similarly, many countries interpret differently whether or not PEPFAR funds can be used for family planning, and this is reflected in differential implementation on the ground. Speaker Tom Kenyon of OGAC added that often there has been misinterpretation or overinterpretation of what could not be done, particularly with regard to implementation of condom programming. He noted that PEPFAR has had to modify guidance because there was a retreat from programming using condoms and other prevention approaches.

Workshop participants stated that because earmarks have a political basis, as opposed to an empirical or a scientific basis, there may be less value in evaluating earmarks or in combining their evaluation with a more rigorous scientific evaluation. Speaker Jim Sherry of George Washington University noted that room is needed for political decision making about what is most important in terms of the evaluation agenda. Workshop participant Naomi Seiler from the U.S. House of Representatives Oversight Committee warned that it may be risky to mix scientific and objective evaluation with evaluation of more difficult political questions, such as evaluating a policy that prevents a certain type of intervention from taking place. As speaker Compton noted, "An evaluation cannot answer what goals are most important; that is a political decision."

Management and Financing

Workshop participants also encouraged the design of impact evaluation measures for the management and financing of the PEPFAR program across a variety of levels. One participant urged the evaluation of the impact of PEPFAR's practice of awarding hundreds of small grants to community organizations. Speaker Sara Pacqué-Margolis of the Elizabeth Glaser Pediatric

AIDS Foundation encouraged the development of impact indicators for performance-based financing systems. Speaker Agnes Binagwaho of the Rwanda National AIDS Control Commission called for a development of metrics to assess the effectiveness and results of program comanagement, such as that piloted between PEPFAR and the government of Rwanda. Speaker Bertozzi remarked that evaluation could be used to test the effectiveness of different technical assistance approaches in different contexts. He noted that PEPFAR, the World Bank, and The Global Fund have very different approaches in terms of levels of technical assistance, with PEPFAR providing intensive technical assistance and The Global Fund providing countries with funds to purchase their own technical assistance. Evaluation may reveal that some approaches may work more efficiently for certain contexts; for example, where there is high local capacity, The Global Fund approach might be more efficient, and in a country with lower capacity, the PEPFAR approach might be more efficient.

HEALTH IMPACTS

Workshop participants suggested the development of impact evaluation measures for tracking change in health, including indicators specific to HIV/AIDS as well as more general indicators.

HIV/AIDS-Specific Health Impacts

Prevalence, Incidence, Morbidity, Mortality, Longevity

Many workshop participants suggested that future impact evaluation of HIV-specific factors move beyond measurements of those accessing care and treatment to measurement of longer term health status. Because antiretroviral (ARV) treatment may achieve viral suppression in a small percentage of patients receiving drugs, observed speaker Bertozzi, measuring the number of people on antiretroviral therapy (ART) may not be enough to determine the lifesaving effectiveness of a program. Speaker Jaime Sepúlveda of the University of California–San Francisco urged the use of state-of-the-art methods to measure prevalence and incidence of HIV infection. Speaker Field-Nguer spoke of the importance of measuring HIV/AIDS mortality and morbidity, and speaker Savannah Lengsfelder from the U.S. Senate Committee on Foreign Relations called for the development of measures of PEPFAR impact on longevity.

Prevention of HIV Transmission

The full range of impacts of treatment on HIV transmission—both biological and behavioral—should be evaluated, urged discussant Over (Table 2-1). In terms of biological effects, ARV treatment can reduce viral loads, but it can have a negative impact on prevention because of resistance that comes about through imperfect adherence to drug regimens and a greater period of infectivity of longer living, HIV-positive individuals. He and speaker Mwiindi urged the development of indicators to measure the completeness of viral suppression, the level of drug adherence, and the level of resistance created by PEPFAR-supported treatment programs. In terms of behavioral effects, treatment programs can on one hand encourage health-seeking behaviors, including prevention behaviors, but those receiving ART as well as HIV-positive and HIV-negative individuals in the community may on the other hand engage in more risky or other adverse behaviors

TABLE 2-1 Possible Effects of ART on HIV Transmission

	Direction of Effect	
Type of Effect	Beneficial (Slow transmission)	Adverse (Speedy transmission)
Biological	Reduce infectiousness. ART may lower viral loads and may therefore lower the risk of transmission per sexual contact.	Select for resistance. Imperfect adherence to ART selects for resistant strains of the virus, which can then be transmitted.
		Longer duration of infectivity. The greater longevity of HIV-infected people taking ART has the unintended negative consequence of increasing the period in which the patient can transmit the virus.
Behavioral	Encourage prevention, especially diagnostic testing. ART may increase the uptake rates of prevention activities, particularly voluntary counseling and testing.	Offsetting behavior. People receiving ART, and HIV-positive and -negative people in the surrounding community, may engage in more risky behaviors than they would if ART were unavailable.

SOURCE: Over et al., 2007.

to preserve eligibility for programs than they would have if ART were not available. The extent to which treatment has contributed, through such behavioral channels, to widening the gap between the number of people on treatment and the number of people newly infected can be measured through evaluation. Speakers Mwiindi and Moore, along with moderator Levine, emphasized the importance of evaluating PEPFAR's success in prevention of mother-to-child transmission of HIV (PMTCT) and in preventing HIV infection in children, adolescent girls, and women.

Quality of Life

Many speakers identified an important need to broaden HIV-specific indicators to include those measuring quality of life, which speaker Sepúlveda described as an evaluation of overall physical, mental, and social well-being of people affected by HIV/AIDS. According to speakers Field-Nguer, Price, and William Holzemer of the University of California–San Francisco, other attributes of quality of life that could be measured through impact evaluation include restored productivity; effective management of chronic, secondary symptoms of ART patients; palliative care and use of analgesics to support dying with dignity; and family grieving.

Behavioral Change

Another major area of HIV/AIDS-specific health indicators that could be tracked is human behavioral change, which, according to workshop speaker Sepúlveda, includes the rates of both risky and protective behaviors. Speaker Tsentas enumerated several measures that could be developed to track changes in human behavior, including retention of patients on treatment, protected sexual intercourse, increased circumcision of males, and participation in needle-exchange programs. Discussant Over emphasized the importance of evaluating the impacts of treatment on prevention behaviors, including VCT. He argued that treatment may itself encourage or "disinhibit" risky behavior because it creates a perception that HIV/AIDS is now curable. He noted that such an effect has already been observed, for example, as decreases in condom use among sex workers in Kenya following announcements of "false cures" (Jha et al., 2001). This phenomenon is further discussed later in this chapter under "Unintended Impacts." Speaker Martha Ainsworth of the World Bank suggested that impact evaluation could be used to assess the effect of knowledge on behavioral change. She cited an example from Thailand in which the mere dissemination of knowledge that 44 percent of sex workers in Chiang Mai were infected with HIV had a dramatic effect on sexual behavior. Speaker Compton suggested that evaluation could be used to assess the impacts of a lack of investment in

specific behavioral change interventions, such as clean-needle distribution programs for intravenous drug users, which currently cannot be supported using PEPFAR funds.

Another workshop participant wondered if evaluation could assess the "dose response" of interventions required to maintain a behavioral change over time. Speaker Shannon Hader of OGAC asserted that measuring behavioral outcomes is much more difficult than measuring treatment outcomes. Tools and methods to measure and understand behavioral change outcomes are described further in Chapter 4, "Methodological Challenges and Opportunities in Evaluating Impact."

Stigma and Discrimination

A final area of HIV-specific assessment, noted speakers Sepúlveda and Field-Nguer, is stigma and discrimination. To what degree have PEPFAR's interventions contributed to promoting the acceptance of people living with HIV/AIDS in the community and to reducing stigma? Monitoring and evaluation (M&E) should consider the impact of stigma on a variety of factors among both people living with HIV/AIDS and health care workers, suggested speaker Holzemer and discussant Timothy Fowler of the U.S. Bureau of the Census. These include participation in HIV testing, use of services (that is, antenatal care), poor treatment by health care providers, adherence to medications, health status, and quality of life, such as loss of social support, isolation, violence, and limiting social interactions because of fear.

Other Health Impacts

Future impact evaluation of PEPFAR could also consider other health indicators in addition to those related to HIV/AIDS. Evidence from the Rwanda context, noted speaker Allen Moore, suggests that many health indicators beyond those specific to HIV/AIDS have improved as a direct outcome of PEPFAR investments. (This case study is described in greater detail in Chapter 4, "Measuring Impacts of Health Systems Strengthening.") Speaker Savannah Lengsfelder also spoke of the importance of evaluating any negative impacts of PEPFAR investments on other health conditions. Speakers Sepúlveda, Binagwaho, and Field-Nguer emphasized the importance of examining the effects of PEPFAR programs on mortality by all causes, overall survival, and lives saved from other diseases. Understanding the infant and child mortality contribution to overall mortality is an important piece of this. Speaker Pacqué-Margolis pointed out that attributes of reproductive health should also be monitored, including fertility, unintended pregnancy, and intended pregnancy. As for HIV/AIDS patients,

quality-of-life indicators should be tracked, noted speaker Tsentas, as well as the effectiveness of interventions in protecting, educating, and equipping people with tools they need to take care of themselves.

Speaker Lengsfelder suggested that data on other health indicators be disaggregated to show if differential effects are occurring in specific populations and in focus countries versus nonfocus countries.

IMPACTS BEYOND HEALTH

What have been the effects of PEPFAR interventions in the areas of gender equality, child welfare, security and development, and institutional change? Workshop participants discussed a number of measures of HIV/AIDS interventions beyond health that could be tracked through impact evaluation.

Gender-Focused Activities

Numerous speakers called for an evaluation of gender-focused activities, particularly those aimed at empowering women and girls. Gender-related dynamics occur in both men and women and influence their risk of contracting HIV, noted speaker Julie Pulerwitz of the Population Council. For men, gender norms encourage multiple sexual partners and early sexual debut. For women and girls, power imbalances result in an increased vulnerability to HIV/AIDS. As speakers Glenzer and Tsentas noted, power imbalances may be reflected in factors such as early and child marriage, weaker ability to negotiate sexual relations, susceptibility to pressure to engage in transactional and intergenerational sex, vulnerability to sexual violence, poor education, lack of economic opportunities and exclusion from control of strategic economic assets, exclusion from decision-making processes and from patronage networks, lack of property and inheritance rights, and lack of legal and enforceable rights.

Indicators for tracking such factors could be developed, suggested Glenzer, as well as indicators to measure the effectiveness of PEPFAR in addressing the underlying causes of women's and girls' vulnerability to HIV/AIDS. Glenzer also noted that building the analytical skills and competencies of men and women themselves around what is happening in their societies is a way to accelerate change in power structures and accelerate improvements, and the extent to which interventions build such capacities is something that can be measured through evaluation. Speaker Rachel Glennerster of the Abdul Latif Jameel Poverty Action Laboratory suggested that impact evaluation be used to identify the most effective messages in persuading teenage girls to change their behavior, and workshop moderator Levine suggested that programs be evaluated to assess their effective-

ness in reducing the spread of HIV infection to women and girls. Speaker Mwiindi introduced a nuance that the targeting and packaging of gender interventions should also be assessed through impact evaluation. Interventions aimed at sensitizing men may be perceived more positively and may be more effective in some cultures, he noted, than interventions targeting women.

Child Welfare

Several speakers highlighted the importance of evaluating parameters related to child welfare. Evaluation could measure the effectiveness of PEPFAR in raising the ability of adults to parent, suggested speaker Paul De Lay of the Joint United Nations Programme on HIV/AIDS (UNAIDS). Another participant observed that evaluation measures could track whether PEPFAR interventions had resulted in healthier, better nourished, and better educated orphans and vulnerable children who end up leading more productive lives as adults.

Security, Development, and Poverty Alleviation

Parameters of a country's security and development could be tracked using impact evaluation. Speakers Binagwaho and De Lay suggested that impact evaluation be designed to track the effects of PEPFAR interventions on peace and security in a nation, poverty alleviation and economic growth, and general national development.

Institutional and Societal Changes

Speakers De Lay and Field-Nguer suggested that impact evaluation of PEPFAR also focus on tracking how societies and institutions are changing and improving. Policy changes such as property rights, inheritance laws, and human rights could be monitored, as well as changes in the contextual environment more indicative of supportive systems, such as political will, community ownership, food and water security, engagement of vulnerable populations, destigmatization, and antidiscrimination measures. Speaker Compton raised the question of what technical assistance is doing to help change political views and priorities in a country.

IMPACTS ON SUSTAINABILITY, CAPACITY BUILDING, AND HEALTH SYSTEMS STRENGTHENING

Workshop participants emphasized the importance of monitoring both the process and results of building local capacity, particularly in the area

of strengthening health systems. Discussant Ambassador Kolker asserted that building local capacity is a key to scaling up programs and that an ideal for PEPFAR is to have countries with plans and capacity, for which all that is needed for national scale-up is funding. Building local capacity is an important prerequisite for handing off important aspects of the PEPFAR program, he noted. Speakers Lengsfelder and Tsentas raised the questions of whether PEPFAR's interventions were motivating and empowering partner countries to develop their own aggressive treatment and prevention strategies and contribute to the development of capable and self-sustaining national health systems and whether the process of building capacity was responsive to local needs.

Beyond measuring the effectiveness of the capacity-building process, evaluation can help to assess the results of capacity building. Evaluation can help to determine the impact of improved personnel staffing, training, and additional equipment on reinforcing the health system in general, noted speaker Binagwaho. Evaluation can also help define the extent to which overall improvements in the health system are attributable to PEPFAR's capacity-building inputs as opposed to other external changes in the environment and the populations served, suggested speaker Pacqué-Margolis. She added that evaluation can also inform us about any worsening of outcomes in health systems as a result of interventions.

Needs for tracking changes in the capacity of the health care system were identified in the areas of health care workforce, infrastructure, institutions, quality of care, the knowledge base, and the national-level health agenda.

Health Care Workforce

Measuring the Effects of PEPFAR on Workforce Shifts

The HIV/AIDS epidemic and interventions have had a dramatic effect on the health care workforce in many countries, noted speaker Sepúlveda. Health care workers themselves are among the populations that have become infected with and succumbed to the virus. In addition, "brain drain" has attracted health care workers both externally to higher salaries in developed countries and internally to generously funded vertical health programs in developing countries. Evaluation of PEPFAR should be designed to include an examination of the impacts of such shifts on the health care system, urged Sepúlveda.

Impact of PEPFAR's Training Approaches

Speaker Holzemer recommended that evaluative strategies be used to measure the impacts of PEPFAR's investments in different types of health care workforce training approaches. Distinguishing between two types of PEPFAR training strategies—*in-service education*, or developing skill sets across health provider groups in the existing workforce (that is, physicians, advanced-practice nurses, nurses, nurse assistants, community-based workers), and *preservice education*, or incorporating more people into the workforce—Holzemer suggested that evaluation be used to assess the qualifications, skills, competencies, and quality of health care workers trained using each approach. Impact evaluation could be used to test the assumptions behind the in-service training approach, in which skills are assigned to the lowest level worker possible. He noted that an in-service, or skill transfer, training approach assumes that it takes too long to train more nurses and other health care workers, and this urgency may not be justified. He added that the in-service training strategy also devalues clinical judgment and may result in a generation of poorly prepared health care workers who are insufficiently supervised and trained. Citing a lack of culture of continuous medical education in many African countries, speaker Mwiindi stressed the importance of evaluating the degree to which in-service training links new research and new findings to training programs.

Holzemer suggested that evaluation also be used to assess the effectiveness of PEPFAR's numerous "twinning"-based training programs, which involve the participation of partnering individuals and institutions from the United States. These include programs such as volunteer-based programs, institution-based partnerships, peer-to-peer collaborative relationships, professional exchanges and mentoring, and nonprescriptive demand- and process-driven partnerships. Holzemer pointed out that the model for such volunteer programs assumes that participants can afford financially to miss work for 3 weeks to serve and to train; however, many potential volunteer nurses are single parents, have families, and cannot afford to take time off from work. The assumptions and logic of such programs should be evaluated, Holzemer noted.

Speaker Field-Nguer suggested that it would be worthwhile to also assess the effectiveness of other workforce strengthening strategies, such as postgraduation fellowships with organizations that can mentor and give technical assistance, as well as recruitment of retired nurses back into the workforce. The effectiveness of training HIV-positive patients to serve as community health care workers could be evaluated, suggested speaker Mwiindi.

Possible measurements of workforce training programs should go be-

yond the number of persons trained, noted Holzemer, and could include the following:

- Length of time to train
- Degree to which clinical facilities are strengthened
- Impact of the least prepared worker on clients and patients
- Degree to which the capacity of qualified workers is developed
- The effectiveness of training and supervising the lowest level workers
- The degree to which the strategy provides opportunities for advanced professional development
- The degree to which the training program addresses gender inequality in the work environment
- Unintended consequences of the training, such as in-service training diverting workers from their jobs because higher pay is offered
- Knowledge, competencies, attitudes, and skills; types of positions taken; and quality of work environment

Impact of PEPFAR's Programs to Support Health Care Workers

A number of PEPFAR programs exist to provide support to professional and community-based caregivers and their families, including programs that provide HIV testing, treatment, and care for infected health care workers. Speaker Holzemer suggested that the effectiveness of such programs be evaluated. Metrics for evaluation of the effectiveness of these interventions might include, for example, whether services for health professionals are provided at different times of day from regular treatment programs, given professionals' desire not to wait in line and mix with clients.

Sustaining Local Workforce Development Systems

Speaker Holzemer noted the importance of evaluating how workforce development strategies complement and contribute to the sustainability of those systems in the country that controls the workforce. For example, there is strong potential for PEPFAR interventions to build the capacity of national regulatory bodies, professional associations (nursing and physicians' associations), educational accrediting associations, and government ministries. Through collaboration with such institutions, Holzemer noted, the careers of health care workers over time, from registration through retirement, could be tracked.

Institutions

Impact evaluation should be designed to assess the effectiveness of PEPFAR's institution-building efforts, said speaker Sepúlveda. He added that tracking change in the ability of community-based organizations to respond to the epidemic is an important consideration for impact evaluation design. Speaker Mwiindi added that evaluation should also consider the degree to which nontraditional institutions have been engaged or integrated in capacity-building efforts. He noted that there is underused potential for strengthening the capacity of the health system by involving institutions such as the church sector, which plays a substantial role in health provision in Africa, but is not traditionally recognized as a health institution. Mwiindi cited an example of a PEPFAR pilot site in Gezabi, Kenya, which did not have enough patients to receive drug treatment until religious leaders were engaged and trained to help identify and recruit patients.

Infrastructure

Speakers Sepúlveda and Mwiindi spoke of the need to evaluate the improvement of public health infrastructure, giving as an example the development of local supply chain management and drug delivery systems. Indicators could be developed, suggested Mwiindi, to measure the extent to which drug forecasting procedures and structures are in place to match resources to the need.

Quality of Care and Service Delivery

Quality of care and service delivery was highlighted as another aspect of health systems strengthening that could be evaluated. Speakers Mwiindi and Field-Nguer suggested a variety of indicators that could be used to track the effectiveness of PEPFAR interventions in the area of quality improvement. These include the quality and appropriateness of service delivery, and existing gaps, in the areas of prevention, care, treatment, support, and mitigation; ARV retention rates; levels of client satisfaction; appropriateness of referrals; and improved community attitudes toward people living with HIV/AIDS. Both speakers emphasized the importance of rational prescription behavior as a metric of quality of care. Field-Nguer commented that even within the same clinic, it is a challenge for health care workers to prescribe the right common antibiotic for the same syndrome for many sexually transmitted diseases. The more complex ARVs represent a further challenge, she said. Speakers Hader and Compton addressed a need for the development of broader metrics of quality. Metrics should be developed for assessing the quality of methodologies and of data, noted Hader.

In addition, noted Compton, information about the process of achieving quality should be collected through evaluation, including factors such as timeliness, access, follow-up, leadership, and management.

Knowledge Base

An expanded knowledge base is a critical determinant of PEPFAR's success, noted speaker Sepúlveda. Impact evaluation measures should be designed to track how PEPFAR has contributed to the development of an evidence base that includes information on what works best and how programs should be implemented, scaled up, and sustained. Moderator Levine asserted, however, that tensions may exist between building knowledge and responding to a serious public health problem. The appropriateness of the balance between implementation and strengthening the knowledge domain is an element that can be assessed through impact evaluation.

National-Level Health Agenda

Speakers Nils Daulaire of the Global Health Council and Sepúlveda broadened the discussion of evaluating PEPFAR's impacts on health systems strengthening to include assessment of how well PEPFAR has integrated with other health issues and has contributed to the change in national-level health agendas and priorities over time. Sepúlveda posited that a possible outcome of explicit PEPFAR interventions is to drive desired improvements into the health system, including incremental incorporation of other health priorities into country agendas, through what he termed a "diagonal" approach, a juxtaposition of both horizontal and vertical approaches. Daulaire added that PEPFAR support for broad maternal and child health systems should be monitored because such systems are key to the foundation for effective health systems overall.

COORDINATION AND HARMONIZATION

How integrated are PEPFAR systems with existing systems and programs, and how well is PEPFAR aligned with country priorities and plans? These are among the questions raised by workshop participants, who emphasized the importance of measuring the coordination and harmonization of program implementation. Participants also emphasized that the design and implementation of the impact evaluation itself should be coordinated and harmonized; this is described further in Chapter 3.

Valuable synergy can be achieved through coordination and harmonization. Moderator Levine observed that coordination and harmonization can show which programs or approaches have impact on specific outcomes.

Coordination and harmonization also bring together the different strengths of diverse actors involved in a program. Discussant Ambassador Kolker hoped that U.S. involvement and leadership in the HIV/AIDS response would motivate other donors to fill gaps in areas in which they have a comparative advantage relative to the United States. He observed that The Global Fund to Fight AIDS, Malaria, and Tuberculosis (The Global Fund), UNAIDS, and the World Bank have different advantages, histories, strengths, and implementation approaches that can be brought to the table in the response to HIV/AIDS. Although the United States is delighted to be in the lead, he noted, it is also delighted not to be the only player. Other workshop participants echoed this sentiment. For example, in comparing the U.S. response to that of DFID, Julia Compton observed that DFID's use of a more process- and sustainability-orientated approach, as opposed to an emergency-orientated approach, might create difficulties in achieving results quickly. Speaker Jody Kusek of the World Bank noted that her institution has a comparative advantage in helping countries build monitoring systems. Discussant Kolker and speaker De Lay cautioned that although harmonization, alignment, and process are important means to an end, they are not the exclusive determinants of a successful program.

A useful framework for considering coordination and harmonization across multiple levels, suggested by speaker Tsentas, is outlined in the following three sections on coordination among implementing agencies of the U.S. government, harmonization and alignment with partner countries, and coordination among program-implementing organizations such as donors, humanitarian assistance programs, and country-level actors.

Coordination Among Implementing Agencies of the U.S. Government

The HIV/AIDS epidemic has stimulated an unprecedented interagency response by the U.S. government (see Box 1-1), but this cooperative response has required a harmonization of approaches within the various implementing agencies, observed speaker Kenyon. For example, stated speaker Kathy Marconi of OGAC, within those U.S. agencies, there was initially a lack of consistent targets across focus countries because each government agency had its own reporting system; such systems have had to be coordinated and harmonized through the PEPFAR program.

Among the main needs identified by workshop participants for evaluation of coordination and harmonization among U.S. government agencies are the impacts of complementary interventions that are essential for managing HIV/AIDS specifically and health generally in developing countries but that go beyond narrowly focused HIV/AIDS services. These so-called wraparound programs include investments in areas such as malaria, tuberculosis, nutrition, food security, social security, education, child survival,

family planning, reproductive health, medical training, health systems, and potable water. Speakers Compton and Pacqué-Margolis suggested that evaluation could be helpful in assessing the effectiveness and constraints of the wraparound model, and speaker Lengsfelder suggested that the effectiveness of integration between HIV/AIDS services and complementary services also be measured. Finally, speakers Ainsworth of the World Bank and Pacqué-Margolis suggested that impact evaluation could be used to measure negative impacts of such programs, such as those created by distorted incentives. If provision of a service is contingent on a patient being infected with HIV, noted Ainsworth, there is a risk that someone who needs the service equally or even more may not receive it. If nutritional support programs are discontinued at a site, inquired Pacqué-Margolis, will patients stop seeking HIV services? A similarly dramatic example was described by speaker Binagwaho, in which a Rwandan child interviewed on television said he wished his mother were HIV positive so he could go to school. Potential negative impacts of complementary interventions are further described later in this chapter under "Unintended Impacts."

Harmonization and Alignment with Partner Countries

A second level of harmonization identified for future evaluation by workshop participants is harmonization with partner countries. Workshop participants value harmonization and alignment of PEPFAR with country priorities and plans because they believe this leads to greater success of the overall program. For example, ownership and local engagement, said speaker Daulaire, are critical to both financial and operational sustainability. However, several participants spoke of challenges and constraints to harmonization with partner countries. Compton pointed out that in some cases, depending on national systems for data collection can be problematic if those systems are weak. Discussant Ambassador Kolker added that where national leadership and national ownership exist, donor alignment can work, implying that the absence of national leadership and ownership makes alignment with partner countries more challenging. Speaker Binagwaho summarized that it is unclear what level of alignment is optimal; she and speaker De Lay observed that impact evaluation could help assess the effectiveness of varying degrees of alignment of PEPFAR with national development plans.

Speaker Mwiindi suggested that the extent to which interventions are culturally contextualized is an important component of evaluating harmonization with partner countries. For example, he noted that interventions for orphans and vulnerable children should take into account that child care is a community-based effort in many cultures. Similarly, how messages about male circumcision are communicated should consider that the prac-

BOX 2-1
Comanagement of PEPFAR in Rwanda: A Case Study

In Rwanda, agreements have been created to align PEPFAR, The Global Fund, and the World Bank to Rwanda's national HIV/AIDS plan. For example, all donors must align to the national format for quantification, designed for the district level. Realignment of PEPFAR's 2–7–10 plan to the Rwandan indicators took about one and a half years. The majority of PEPFAR investments are also comanaged with the Rwandan government. Committees that make decisions about the use of funds are chaired by a partner country national. At the technical level, U.S. government institutions, Rwandan government institutions, other donors, and civil society work together. For example, field visits to assess progress are conducted jointly by representatives of PEPFAR, The Global Fund, and the government of Rwanda. Local institutions also set policies and standards for implementation. Plans and projects are open and accessible at the community level, and the government of Rwanda shares progress among donors in a transparent fashion. The comanagement approach is a key to ownership and success because the host country is part of the decision and leads the decision-making process.

SOURCE: Binagwaho, 2007.

tice is not culturally accepted in certain areas. Citing the experience of comanagement of PEPFAR between the U.S. government and the government of Rwanda (see Box 2-1), speaker Binagwaho described other measures that could be tracked to evaluate whether harmonization mechanisms have provided effective and demonstrable results. These include the openness and accessibility of donor plans and projects at the community level, the presence of instruments and structures for joint decision making, and the degree of information sharing and joint implementation among partners.

The alignment of PEPFAR with national priorities in Uganda was another example discussed. Citing a perceived criticism in the IOM committee's report that alignment had been difficult in the early years of the Uganda partnership because of the many funding earmarks and restrictions, discussant Ambassador Kolker acknowledged that although U.S. efforts were not a perfect match to local priorities, they were congruent with the national plan. The United States does not claim to be the only actor in the field, and PEPFAR does not claim responsibility for everything, he stated; in the Uganda case, the United States was able to match its unique capabilities and expertise to local needs and opportunities. Kolker added that about

half of the focus countries have taken advantage of abstinence waivers,[2] a mechanism that provides some flexibility in the coordination process. The existence and effectiveness of such mechanisms might be included in future evaluations of the impact of PEPFAR's harmonization efforts with partner countries.

Coordination Among Program Implementers

A third area of coordination identified for prospective evaluation is the coordination among program implementers. Workshop discussant Ryan suggested that the development and effectiveness of a variety of coordination tools, or mechanisms for fostering the exchange of learning, be evaluated. Speaker Field-Nguer gave an example of such a coordination tool: an implementers' group that has been formed, through John Snow International and World Learning with Global Health Council, to foster learning about what is happening and what could be improved. This group has been using the experience of implementers in the field to influence the reauthorization process. Speaker Marconi described the development of two coordination tools whose effectiveness could be evaluated: (1) the PEPFAR extranet, a mechanism for information sharing accessible to all U.S. government employees, and (2) systematic literature reviews on the latest intervention research. The systematic reviews are conducted by the Cochrane Group and circulated to PEPFAR partners all over the world.

Speakers Mwiindi and Compton suggested that harmonization of the drug procurement system can provide a useful case study for assessing coordination among program implementers. Mwiindi noted that PEPFAR's entry into the drug supply chain—and the program's exclusive sourcing of U.S. Food and Drug Administration (FDA)–approved drugs or branded drugs—has resulted in parallel streams of drug qualification, delays in drug delivery, and inconsistency with national protocols. Compton called for an impact evaluation of the policy requiring FDA approval of drugs distributed through PEPFAR. Mwiindi added that the effectiveness of PEPFAR's harmonization with existing drug procurement systems also be evaluated, which might include tracking measures such as drug quality, drug cost, reliability of the drug supply, and level of engagement with organizations on the ground, such as religious groups, that function in the existing health care delivery and drug supply system.

[2]PEPFAR partner countries can apply for a waiver that allows them to reapportion their prevention funds among abstinence–faithfulness–condom use (ABC) interventions to reflect the nature of the local epidemic.

SUSTAINABILITY IMPACTS

Workshop participants appealed for the development of evaluation measures that would assess PEPFAR's effectiveness in evolving into a long-term, sustained response. Moderator Levine, along with speakers Moore and Gootnick, emphasized that making the transition to sustainability will be a challenge for what Moore termed a "hurry-up effort" that was legislated with a sense of urgency and began as an emergency response. Other participants stressed the importance of program sustainability given the long-term nature of the HIV/AIDS disease. Discussant Phillip Nieburg at the Center for Strategic and International Studies noted the importance of acknowledging that HIV/AIDS is neither an epidemic nor a pandemic, but endemic, which implies a sustainable response. Speaker De Lay also underscored the need for PEPFAR to ensure effective chronic care for infected persons into the future.

Measuring the financial sustainability of PEPFAR was a major focus of discussions, and speaker Compton warned that the financial implications for PEPFAR not continuing into the future would be tremendous, given the substantial financial burden that would be passed on to countries. Speaker Mwiindi suggested that evaluation could help prompt the development of and measure the progress toward a clear exit strategy. Measures could be developed, for example, of the degree to which the PEPFAR program has leveraged additional resources by other donors and by national governments, remarked workshop discussant Sherry.

Workshop participants also emphasized the importance of evaluation for long-term learning about the program. Moderator Levine suggested that evaluation can help to generate a technical consensus and a stock of knowledge about what works. Speaker Ainsworth added that information on what combination of services is most effective—a potential product of evaluation—can help to improve the efficiency and sustainability of the program.

Evaluation of PEPFAR's impacts on sustainability needs to include measures of capacity building and promotion of local independence, workshop participants said. Speaker Compton noted that evaluation should assess the degree to which PEPFAR is making a sustained contribution to institutions and systems that include research, M&E, policy, budgeting, planning, and programming systems, as well as the degree to which PEPFAR is sustaining public-sector and voluntary staffing. Speaker Pulerwitz advised that evaluation of sustainability also take into account the degree to which "least dependency" practices are in place that are conducive to local implementation, ownership, and coordination with national systems and structures.

EQUITY AND FAIRNESS IMPACTS

Speaker Norman Daniels of the Harvard School of Public Health focused his remarks on the importance of evaluating the fairness of the PEPFAR program. He introduced the concept of fairness and defined its subcomponents—equity, accountability, and efficiency. He provided examples showing the trade-offs and tensions between the competing goals of equity and efficiency in decision making. Finally, he offered guidance on indicators that can be developed to examine issues of equity as PEPFAR programs are scaled up.

Defining Fairness

Aspects of fairness are central in decisions to scale up ART treatment and prevention, noted Daniels. In the context of domestic and international health policy, the goal is to improve population health, as measured in the aggregate, but also to distribute the benefits of health policy in a fair way and reduce disparities. These dual goals may conflict with each other, however, due to conflicting priorities. A framework integrating concerns about equity, accountability, and efficiency can be used for evaluating the fairness of health interventions in developing countries (Box 2-2). Benchmarks can be developed for each component and indicators generated to measure the improvement that a particular intervention produces for each of these components.

Trade-offs Among Competing Goals in Evaluating Fairness

The evaluation of fairness involves not simply the maximization of a specific outcome, but the way in which resources are being used to achieve several goals, which may at times conflict with each other. Trade-offs can exist, when attempting to achieve the goals of equity and efficiency, between getting the best outcomes with scarce resources and giving people fair chances at some benefits.

Several examples of trade-offs in evaluating fairness were shared.

World Health Organization 3 by 5 Initiative

Issues with equity implications have arisen in the context of scale-up decisions of the World Health Organization's (WHO's) 3 by 5 Initiative,[3] such as cost recovery, eligibility criteria for patient selection, site selection,

[3]The WHO 3 by 5 Initiative was a global target to provide ART to 3 million people living with HIV/AIDS in low- and middle-income countries by the end of 2005.

BOX 2-2
The Elements of Fairness

- **Equity:** A key aspect of equity is the absence of unjustifiable inequalities or disparities across demographic groups. Such inequalities might include gender inequality, urban–rural inequality in access, child versus adult inequality, inequality of vulnerable and stigmatized groups, and inequality of immigrant and migrant populations, which are large in some high-prevalence countries. Benchmarks focusing on various aspects of equity include the equity for exposures to risk through intersectoral public health issues, financial and nonfinancial barriers to access to care, different levels of coverage in different parts of the health system, and equity in financing.

- **Accountability:** The acceptance of responsibility for and readiness to justify decisions, acts, or failures to act has both value as a means to achieving performance and intrinsic value, in that people want to know something about how decisions were made. Accountability works at three levels. "Upward" accountability to congressional leaders is important for justifying how taxpayer money was spent and whether the goals authorized by decision makers were achieved. "Horizontal" accountability among funders and donors providing technical assistance is important for cooperation. A "downward" accountability refers to the responsibility to populations affected by the decisions and implementation of the program. Accountability to program beneficiaries reinforces ownership of goals within a population, transparency about how the goals were established, and commitment to sustaining achievement around those goals over time. Benchmarks for accountability include democratic accountability and empowerment, and patient and provider autonomy.

- **Efficiency:** Efficiency is the use of scarce resources in a way that gives value for money. Benchmarks for efficiency include both clinical- and administrative-level efficiencies.

SOURCE: Daniels et al.,1996.

and selection of practitioner groups or health workers (Daniels, 2005). In the case of cost recovery, there are concerns about barriers to access created by charging for drugs, but user fees are arguably critical for sustainability of a program, a key element of efficiency. The issue of site selection was similarly controversial. Although mobilization of resources and trained personnel in tertiary care centers may be technically the most efficient solution for rapid scale-up, a concentration of service delivery in areas where the largest numbers of people can be reached most rapidly leaves people in rural areas without a fair chance of any benefit because of poor access to services.

Integration of Treatment with the Health System

Another example of the tension that exists between health maximization and equity concerns the integration of ART treatment services with the rest of the health system. Despite direct investments in health system improvement, there is concern about unintended effects of PEPFAR implementation that may undermine other health programs in the form of negatively affecting personnel distribution in countries or draining parts of the health system toward politically driven programs. Daniels said that while AIDS interventions could be steering resources disproportionately away from other health problems, a focus on health system strengthening from the AIDS effort could also broaden the assistance with regard to other diseases.

Equal Access to Care Through Randomized Trials

Speaker Glennerster raised some of the ethical tensions that exist between the goals of accessing care and generating knowledge about what interventions are effective through randomized trials. Because randomized trials traditionally require a control group, there is an ethical dilemma of excluding people from access to a program that might save their lives.

Prevention and Treatment

Another workshop participant raised the example of tension between the prevention and treatment approaches of PEPFAR, that is, between the people living with the disease who will die without treatment and those who will benefit from effective prevention programs—the uninfected population plus future generations. Daniels observed that while the ethical argument for prevention was initially based on "best outcomes" because it assumed that infected people would not have any chance of benefit and that not as many people would be helped overall if treatment were provided, that context has changed by the decreasing cost of treatment and by increasing political pressures favoring treatment.

Indicators for Evaluating Fairness

Daniels provided guidance on two types of measures that could be developed to assess the fairness of PEPFAR: (1) evaluation of the process of goal setting and (2) implementation and evaluation of the level of the program's integration within the health system. To better understand and address basic aspects of equity, Daniels stressed, it is important in any evaluation to disaggregate data and to collect it uniformly. Disaggregated data can tell us much about inequality related to gender, urban–rural access, age,

employment status, vulnerable and stigmatized groups, and immigrant and migrant populations. For example, noted speaker Binagwaho, evaluation of disaggregated data can tell us about the degree to which all geographic regions of a country and marginalized groups—prostitutes and prisoners—are benefiting from services. In Rwanda, she mentioned, prisoners now have access to the same prevention, care, and treatment services as communities, thanks to a successful national plan. The effectiveness of such programs can be assessed only if data on populations served are disaggregated.

A Fair Process in Goal Setting and Implementation

Processes for airing disagreements, surveying stakeholders for their input, and finding fair and legitimate solutions are constructive mechanisms for resolving tensions and for setting and implementing goals (Daniels and Sabin, 2002). Such processes have been used by the National AIDS Commission in Malawi and by planners of Mexico's national insurance program. In Malawi, the National AIDS Commission set one of the best examples of a public ethics discussion by holding public hearings, involving stakeholders as members of the AIDS Commission, and publishing reports on decisions taken, including reasons why minority positions were not adopted (Daniels, 2005). Speaker Binagwaho contributed yet another example of a public ethics discussion from the Rwandan experience, in which people living with HIV/AIDS can themselves have a voice through their election to district-, provincial-, and national-level decision-making bodies. It is possible to develop indicators to measure whether these types of processes are occurring as decisions are made about program implementation.

Integration with the Health System

Assessment of unintended fairness impacts requires monitoring of the extent to which a particular scale-up program is integrated with the health care system. A key monitoring indicator is population-disaggregated, georeferenced information about the health system level of the site at which people receive treatment. Such data enables identification of where the benefits of scale-up are being delivered and to what parts of the population.

Daniels noted that the issue of brain drain of trained health workers can be an element of evaluating PEPFAR integration with the broader health system. Useful information would include data on health personnel in parts of the health system that are adjacent to the ones where scale-up sites are being established. Such information would show whether health workers are being pulled out of places where services are needed.

How developed countries meet their health care needs with the training of health workers and the impact of not meeting health care needs in devel-

oping countries can also be explored through evaluation. Daniels noted that the United Kingdom has now put in place a specific code to correct some of the drain on human resources from the developing countries, including making monetary contributions to Malawi to eliminate some factors that were driving health workers out of the country. The existence and effectiveness of such compensatory mechanisms could be assessed through impact evaluation.

UNINTENDED IMPACTS

Workshop participants discussed the need for evaluation to assess whether PEPFAR has had unintended consequences. Speaker Compton pointed out that both positive and negative synergies can occur and that methodologies are needed for capturing contextual data that will allow detection of these unintended impacts. Workshop participants discussed potential unintended effects of PEPFAR on program integration, diversion of resources from other health areas, corruption, access to services, adverse and high-risk behavior, nutrition, and reproductive health and family planning.

Program Integration

One workshop participant observed that evaluation could be used to assess the unintended impacts of earmarking of prevention funds on program integration. The participant hypothesized that earmarking of prevention funds may have resulted in more pipelining of funds, which in turn may have prevented integration of programs on the ground as separate groups of contractors arose to conduct AB programs independently of condom use programs, but few integrated abstinence–faithfulness–condom use (ABC) programs have arisen.

Diversion of Resources

Speakers Compton and Lengsfelder suggested that evaluation be used to consider unintended effects of PEPFAR on other diseases or health care areas. If shifts in emphasis have occurred, asked Lengsfelder, which areas are receiving the least attention? Compton suggested that the budgets and outcomes of malaria and less popular diseases should be tracked. Speaker Daulaire brought up the specific example of monitoring the change in allocation of resources to family planning, which he believed should be more actively sought as part of PEPFAR's HIV/AIDS prevention and control strategy. While HIV programming from the United States has increased from $120 million a year in the mid-1990s to $5.3 billion, he noted, fam-

ily planning funding in the same period has been reduced 14 percent, even though the number of women of childbearing age has increased by 30 percent.

Workshop participants also discussed the importance of monitoring unintended impacts of PEPFAR on aspects of the broader health care system, such as the workforce and infrastructure. Compton suggested that evaluation assess the macroeconomic impacts of PEPFAR funding on brain drain from essential services resulting from hiring and per diem practices. Speaker Mwiindi proposed that unintended effects of PEPFAR's large, multicountry supply chain program on existing supply chains be evaluated. Possible outcomes such as trading volume imbalance or brain drain from existing supply chains to the new system could be assessed.

Discussant Ambassador Kolker suggested that evaluation be used to assess the extent to which PEPFAR's investments in other diseases, in the health system, and in complementary development services (antipoverty programs, family-planning services, nutrition, women's rights) have served as a counterweight to the diversion of attention from other areas.

Corruption

Speaker Compton suggested that evaluation be used to assess the effects of PEPFAR on corruption. She asked whether corrupt practices, such as double counting of infrastructure, are occurring in the context of PEPFAR implementation.

Access to Services

Speaker Pacqué-Margolis expressed interest in measuring possible positive or negative effects of PEPFAR on the access and use of other services. For example, are wait times becoming longer and decreasing access, or are access and use of other services in the context of HIV/AIDS programs improving?

Adverse and High-Risk Behavior

Discussant Over noted that HIV/AIDS treatment programs may be inadvertent incentives for adverse behaviors, such as lack of drug adherence to maintain low CD4 (cluster of differentiation antigen 4) counts[4] in order to qualify for disability, or expression of desire to become HIV positive to

[4]CD4 is the receptor for HIV predominantly found on the surface of T-lymphocytes. The CD4 count is the number of helper CD4 T-lymphocytes in a cubic millimeter of blood. The absolute CD4 count declines as HIV infection progresses.

qualify for programs. Effects such as these should be monitored through impact evaluation. As Over described previously in this chapter in the "Behavioral Change" section of "HIV/AIDS-Specific Health Impacts," the availability of treatment may also encourage high-risk behavior because of the perception that AIDS is now treatable. Speaker Compton reinforced this argument by drawing from the experience in Zambia, where there is anecdotal evidence that abstinence programs have been successful in increasing the age of first sexual relations but have simultaneously decreased the amount of condom use. Compton asked if the emphasis on abstinence and faithfulness approaches may be unintentionally increasing the stigma of condom use.

Nutrition

Several speakers raised the importance of evaluating unintentional impacts of complementary programs, such as nutrition and food aid. Speaker Binagwaho mentioned the example of a nutrition program that pushed some HIV-positive families to have a child in order to obtain food; evaluation may shed light on the relationship between nutrition programs and childbearing. Speaker Compton noted that the effects of food-aid distribution by PEPFAR on local markets also could be evaluated.

Reproductive Health and Family Planning

Unintended impacts of PEPFAR on reproductive health and family planning was identified by workshop participants as an area of interest for evaluation. Several participants noted that PEPFAR's reproductive health programs may have had some positive synergies with HIV prevention, treatment, and care programs. One participant noted that family planning was the most cost-effective way to prevent mother-to-child transmission and urged impact evaluation to look more closely at the links between reproductive health and changing sexual behavior for the prevention of infection. Speaker Kenyon observed that through HIV counseling and testing procedures, more women than ever are receiving needed attention during pregnancy. At the same time, Kenyon noted an example of the potential for negative synergies between treatment and reproductive health in that women on ART in one country have been getting pregnant at a high rate. He observed that family planning guidance had not been updated in that country since 1995 but needed to be revisited in light of HIV and availability of ART. Discussant Over highlighted the phenomenon of continued or restored fertility during treatment, observing that treatment programs may thus create more orphans. Citing data from work in India and other countries, Over posited that while treatment programs can avert orphan-

hood years among a woman's current children, new orphanhood years can also be created among children born during the extended life span of a woman on treatment.

Speakers Daulaire and Kenyon suggested that evaluation could be used to assess positive or negative synergistic effects of PEPFAR's policies not to procure family planning commodities and contraceptives. While Daulaire argued that nonprocurement of such commodities may undermine HIV/AIDS prevention and control, Kenyon asserted that such investments may divert resources from, and thereby weaken, HIV-related activities.

3

Designing an Evaluation That Incorporates the Guiding Principles of Coordination, Harmonization, and Capacity Building

Workshop participants focused their discussions not only on *what* should be evaluated, but on the ideal *process* of designing and conducting the evaluation itself. Participants stressed the importance of an earnest effort to design and conduct the evaluation in a way that truly incorporates the principles of coordinating evaluation efforts among global partners, harmonizing with evaluation needs of country partners, and contributing to strengthened local evaluative capacity. Many lessons for designing the process of evaluation can be learned, participants noted, from previous experience with harmonization, coordination, and capacity building in the context of implementation of human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) interventions through the President's Emergency Plan for AIDS Relief (PEPFAR). This chapter summarizes the benefits, costs, and opportunities of coordination and harmonization, as well as for capacity building in evaluation.

BENEFITS, COSTS, AND OPPORTUNITIES OF COORDINATION AND HARMONIZATION IN EVALUATION

Benefits of Coordination and Harmonization in Evaluation

Workshop participants outlined the value and benefit of coordination and harmonization in impact evaluation. Discussant Jim Sherry of George Washington University noted that coordination and harmonization in evaluation design and implementation are important in influencing

others' work, which is in turn critical for addressing strategic questions at a broader policy- or program-practice level. It is useful to distinguish, he noted, between an internal, institution-specific evaluation that relates to tracking needs for program survival and direction change and a shared international evaluation that relates to influencing how other actors use resources and to strengthening overall capacity. Given the high program transaction costs of evaluation for countries and partners, collaborating as much as possible also minimizes work and ensures more efficient use of funds, observed speakers Julia Compton of the UK Department for International Development (DFID) and Agnes Binagwaho of the Rwanda National AIDS Control Commission. Speaker Sara Pacqué-Margolis of the Elizabeth Glaser Pediatric AIDS Foundation further underscored the importance of coordination to ensure maximum use of evaluation dollars, noting that successful completion of impact evaluations involves high human and financial resource costs, an extensive time frame, and serious commitment. Speaker Mary Lyn Field-Nguer of John Snow, Inc., pointed out that the strengths and perspectives brought to the table by focus-country government partners, implementing partners, and other stakeholders are a further benefit. She and other workshop participants outlined the value each stakeholder brings to the evaluation process.

Partner Countries

The value of engaging partner countries is their accountability to their citizens, observed Field-Nguer. Government partners were dealing with the challenges of service delivery to their populations long before PEPFAR and can provide a critical perspective on health system issues, such as health care workforce and supply chain issues, and how PEPFAR is addressing these, she added. Even in the context of an emergency situation, involvement of partner countries in the design of evaluation questions and methodology can improve the quality of the overall evaluation design and the interpretation of evaluation findings, noted speaker Binagwaho. Speaker Kathy Marconi of the Office of the U.S. Global AIDS Coordinator (OGAC) stressed that involvement of partner countries can support the process of developing evaluation priorities that are actually meaningful to the countries, and this has implications for sustainability.

Implementing Partners

The value of engaging implementing partners is their familiarity with program data and lessons and their understanding of the challenges of delivering services across a continuum of care, stated Field-Nguer. Coordination with implementing partners, with their knowledge of the lessons of

decades of AIDS work, for example, can contribute to avoiding mistakes, such as repetition of research to answer questions already answered and implementation of program strategies that do not work. Implementers add value not just as reporters of data and information, but also as users of data and information, and hence need to be involved in the evaluation design process, observed speaker Nils Daulaire of the Global Health Council.

Beneficiaries and Other Stakeholders

Local people add value to the evaluation process because they can place change in context, said workshop participant Joanne Capper of the U.S. Peace Corps. Speaker Field-Nguer added that service beneficiaries are the experts on program impact and are an audience that can define the prevention, care, and treatment service parameters of acceptability, accessibility, and affordability. In the end, patients, clients, and community members define the effectiveness of care services. However, she noted, few studies are available about the community level of knowledge.

Field-Nguer identified specific groups of stakeholders who can add value to the design phase of impact evaluation through choosing the questions and the methodologies. These groups include the following:

- People living with HIV/AIDS
- Women
- Youth, including those in and out of school
- Other populations at risk
- Government ministries—beyond ministries of health—and local management units
- Health care workers at all levels in urban and rural settings
- Health facilities
- Nongovernmental and community organizations working inside and outside facilities
- Community leaders
- Religious leaders

Constraints of Coordination and Harmonization in Evaluation

High transaction costs are among the greatest constraints to coordination and harmonization in evaluation, workshop participants said. There is tension between the benefits of coordinating and taking advantage of existing synergies by linking with others' evaluation work and the costs of that coordination, remarked workshop moderator Ruth Levine of the Center for Global Development. "The mere sharing of information is a huge task that can be all consuming," observed Ambassador Jimmy Kolker, OGAC.

In addition to being costly, the processes of design, consensus building, commitment, planning, and working with host countries are very complex, asserted speaker Pacqué-Margolis.

Speaker Paul De Lay of the Joint United Nations Programme on HIV/AIDS (UNAIDS) pointed out that the transaction costs of inclusiveness and consensus building are particularly pronounced when immediate results are desired, as was the case for PEPFAR. However, he noted, in order to sustain desired results, harmonization, integration, and sharing across partners will be necessary. Overtaxing or overextending evaluators with efficiency-level questions was a further constraint to coordination identified by discussant Sherry. Questions at this level prevent partners from influencing others' work and being influenced. He emphasized that the focus of coordination and harmonization should be on strategic, broad, or program-level questions.

Institutionalizing accountability to program beneficiaries through coordination may be constrained when there is a large power imbalance such as that existing between donors and partner countries, observed another workshop participant. When countries depend on donor resources, which could potentially be removed, there is concern that countries may not stand up to donors and speak up for what they want, particularly if a strong coordinating plan and leadership are not in place. Institutionalization of such accountability, noted speaker Norm Daniels of the Harvard School of Public Health, depends heavily on the effectiveness of the national coordinator and the national plan. Donors who view country ownership as a desirable objective also must be open to acknowledging what counts as fair and reasonable at the national level.

Opportunities for Coordination and Harmonization in Evaluation Efforts

Workshop participants described a number of opportunities for coordination and harmonization with global partners and country partners in evaluation design and implementation.

Coordinating Evaluation Design, Conduct, and Results with Global Partners

Workshop participants strongly articulated the need for PEPFAR to coordinate its evaluation efforts with other global partners. PEPFAR should not just be learning from its own evaluations, said speaker Rachel Glennerster of the Abdul Latif Jameel Poverty Action Laboratory, it should be looking at work being done elsewhere. Another participant stressed that such coordination should include divulging and

sharing negative results with global partners. Speaker De Lay noted that in parallel to PEPFAR's evaluation efforts, The Global Fund to Fight AIDS, Malaria, and Tuberculosis (The Global Fund); the World Bank; and UNAIDS are conducting evaluative efforts, and bilateral efforts, such as DFID's AIDS program, are also undergoing evaluation. Although these different evaluations address unique issues, constituencies, and time frames, there is strong potential for sharing data, approaches, and evaluation research. De Lay suggested that outcomes of PEPFAR evaluative efforts be tied to and shared with other global and bilateral evaluations of AIDS institutions and initiatives. Several of these efforts are described in greater detail below.

The Global Fund. Speakers De Lay, Compton, Marconi, and John Novak of the U.S. Agency for International Development (USAID) supported the suggestion that PEPFAR collaborate with The Global Fund evaluation, particularly with regard to broader questions such as systemwide effects. As De Lay noted, The Global Fund is just starting a 5-year evaluation with a specific set of questions about the effectiveness of the funding models, level of partner support, and technical assistance. The Global Fund evaluation will devote nearly $15 million of the $17 million total evaluation budget to research, using prospective study survey evaluation research methods. Discussant Kolker noted that OGAC is currently in close contact with The Global Fund impact study organizers. Speaker Theresa Diaz of the U.S. Centers for Disease Control and Prevention (CDC) added that an OGAC technical group will work with The Global Fund impact evaluation technical group to review all methodologies and analyses used. In every PEPFAR country, a group of government contacts has been designated as part of the task force working with The Global Fund on the impact evaluation, she said.

UNAIDS. Speakers De Lay, Marconi, and Stefano Bertozzi of the National Institute of Public Health, Mexico, along with discussant Kolker, suggested that PEPFAR engage collaboratively with UNAIDS. According to De Lay, UNAIDS had a major evaluation in 2001–2002 and will now start a second evaluation. These evaluations have focused on the role and impact of UNAIDS cosponsors in a changing environment. Bertozzi noted the particular value of engaging with the economics reference group at UNAIDS; Marconi mentioned the strengths of the UNAIDS modeling reference group in developing global-level impact measures, such as those for stigma and gender discrimination. Kolker noted that the United States is the main supporter of the UNAIDS monitoring and evaluation (M&E) reference group.

World Bank. PEPFAR collaboration with the World Bank on evaluation may be beneficial in developing an integrated HIV/AIDS research agenda, noted speaker Jody Kusek of the World Bank. She said the World Bank evaluation will focus on the impact of technical assistance such as treatment, scale-up of treatment facilities, prevention, programs that affect policy instruments (such as cash transfer policy instruments as incentives for behavioral change), and socioeconomic impacts of HIV/AIDS; the effectiveness of HIV/AIDS programs in achieving goals; and the design of new investments to ensure that impact can be assessed.

OECD. Collaboration with the Organization for Economic Cooperation and Development (OECD) on evaluation may be helpful in the area of HIV/AIDS that deals with gap analysis and meta-evaluations, speaker Compton said.

WHO. Discussant Kolker reported that OGAC has been involved in the annual meeting on the World Health Organization's (WHO's) HIV/AIDS impact evaluation.

Coordinating with partners beyond the AIDS community. Discussant Sherry suggested that drawing on evaluation expertise beyond the AIDS community, such as in the areas of sustainability and broader development issues, may be beneficial to the evaluation of PEPFAR.

Coordinating Evaluation Design, Conduct, and Results with Country-Level Partners

Workshop participants described several opportunities for harmonization of evaluation with country partners. Speaker Pacqué-Margolis asserted that the evaluation function should be prioritized in country operation plans; partners should be included in evaluation planning; and funding mechanisms should promote harmonization. Workshop discussant Phillip Nieburg of the Center for Strategic and International Studies envisioned that country partner organizations would be given an opportunity to review drafts and submit comments and concerns in the evaluation process. Speaker Binagwaho and workshop participant David Stanton of USAID highlighted the importance of community-level interpretation of evaluation results.

Several speakers pointed out that impact evaluation should mirror, and perhaps draw lessons from, the experience in coordinating and harmonizing with partner countries in program implementation. Field-Nguer reminded participants that the guiding philosophy of PEPFAR, the "three ones," includes the concept of a single monitoring and evaluation plan shared with partner countries. Speaker Daulaire commented that lessons about

coordination and harmonization also can be drawn from the experience in universal childhood immunization and smallpox eradication, in which stakeholder engagement was done effectively and with a view toward sustainability.

Specific opportunities for country-partner harmonization on evaluation are described below.

Joint field evaluations. Discussant Kolker remarked that aid effectiveness could be improved if common program and sectoral approaches used joint evaluation visits based on a single national plan and a single monitoring and evaluation system. Speaker Binagwaho offered the Rwandan experience—in which partners from PEPFAR, The Global Fund, and the government of Rwanda routinely conduct joint field visits for evaluation purposes—as a potential model.

Centralized funding and data aggregation. Echoing points made in Shannon Hader's presentation about the value of a central coordinating hub for impact evaluation, speaker Pacqué-Margolis argued that a dedicated, central funding source is operationally better suited to coordinate the evaluation effort across countries. She noted that funding mechanisms that send money to the country level do not promote harmonization on evaluation because countries often have little money left over for M&E. Pacqué-Margolis further stated that evaluation findings need to be aggregated across countries—also by a central coordinating unit—to be meaningful and to see patterns.

Knowledge management. Speaker Field-Nguer spoke of the opportunity to use evaluation coordination to develop a mechanism for knowledge management, both for existing data and for methodologies to gather new data.

Harmonization of fairness-monitoring approaches. Speaker Daniels suggested that harmonization be used to work toward agreement on which aspects of equity, accountability, and efficiency will be part of the M&E program. He noted that such discussions provide an opportunity to clarify the overall program objectives.

BENEFITS, CONSTRAINTS, AND OPPORTUNITIES OF BUILDING CAPACITY IN EVALUATION

Benefits of Building Capacity in Evaluation

Workshop participants emphasized the importance of designing an evaluation that itself strengthens local capacity. In-country capacity is

needed not just for service delivery, noted speaker Pacqué-Margolis, but for M&E and for continuous quality improvement, advocacy, national planning, and budgeting. Building local evaluative capacity has particular benefit as PEPFAR undergoes a transition from an emergency program to a long-term, sustained program, speakers noted. In an emergency response, observed speaker Bertozzi, the time to build local capacity on evaluation issues is limited. In contrast, a long-term response involves, for example, not just training existing health practitioners but also educating new ones, and not just involving local researchers in a project, but also building the capacity of local health researchers to do prospective evaluations. Field-Nguer added that local capacity and systems to collect, analyze, and use program information are of critical importance to program success and sustainability and should be built into the process of PEPFAR program implementation and impact evaluation.

Constraints to Building Local Capacity in Evaluation

Lack of systems for gathering data, inadequate funding mechanisms, and poor stakeholder engagement are among the constraints to building local capacity for evaluation, workshop participants said. Speaker Kusek observed that many countries are unable to take advantage of the 5 percent to 10 percent of total project budget available for evaluation under World Bank loans because they often don't have systems in place for gathering data. There is powerful competition for resources between the development of monitoring systems and the implementation of the program, she said. Pacqué-Margolis observed that although funding mechanisms should develop and promote capacity for evaluation, information dissemination, and advocacy, they often fall short. Funding earmarks and directives limit the prioritization of M&E, operations research, clinical research, and advocacy in Country Operation Plan planning. Poor or distorted engagement of stakeholders is another constraint to building local capacity for evaluation. Many partners, including country-level stakeholders, do not have a place at the table in defining the research agenda and conceptual frameworks, noted Pacqué-Margolis. In contrast, researchers tend to drive the evaluation agenda according to their own interests, noted Kusek, using resources to conduct more narrowly focused impact evaluations that are less helpful.

Opportunities for Strengthening Evaluative Capacity

Workshop participants suggested a number of opportunities for strengthening local capacity to conduct impact evaluation. Diaz suggested that donors could disseminate suggested methodologies, offer technical assistance, conduct training workshops, and engage in one-on-one mentor-

ing. Binagwaho added that the recruitment and training of independent consultants in partner countries is another way to develop competency in evaluation methods that will remain in the country. Country-driven decision making and priority-setting processes were also suggested as potential models or tools for strengthening local evaluative capacity. Binagwaho suggested that the experience of comanaged decision making in Rwanda potentially could be applied to evaluation design. Workshop participant Laura Porter from CDC suggested that a country-driven evaluation priority-setting process—drawn up by a multidisciplinary team—could be another mechanism for building local capacity. She noted that a need for such a process was articulated at a recent meeting of the M&E Reference Group Evaluation Subgroup.

4

Designing an Impact Evaluation with Robust Methodologies

This chapter summarizes workshop discussions on methodological issues related to impact evaluation design for the President's Emergency Plan for AIDS Relief (PEPFAR) and is divided into three sections. In the first section, a diverse set of case studies of conceptual models and methodological approaches are presented from previous large-scale evaluations—from the World Bank, the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute of Technology (Poverty Action Lab), the UK Department for International Development (DFID), the Cooperative for Assistance and Relief Everywhere, Inc. (CARE), and The Global Fund to Fight AIDS, Tuberculosis, and Malaria (The Global Fund). In the second section, methodological challenges and opportunities of impact evaluation are described for the measurement of outcomes and impacts specific to human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS), for the measurement of more general outcomes and impacts, for attribution and accounting, and for the aggregation of impact results. The third section summarizes themes common to the approaches.

CONCEPTUAL MODELS AND METHODOLOGICAL APPROACHES: CASE STUDIES

Impact evaluations require the development of a conceptual model. The model must be defined, the inputs and outcomes measured, and assumptions and conversion factors determined. For prevention of mother-to-child transmission of HIV (PMTCT), noted speaker Sara Pacqué-Margolis of

the Elizabeth Glaser Pediatric AIDS Foundation, there is a clear, logical pathway between access to services, counseling and testing, test results, prophylaxis by women and infants, and aversion of infections. Assumptions and conversion factors to be determined for PMTCT can include questions like the following: What regimens are taken and how effective are they? Are they actually consumed and when? What is the rate of transmission during labor and delivery? What is the rate of prevention of infections in HIV-negative women who come in for counseling? What is the level of infection transmitted through breast milk? Speaker Carl Latkin of the Johns Hopkins School of Public Health cautioned that although models of change are needed to guide interventions, sometimes they don't explain findings. Models are practical heuristics but should not be blinders, he noted; we should not let models narrow the way we look at change.

Impact evaluations also require the use of methodological approaches. These can include quantitative, qualitative, and participatory methods and theory-based program logic. Examples of impact evaluation methods, provided by speaker Mary Lyn Field-Nguer of John Snow, Inc., include client satisfaction interviews and surveys, exit interviews, mystery clients, targeted intervention research, focus groups, and key informant interviews.

The following case studies describe the experiences from evaluations of five HIV/AIDS assistance programs run by the World Bank, Poverty Action Lab, DFID, CARE, and The Global Fund. Conceptual models and different evaluation methodologies are described in the context of each study.

World Bank Evaluation of HIV/AIDS Assistance Programs

Workshop speaker Martha Ainsworth, lead economist and coordinator of the Health and Education Evaluation Independent Evaluation Group at the World Bank, described the approach and methodologies used in an independent evaluation of the World Bank's HIV/AIDS assistance programs. The evaluation assessed $2.5 billion of World Bank investments in HIV/AIDS prevention, care, and mitigation programs between 1988 and 2004 in 62 developing countries. Two objectives of the evaluation were defined: (1) to evaluate the development effectiveness—or relevance, efficiency, and efficacy—of HIV/AIDS assistance in terms of lending, policy dialogue, and analytic work at the country level relative to the counterfactual, or absence of a Bank program and (2) to identify lessons to guide future activities.

Ainsworth shared the World Bank's experience in prioritizing what to measure in evaluation. Although the World Bank has a large portfolio of complementary programs in education and agriculture, indicators were narrowed down to only those with direct HIV/AIDS outcomes and impacts. In addition, identifying how lessons from completed assistance were still relevant to new approaches posed a challenge, given that three-quarters of

the HIV/AIDS assistance programs being evaluated were still in progress. In assessing a long-term, ever-changing implementation approach over time, therefore, the World Bank evaluation was designed to select those issues that were common to all projects, such as political commitment, setting strategic priorities, multisectoral responses, ministry of health role, use of nongovernmental organizations (NGOs) in implementation, and monitoring and evaluation (M&E). The World Bank evaluated the projects completed in the past and examined those issues relevant to ongoing projects. Through this approach, the assumptions and design of the ongoing portfolio were analyzed and prospectively evaluated. The World Bank was able to consider design issues and point out where risks had been mitigated and where problems could be addressed through midstream adjustments.

The World Bank evaluation drew on a number of methodological approaches. As Ainsworth noted, the World Bank does not rely exclusively on a single source of information, but rather uses different types of evaluations already occurring in the context of the work, such as midterm reviews, completion reports, and annual reviews. Evaluation methods used include the following:

- **Results chain documentation:** Inputs, outputs, outcomes, and impact of government, the World Bank, and other donor efforts were gathered.
- **Time lines:** The documentation of timing of efforts was collected, although in many activities this type of M&E information is lacking.
- **Interviews:** Some information was elicited from interviews of stakeholders, other donors, people and staff involved on the ground, and government implementers.
- **Desk work:** The following were collected and analyzed: literature reviews; archival research; interviews on the time line of World Bank response; an inventory of analytic work; a portfolio review of health, education, transport, and social protection sectors; and background papers on national AIDS strategies.
- **Surveys:** Surveys were conducted of staff members, audiences for analytic work, project task team leaders, and country directors.
- **Field work:** Project assessments and case studies—chosen to reflect different levels of experience and where interventions worked or did not work—were collected and reviewed. For example, a project in Indonesia, canceled because the World Bank intervention occurred before anyone was visibly ill, was chosen for the evaluation, as was a project in Russia, where only policy dialogue and analytic work were conducted.

Use of Randomized Controlled Trial Methodologies to Evaluate HIV/AIDS Programs

Rachel Glennerster, executive director of the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute of Technology, described the application of randomized controlled trial methodology to HIV/AIDS program evaluation. She described the advantages and disadvantages of randomized trial methodologies and then discussed the results from two case studies in which randomized methods were used, an evaluation of an HIV education program in Kenya and an HIV status knowledge program in Malawi.

Advantages and Disadvantages of Randomized Evaluations

To know the true impact of a program, one must be able to assess how the same individual or group would have fared with and without an intervention. Because it is impossible to observe the same individual in the presence or absence of an intervention simultaneously, comparison groups that resemble the test group are commonly used. Common approaches for selecting comparison groups include a "before and after" approach, in which the same group of individuals are compared before and after exposure to an intervention, and a "cross-sectional" approach, in which, at a single point in time, a group of countries or communities in which an intervention has occurred are compared to a "non-intervention" group. However, programs are usually started in particular places at certain times for a reason, and they are usually established with the countries, communities, schools, and individuals most committed to action. Therefore, estimates of program impact may be biased because it is difficult to find a comparison group that is equally committed to those where the program was established. This may in part explain why projects typically work well in a few places, but fail when scaled up. In randomized controlled trials, like medical clinical trials, those who receive the treatment and the control group are selected randomly. By construction, those who receive the proposed new intervention are no more committed, no more motivated, no richer, and no more educated than those in the control group. Randomized trials produce results that are freer from bias than other epidemiological studies. Randomized evaluations can be used to test the efficacy of interventions before they are eventually scaled up to the national level.

Randomized trials conventionally have been used to look at drug effectiveness, but are also being applied to other areas where they are not commonly used. For example, randomized trials can be used to investigate social patterns, such as what messages are most effective in changing the sexual behavior of young girls.

There is a perception that randomized evaluations are difficult both to implement and to integrate with what is going on at the ground level, but with innovations in randomization over the past 10 years, randomized studies are less intrusive and less like more formalized clinical trials. Several mechanisms exist to more naturally introduce randomization into the way a government works or with the way an NGO works on the ground, including the following:

- **Lottery:** Randomization can be introduced through a lottery if a program is oversubscribed.
- **Beta testing:** Randomization can be introduced through small-scale experimentation of methods before scaling up to the national level.
- **Randomized phase-in over time and space:** Capacity or financial constraints may limit the ability to introduce interventions in all communities immediately. The order in which a program is phased in can be randomized, allowing for an assessment of effectiveness to be made during the phase-in period.
- **Encouragement design:** Often, national programs that are up and running do not have 100 percent adoption; the impact of such programs can be evaluated by randomly encouraging some people to participate in the program.

Several of these mechanisms simultaneously help to address some of the ethical questions surrounding randomized design—the exclusion of people from having access to care or programs that might save their lives. In the randomized phase-in approach, all individuals will ultimately benefit from the intervention; under the encouragement design, no one is denied care.

A disadvantage of randomized evaluation is that it cannot be done after the fact; it must be implemented with the program. Institutional constraints are another disadvantage to randomized evaluation that sometimes make it more difficult to engage with partners in an intensive way. One workshop participant noted that randomized controlled trials can be difficult to translate from the individual level to the community level, where interventions are more complex. Glennerster acknowledged that randomized controlled trials can be improperly designed and can thereby generate incorrect results.

Using Randomized Trials to Evaluate HIV/AIDS Education Programs in Kenya

Randomized trial methodology was used to evaluate a Kenyan HIV/AIDS education project, a collaborative effort among the government of Kenya, a local NGO, U.S. universities, and Jomo Kenyatta University in

Kenya. The method was used in randomly chosen schools to test a range of education strategies for their effectiveness in getting children to understand messages about the risks of HIV. These strategies included training teachers in a new HIV/AIDS education curriculum, reducing education costs to encourage young girls to stay in school, holding debates about whether or not to teach about condoms in primary schools, holding essay competitions about protection from HIV, and telling children about relative infection rates by age, including the dangers of sexual, gift-exchanging relationships between young girls and older men (sugar daddies), the greater likelihood of older men to be infected than younger men, and the greater likelihood of girls to be infected than boys. Upon implementation of each program, the evaluation tracked observed changes in behavior, including school dropout rates, marriage, pregnancy, and childbirth, as determined by community interviews. Follow-up studies are also tracking HIV infection rates under each type of intervention.

Results from the trial are shown in Figure 4-1.

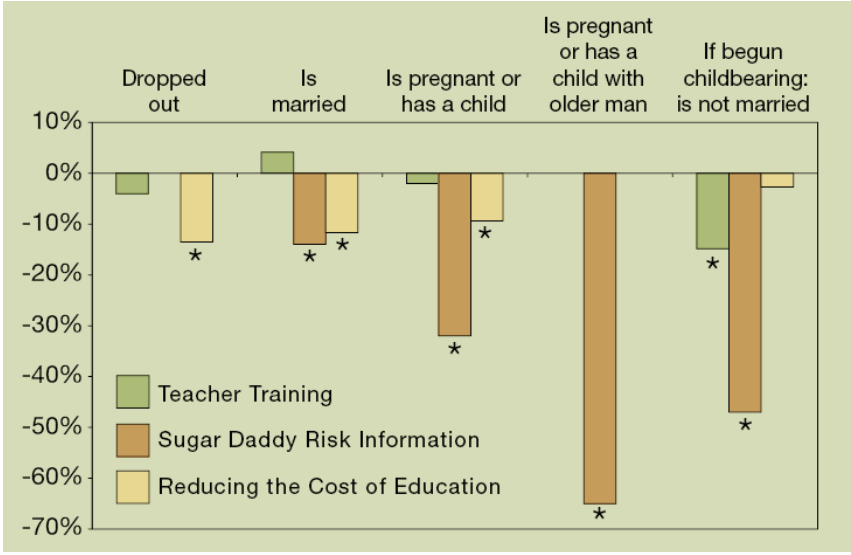

FIGURE 4-1 Impacts of alternative HIV/AIDS education strategies on girls' behavioral outcomes.
NOTE: ★Indicates that the difference with the comparison group is significant at 10 percent.
SOURCES: Duflo et al., 2006, and J-PAL, 2007.

The teacher training in the national curriculum had little effect on school dropout rates, marriage, and childbirth, although girls from schools where the training was conducted were more likely to be married if they had a child, and there was a slight effect on increasing tolerance of those with HIV in schools that underwent the training. Reducing the cost of education was found to be an effective strategy for reducing dropout, marriage, and childbirth rates. Education programs about the dangers of sexual relations with older men, or sugar daddies, led to a 65 percent drop in pregnancies or childbirths with older men and no increase in childbearing with younger men. Self-reported data indicated a shift between having relationships with older men to having relationships with younger men. Self-reported data from the boys in the group indicated increased condom use, potentially because boys had learned that girls were much more likely to be infected than boys. Results of the debate and essay interventions remain to be tested with outcome data; currently, only self-reported data exist, which can be very biased. On the basis of the costs of the interventions, the evaluators were able to calculate a cost-per-childbirth-averted rate for each intervention, with the education program about older men being the most cost-effective intervention, at $91 per childbirth averted, compared to $750 per childbirth averted for interventions to reduce the cost of schooling.

Using Randomized Trials to Evaluate HIV Status Knowledge Programs in Malawi

Although half of HIV/AIDS prevention spending in Africa focuses on HIV testing, many of those tested do not come back to pick up their results. A study conducted in Malawi used randomized evaluation to test the impact of campaigns promoting knowledge of HIV status (Thornton, 2007). Only 40 percent of those tested for HIV returned to collect their results, but the study showed that a small incentive—only 10–20 cents, or a small fraction of the daily wage—was enough to increase results collection by 50 percent. The study went on to test whether or not knowledge of status changed behavior. In follow-up interviews with those who had and had not received encouragement to pick up their test results, people were given the opportunity to buy subsidized condoms and the money to buy them. In comparing the treatment group (those encouraged to and therefore more likely to know their status) with the control group (those who were not encouraged and thus less likely to know their status), the study found that the knowledge of HIV status had virtually no impact on whether people purchased subsidized condoms, even when they were given the money to buy them. Only HIV-positive individuals in long-term partnerships were more likely to buy condoms if they knew their status, and few bought subsidized condoms.

Glennerster cautioned that if randomized methodologies are not used and if studies survey only the sample that returns for test results, it may appear as if knowledge of status is effective in reducing HIV incidence. A randomized methodology allows researchers to tease out proper attribution for the perceived success of a program. Glennerster also noted that the use of plausible correlation approaches—suggested by workshop speaker Paul De Lay of the Joint United Nations Programme on HIV/AIDS (UNAIDS) as a more practical methodology applicable to work at the country level—without doing a full trial can also lead to the wrong policy conclusion. With millions of dollars being invested in knowledge-of-HIV-status programs, it is worth testing whether they are effective in reducing incidence, she concluded.

DFID Evaluation of the National HIV/AIDS Strategy

Speaker Julia Compton, senior evaluation manager of the Evaluation Department, DFID, described a recent evaluation of the UK national HIV/AIDS strategy, "Taking Action," a comprehensive and far-reaching $3 billion, 5-year effort launched in 2004, which included a substantial overseas investment component. This national strategy cuts across the UK government and involves six priority areas. The following four objectives were defined for the evaluation:

- Developing recommendations for improving implementation
- Developing recommendations for how to measure success: indicators
- Developing recommendations for a future UK strategy on HIV and AIDS
- Developing recommendations for other UK government strategies

Through an extensive consultative process, DFID identified 13 evaluation questions focusing on inputs and processes specific to decisions, for example, the usefulness of spending targets and the effectiveness of country-led approaches.

The evaluation used several methodologies. Seven case studies of countries were conducted and three working papers were developed to gain an understanding of spending, M&E frameworks, and challenges in reaching women, young people, and vulnerable groups.

The evaluation was a heavily consultative process; in fact, the process of communications and consultations during the evaluation process may have had greater impact on changes in the strategy than the actual evaluation data, remarked Compton. The process of evaluation motivated DFID to make changes needed to achieve positive results. Compton cautioned that concentrating too narrowly on the data—at the expense of communi-

cation and understanding what policy makers want—may result in missed lessons from evaluation.

A major challenge to the DFID evaluation was the declining quantity and quality of data collected at projects in-country. Because DFID relies heavily on country-led approaches and country systems to collect data, this was a major constraint to the evaluation.

CARE Evaluation of Women's Empowerment Programs

Kent Glenzer, director of the Impact Measurement and Learning Team at CARE, described the approach and methodology of a multiyear evaluation of the impact of women's empowerment interventions. The evaluation is a $500,000 effort assessing interventions at field sites in more than 40 countries, plus 900 other projects through secondary data. This evaluation is being conducted to inform organizational change at CARE, a private, international humanitarian organization with a focus on fighting global poverty.

CARE uses a literature-based theory of social change and defines the concept of empowerment as a process of change in women's agencies, social structures, and relations of power through which women negotiate claims and rights. CARE's approach for evaluating complex systems, such as women's empowerment, involves bringing together experts—internal, external, and local—and coupling M&E with project implementation. In CARE's experience, local actors know and understand systemic changes better than external experts; therefore, CARE's role is to bring actors—most importantly women and girls—together over the long term to discuss systems changes, develop hypotheses, and build collective knowledge about change.

CARE is tracking change across 23 categories of women's empowerment. Indicators—including those developed by local men and women—are developed at multiple levels for each category and include measures of individual skills or capabilities; measures of structures such as laws, family and kin practices, institutions, and ideologies; and measures of relational dynamics, such as those between men and women and between the powerful and less powerful. Although across the sites the indicators are different, broad patterns can be compared relating to where and how change is happening.

The following attributes of a successful evaluation approach, from the perspective of CARE, were outlined:

- Evaluation is a long-term learning experience that should unite relevant actors.
- Evaluation should be flexible enough so that different dependent

variables can be specified in different contexts, but should be designed to permit comparison of variables across contexts.

- Centrally planned, mixed-method evaluation designs work best.

The Global Fund Evaluation

Stefano Bertozzi, member of the Technical Evaluation Reference Group of The Global Fund, described a 5-year evaluation plan for The Global Fund, which will focus on 8 countries in depth, plus 12 others using secondary information. The evaluation is a "dose-response design," meaning it will look for correlations between intensity of project implementation and changes in trends of the HIV/AIDS epidemic in terms of survival of infected individuals and prevention of new infections.

The plan includes evaluation of the following three major topics:

- **Organizational efficiency:** Operations, business model, and governance structure in The Global Fund, which are based on technical reviews of country-generated proposals with little country presence other than auditing firms, will be evaluated.
- **Partnership environment effectiveness:** Country and grant performance will be evaluated, including the effectiveness of mobilization of technical assistance and effectiveness of country-coordinating mechanisms.
- **Health impact:** The health impact of The Global Fund on the three diseases it covers (HIV/AIDS, TB, and malaria) will be evaluated.

MACRO International Inc., Harvard University, the World Health Organization (WHO), and Johns Hopkins University are implementing the evaluation, and data collected by MACRO through Demographic and Health Surveys-Plus (DHS+)[1] will serve as the baseline assessment. The limited budget of the evaluation will not permit the conduct of large-scale surveys.

METHODOLOGICAL CHALLENGES AND OPPORTUNITIES IN EVALUATING IMPACT

Workshop participants described methodological challenges and opportunities in evaluating the impact of PEPFAR, including those in measuring outcomes and impacts specific to HIV/AIDS, measuring broader impacts and outcomes, attributing results, and aggregating the results of impact evaluation. The discussions were wide-ranging and touched on many chal-

[1]Demographic and Health Surveys including HIV prevalence measurement are known as "DHS+."

lenges and opportunities, but were by no means an exhaustive or prioritized list of considerations or an in-depth analysis of any one of them.

Measuring HIV/AIDS-Specific Outcomes and Impacts

HIV/AIDS-specific outcomes and impacts include the measurement of HIV prevalence, incidence, infections averted, mortality rates, development of drug resistance, orphanhood prevention, behavioral change, and stigma and discrimination. Workshop participants described methodological challenges and opportunities in each of these areas.

Measuring Change in HIV Prevalence

HIV prevalence is the proportion of individuals within a population infected by HIV during a particular time. It is a function of both the death rate of those already infected and the rate at which new infections occur. Repeated surveillance of pregnant women at antenatal clinic (ANC) sentinel sites is currently the most common method for measuring changes in HIV prevalence. Workshop speaker Theresa Diaz of the U.S. Centers for Disease Control and Prevention (CDC) pointed out some of the challenges and limitations of using this approach. Comparison with nationally representative household-based surveys shows that the ANC surveillance method tends to overestimate prevalence, she said, because ANCs are predominantly urban. In addition, the ANC methodology does not take into account other factors, such as the change in use of clinics over time, increased survival, or immigration, which can lead to a change in HIV prevalence. The method is also unreliable for measuring prevalence in areas where epidemics are concentrated in high-risk groups, such as Vietnam.

Diaz noted that a number of new tools are now becoming available to analyze prevalence trends more effectively. CDC uses a suite of methods (chi-square, linear, trend, linear regression, and nonparametric methods) for analyzing prevalence trends using only the most consistent ANC sites and the most recent data. In addition, a second population-based survey of HIV testing will soon be available in some countries to allow analysis of HIV prevalence over time. The collection of data on antiretroviral (ARV) use—both from ANC sentinel surveillance surveys and from the population-based surveys—would allow better prevalence data to be collected, in addition to data on coverage. Finally, methods such as respondent-driven sampling are being standardized for collecting HIV sero-prevalence data among high-risk groups. When such methods use the same sampling methodology in the same place over time, trends can be observed.

Measuring HIV Incidence

Workshop participants described some of the challenges of various methods for measuring HIV incidence, which is the number of new HIV infections within a population at risk over a given period of time. Measuring incidence is difficult, noted Diaz, because symptoms of HIV do not appear until years after infection.

Cohort studies—longitudinal studies of HIV acquisition in a particular group of people—are considered the "gold standard" for measurement of HIV incidence, noted Diaz; however, they may not always reflect the true population incidence, particularly if interventions are taking place within the cohort. Speaker Geoff Garnett of Imperial College, UK, noted a further disadvantage: there may be substantial loss of cohort participants to follow-up. Discussant Timothy Fowler of the U.S. Bureau of the Census (BUCEN) stressed that another major gap in HIV incidence measurement is the lack of empirical data on incidence by age and sex. Laboratory assays (specifically, the branched gp41 peptide, or BED, test) can also be used to distinguish recent infections from long-term infections on the basis of relative levels of anti-HIV antibodies, noted Diaz, but they tend to overestimate the proportion of most recent infections under certain circumstances, for example, in people who have taken ARV drugs immediately before the test.

Finally, the potential for spread of infection and future infection can be measured through modeling techniques, noted speaker Garnett. Factors such as contacts within the population, duration of infectiousness, transmission probability, heterogeneity in risk, mixing patterns, and different types of contact may be included in the model. However, modeling tools are limited by their inability to measure accurately the risks within populations in order to determine timing within an epidemic. Models may not predict reliably at the threshold where spread of infection becomes epidemic, and there is greater sensitivity of a system to small changes.

Speakers described a number of improvements in methodologies for measuring incidence that may provide future opportunities. Diaz mentioned that CDC has developed an adjustment formula for the HIV incidence laboratory assay on the basis of several populations and is now validating the approach in other populations. CDC is also developing a second laboratory assay to increase the specificity of HIV testing so that those individuals who contribute to "false-recent" test results can be more easily excluded. New developments in modeling tools can help to measure incidence. With the availability of the population-based HIV sero-survey data (DHS+), stated Diaz, new models based on these surveys will be able to provide important age-, sex-, and geography-specific incidence information that the ANC surveillance data cannot provide. In fact, noted Garnett, a promising method is to calculate incidence by accounting for mortality in successive prevalence

(DHS+) surveys. He also said that national incidence can be calculated indirectly by fitting models to prevalence data and back-calculating incidence on the basis of data on the survival of HIV-infected people.

Measuring Change in Infections Averted

Speaker Rand Stoneburner, an independent consultant, defined "infections averted" as the difference between expected and actual annual incidence, as shown using modeling techniques that can be empirically validated. BUCEN has been charged with estimating the number of new HIV infections that are prevented during the first 5-year phase of PEPFAR using a population projection model that takes both population change and HIV/AIDS impact into account, said discussant Fowler. The data and assumptions behind the model are taken from the UNAIDS reference group on estimates, modeling, and projections, noted Fowler, and consider factors such as the survival of HIV-positive individuals and whether or not they receive antiretroviral therapy (ART). The BUCEN model has been used to generate baseline estimates of numbers of new cases for the years 2005–2010 and will compare those to what actually occurs, noted speaker Diaz.

But measuring a "nonevent" such as change in infections averted can present significant methodological challenges, observed Diaz. When measuring infections averted in the general population, she noted, one has to assume that people would not have died of HIV infection before dying of other causes and would not have contracted HIV at a later time, but these may not be valid assumptions. Although modeling has been used to some extent to measure the effect of interventions on averting infections, it has limitations because of the gaps in data available in developing countries, she noted. Speaker Garnett also pointed out that large data gaps exist specifically in the areas of efficacy measurement in different epidemiological contexts and in the translation of efficacy to large-scale interventions. For example, noted Diaz, CDC's methodology for predicting infections averted in newborns by comparing mother-to-child transmission of HIV with or without preventive ARVs does not consider epidemiological context factors such as breast-feeding practices, efficacy differences among programs and countries, adherence and proper use, and impacts of counseling.

Alternative models, population surveys, and cross-country comparisons were put forward by workshop participants as possible opportunities for more effectively measuring changes in infections averted. Fowler mentioned that an update of the Spectrum model, which uses prevalence data to calculate past incidence, considers epidemiological contextual factors such as breast-feeding regimes and different ART regimes (UNAIDS, 2007b). Speaker David Gootnick of the U.S. Government Accountability Office cited another model developed by the Futures Group that can be

used to attribute infections averted to specific interventions, such as partner reduction and male circumcision (Stover and Bollinger, 2006). Stoneburner referred to the use of serial HIV population surveys to complement modeling approaches and provide further evidence to deduce changing incidence and infections over time. For example, a population survey done in Uganda in 1987–1988 and again in 2004–2005 was used to show reduced HIV prevalence and incidence in the younger generations over time (Stoneburner et al., 1996). Cross-country comparative analyses of HIV dynamics, risk behaviors, and intervention uptake, noted Stoneburner, can also provide insight regarding the relative effectiveness of interventions.

Measuring Changes in Survival and Mortality Rates

Workshop participants identified a number of challenges in measuring changes in survival—the percentage of a group who are alive for a given period of time after diagnosis or treatment—and in mortality rates—the proportion of deaths in an area compared to the population of that area per unit of time. Speaker Diaz pointed out that measuring changes in mortality rate in the general population has raised questions about whether ART decreases mortality, through increasing chances of survival of HIV-infected individuals, or increases mortality, through increased opportunities for viral transmission to others. She added that "lives saved" and mortality rates may not actually be appropriate outcome measures given that HIV/AIDS treatments may not actually save life, but only delay death.

Other challenges of measuring change of mortality relate to gaps in available data. Many patients are lost to follow-up, and frequently population- and hospital-based mortality data exist but are overlooked, noted speakers Diaz and Stoneburner. Mortality and health surveillance systems in general lack support, infrastructure, and validation, stressed Stoneburner. In many countries, he noted, capturing all deaths and ensuring the accuracy of mortality data are a problem. For example, many countries in sub-Saharan Africa are registering only a fraction of all deaths; nevertheless, appropriate analysis of such data may still be useful in measuring mortality changes over time. Accurately ascertaining the cause of death in the general population is a further problem, noted Diaz. In some countries, however, such as Botswana, both the accuracy and capture of mortality data are thought to be high—90 percent to 95 percent. The variability in data capture and accuracy from country to country may contribute to a disconnect between observed mortality and modeled mortality and may indicate a need to adjust models to better replicate empirical data, observed Stoneburner. He shared an example from Botswana in which a model predicted more deaths than the registration data reported, likely a result of overestimating HIV incidence due to too short an assumed survival in the

model. Speaker Jonathan Mwiindi of the Kijabe HIV/AIDS Relief Program in Kenya asked a question about a case in which increased mortality rates of patients coinfected with HIV and TB during early ARV treatment were perceived to relate to ARV treatment failure rather than the inadequate recognition and treatment of TB. Stoneburner responded that better use of existing TB and ARV surveillance data, including ARV cohort survival analyses, rather than reliance on population-level vital registration data would better identify risk factors for such adverse outcomes and better guide clinical management.

Workshop participants identified a number of opportunities for more effective measurement of mortality and survival rates. Diaz commented that "years of life added" might be a more appropriate measure than "lives saved" among the HIV-infected population because life may only be prolonged for a limited number of years.

Several speakers emphasized the importance of improving the quality of mortality data among both the infected and general populations. Diaz urged more aggressive pursuit of information on patients lost to follow-up and standardized methods for collecting information from hospital records. Several innovative methods for improving mortality data were suggested. Discussant Fowler mentioned that the health metrics network at WHO has done research on verbal autopsies for following up on deaths in households and determining cause of death. One verbal autopsy scheme under consideration is being developed by BUCEN—called sample vital registration with verbal autopsy (SAVVY)—and involves a census of a sample of households from different areas of the country, with follow-up over a period of time when there have been deaths. Census enumerators return to the households to collect information, which is reviewed independently by medical doctors to obtain a cause of death.

Corporate-sector surveillance systems may also provide important data on early impact on mortality of treatment programs, suggested Stoneburner. He shared results from a private-sector mortality attrition study conducted in Malawi showing early impact of ART programs, which was later expanded to a larger corporate-sector study gathering data from seven businesses (see Figure 4-2). Stoneburner acknowledged that although the private sector is very selective—more likely to provide access to treatment at the workplace and to target employees who are well educated and highly motivated to stay on treatment—reduced mortality in the private sector may be an important early indicator of expected response in the general population, once enough people have been treated to see a change in mortality.

Speakers also emphasized the importance of gathering age-specific and population-based mortality data for measurement of mortality impact. Diaz noted that impact on mortality may be useful when measured by concentrating on the young adult population and excluding the most

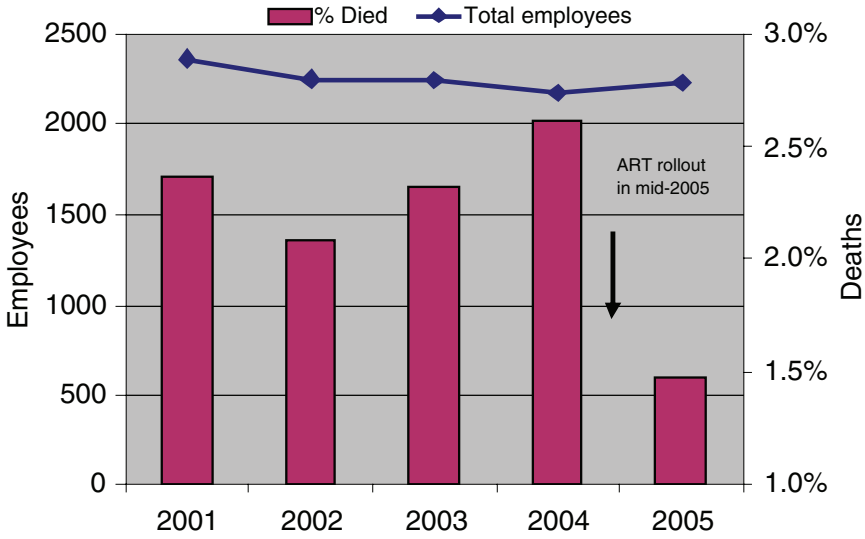

FIGURE 4-2 Private-sector attrition data show evidence of early ART impact on mortality.
SOURCE: Adapted from *Partners In Impact: Results Report*, The Global Fund to Fight AIDS, Tuberculosis and Malaria, March 2007. Based on data provided by the National AIDS Commission Trust of the Republic of Malawi, Principal Recipient for the Global Fund programs.

obvious, non-HIV-related causes of death, such as accidents and maternal mortality. Stoneburner reinforced the importance of gathering age-specific, population-based mortality data, describing a Botswana study that assessed impacts of ART programs on adults mortality and PMTCT programs on infant and child mortality (WHO, 2006). Although effects of ART programs clearly correlated with declining mortality in the 25–54 age category, in the 0–4 age range, no mortality decline was observed despite high use of azidothymidine (AZT) by mothers and infants. The unexpected lack of mortality decline in children could relate to inaccuracies in mortality capture, but more importantly may relate to factors impeding overall intervention effectiveness, such as the added risk of increased infant mortality related to infant feeding practices. In summary, use of available data can identify important changes in the dynamics of impacts of interventions that would not otherwise be detected through routine M&E tools.

Measuring Behavioral Change and the Impact of Behavior-Change Interventions

Workshop participants discussed challenges and opportunities in measuring behavioral change and the impact of interventions that modify risk behaviors. These interventions, noted speaker Latkin, include those to modify sexual behaviors, injection behaviors, and drug-adherence behaviors. Gaps in data and surveillance are a major challenge in measuring behavioral change. Discussant Caroline Ryan, Office of the U.S. Global AIDS Coordinator (OGAC), pointed out the need for more qualitative data and more behavioral surveillance, such as behavioral sentinel surveillance with biomarkers, to be conducted on a more consistent basis in order to understand the drivers of infection. Tools currently available include the DHS+ studies and assorted simulation models, she said. Although these are providing information on who is affected and what the behaviors are, information on *why* specific populations are affected is also needed. Prevention efforts need to have more heterogeneity to be effective because of the substantial variation within populations, commented Ryan.

Evaluation methods are also currently limited in their ability to determine coverage and extent of behavioral change occurring, observed speaker Latkin. While some interventions, such as media communication, result in small changes but large coverage, others, such as behavioral-change counseling, result in large behavioral changes with narrower coverage. More information is needed to determine how much behavioral change and how much coverage are needed to change the course of the epidemic. For some interventions, such as adherence to ARV treatment, incomplete behavioral change has consequences that are even worse than no behavioral change, noted Latkin, because moderate adherence could provide a greater selective pressure for the evolution of viral drug resistance compared to poor adherence. Methods for determining such unintended impacts need to be developed.

A further challenge of measuring behavioral change and effects of behavior-change modification is the potential influence of factors independent of the program intervention. As Latkin pointed out, cultural and structural factors may affect a program, leading to a success of an intervention in one context and a failure in another. For example, he noted, an identical intravenous drug user intervention used successfully in the United States failed in Thailand because of a specific law on drug use in Thailand. Speaker Stoneburner further reinforced the idea that although changes in HIV prevalence may result from behavior-modification interventions, they may also result from other factors that have nothing to do with the intervention. Frequently, such change comes about not because of a specific targeted intervention from an outside agency, but because of a comple-

mentary, indigenous community response, or even from natural epidemic dynamics. For example, in Uganda, declining HIV prevalence correlates with declines in multiple sexual partners, a population-level indicator of behavioral change (Stoneburner and Low-Beer, 2004; MOH and MACRO, 2006) that generally occurred before major externally funded HIV interventions. In contrast, in Botswana, extraordinarily high and stable HIV prevalence is associated with high levels of condom use (80 percent) for the past decade, no evidence of declines in multiple sexual partners, and a plethora of resources and externally funded interventions. However, posited Stoneburner, incipient declines in HIV noted since 2004 may have more to do with mortality breaking up sexual networks than interventions.

In a further example from the Malawi context, in which a similar pattern of declines in HIV prevalence and declines in multiple sexual partners among males was tracked between 1996 and 2004, the association between declines in HIV prevalence and evidence of behavior change is less clear. Despite data suggesting substantial declines in HIV prevalence among youth in urban and semi-urban areas, there is no similar trend in the few rural sites where data are available. Trends in behavioral indicators since 1996 suggest substantial declines in multiple partners overall, but when stratified by residence, the major decline from 1996 occurred among rural rather than urban males. Declines in prevalence among urban youth may be related solely to natural epidemic dynamics, noted Stoneburner, but a more plausible hypothesis is that prevalence declines relate to behavior change preceding the behavior changes observed in rural areas but not captured through the survey.

Workshop participants noted several opportunities for evaluating behavioral change. Latkin urged a shift in measuring change from what is happening at the individual level to what is happening at the social and institutional levels (such as, community, network, or national levels). Change at these levels is necessary to build infrastructure for and sustain risk reduction at the individual level, he noted. Instead of using the individual as the unit of analysis, we should be evaluating a random sample of programs. Latkin also provided guidance on what should and should not be measured to track behavioral change. Knowledge of and contact with programs are indicators that tend not to be associated with behavioral change. Opinions of leaders, impediments to behavioral change, and unintended negative consequences of behavioral change are among the indicators that should be measured. The evaluation of behavioral change should be integrated at multiple levels within PEPFAR as opposed to focused within a program, and it should systematically engage both the scientific community and stakeholders, recommended Latkin.

Speaker Garnett highlighted the availability of other tools for evaluat-

ing behavioral change. Behavior surveys are a useful tool for attributing changes in HIV incidence to specific changes in risk. Such surveys should specifically target younger people 1 or 2 years after sexual debut, he suggested. Sampling this demographic can be used to calculate cumulative incidence of HIV and can serve as an early indicator of the success of HIV infection prevention interventions. HIV prevalence changes in response to behavioral change interventions are more marked among young people (ages 15–19) as compared to older people (ages 40–44).

Modeling is another instrument for linking behavior-change interventions to prevalence and incidence outcomes, noted Garnett. Modeling can simulate the degree of deviation between what can be expected from the natural course of the epidemic and what can be achieved through various interventions. Models can be used to predict what is known as the *counterfactual*, or the outcome that would have occurred had the donor or intervention been absent. Models that combine trends in prevalence and incidence with studies of risk behavior can be a useful tool for retrospectively understanding how interventions might have worked to maximize declines in HIV prevalence. Simulation models show maximal declines in prevalence as high-risk behaviors decrease, leading to reduced incidence and fewer replacements of those HIV-infected people who die (such models must simultaneously take into account the opposite effect of ART treatment in decreasing the rate of decline in prevalence as the death rate of infected persons decreases). If interventions are successful in changing behaviors, incidence is lower, and declines in prevalence are maximized because people dying are not instantly replaced by people with similarly high-risk behavior.

Modeling of the Ugandan HIV/AIDS epidemic has allowed researchers to simulate the effects of various behavioral changes—increased condom use, delayed sexual debut, and decrease in partner change—on prevalence. The observed prevalence data best fit a scenario in which all three behaviors changed at once (Hallett et al., 2007). Similarly, in Zimbabwe, modeling the declines in prevalence also indicates that risk behaviors are changing and leading to decreased incidence. These prevalence results in Zimbabwe have been corroborated through randomized controlled trials conducted at two time points. Surveys conducted in conjunction with the trials demonstrate that declines may have resulted from behavioral changes such as foregoing casual sexual partners and reducing simultaneous partners (Gregson et al., 2006). The success of behavior-change interventions is highly contextual, however. The failure of prevalence levels to decline in Côte d'Ivoire suggests no impact of interventions on risk behaviors.

Measuring Stigma and Discrimination

Stigma and discrimination—negative attitudes, beliefs, and actions toward people who are perceived to have HIV/AIDS and those associated with them—are an important part of the impact evaluation picture, workshop participants noted, but methodologies for studying them are limited. Speaker William Holzemer from the University of California–San Francisco urged the development of more rigorous research and data collection approaches. Most of the literature on stigma and discrimination, he noted, is based on anecdotal evidence, testimonials, and a few qualitative studies. Perceptions of a reduction in stigma and discrimination are based, for example, on observations of increased numbers of patients seeking testing and long lines of people waiting to access ART (Holzemer and Uys, 2004). In a recent review of the literature, not one stigma-reduction intervention trial had any rigorous measure of stigma that could be used to draw a conclusion about a particular intervention. Holzemer emphasized the importance of developing scales to measure stigma; the effects of stigma on infected individuals, families, and health care providers; and the effectiveness of strategies for mitigating stigma. Citing reports he had seen suggesting that women from Mozambique who use antenatal services are automatically assumed to be HIV-positive (IRIN, 2007), discussant Fowler also noted the importance of developing measures that would track the extent to which stigma is a factor in patients who seek other services, such as antenatal care.

New methods for measuring stigma and new sources of data may provide opportunities that will be useful to the future impact evaluation of PEPFAR, workshop participants said. Speaker De Lay mentioned that a new tool for measuring stigma is now available from the International Planned Parenthood Federation (IPPF et al., 2008). This index includes a measure of self-imposed stigma, which can capture the failure of persons living with HIV/AIDS (PLHAs) to access services because of fear of rejection or perception that their future is too limited to justify attempting to access services (such as education) in the long term. Holzemer referred to other new measures now available that can be valid and reliable instruments for measuring stigma (Holzemer et al., 2007). These instruments—reflecting measures of people's perceptions—include 33 factors and are based on the reported frequency of occurrence of verbal abuse, negative self-perception, health care neglect, social isolation, fear of contagion, and workplace stigma based on HIV status.

Although few empirical studies exist, a few correlational studies and new sources of data on stigma are emerging, noted Holzemer. Focus group data collected from five African countries have assessed stigmatization of patients by health care workers among more than 1,500 nurses and 1,500

PLHAs. Studies suggest that among the HIV-infected, stigma has impact on participation in testing, the use of services (such as giving birth at home instead of returning to a clinical facility), adherence to medication, health status, and quality of life (such as loss of social support, isolation, violation, verbal violence, and limiting social interactions). The data indicate clearly that HIV patients are treated poorly by health care providers—nurses, physicians, and others. The new studies have shown that stigma also has impact on the quality of work life and quality of life for health care workers and their families. Health workers and their families may be stigmatized by their neighbors because fear of contagion is an underlying cause of stigma. New data are showing that testing, diagnosis, having the disease, physical manifestations of AIDS, status disclosure, suspicion, and rumors are all triggers to the cascade of stigma events.

Measuring Changes in Orphanhood Prevention

Workshop participants discussed some of the challenges and opportunities in measuring changes in orphanhood prevention—the prevention of the death of one or usually both parents of a child. Speaker Diaz noted that measurements need to distinguish between children who have lost one parent (single orphans) and children who have lost both parents (double orphans) to HIV. In addition, HIV status should be taken into account in the calculation of years of orphanhood averted. Treatment—both of HIV-positive orphans and of HIV-positive parents—can have an impact on orphanhood. Diaz pointed out that ART treatment of orphans actually extends years of orphanhood. Discussant Mead Over of the Center for Global Development observed that although treating HIV-positive parents can reduce orphanhood years of existing children by prolonging parents' lives, it can also generate years of orphanhood among children who are born to HIV-positive parents during treatment. He also mentioned a limitation in the ability to conduct cost-effectiveness analysis of orphanhood-prevention interventions. No method yet exists for expressing "orphanhood years averted" in terms of healthy life years, the usual common denominator for a benefit in cost-effectiveness analysis. Over called for the need to establish a crosswalk between orphanhood years averted and the dollar value of a healthy life year in order to better integrate evaluation of averted orphanhood into cost-effectiveness analysis.

New models are in development to better quantify the impact of treatment and prevention in preventing orphaning of children, noted speaker Pacqué-Margolis. Diaz also offered guidance on potentially useful indicators, including "years of orphanhood averted" and "number of children who reach age 18 before the death of a parent whose life is extended by

ART." Differences in overall numbers of orphans within a given time period with and without ART can also be examined, said Diaz.

Measuring Change in the Development of Drug Resistance

Workshop participants noted the importance of evaluating the development of viral resistance to drugs—the evolved capability of HIV to withstand a drug to which it was previously sensitive. Speaker Diaz stated that two strategies—both used by PEPFAR and WHO—are available to measure drug resistance: threshold surveys and therapy monitoring. The threshold survey can be used to assess transmitted HIV infection using blood tested at ANC sentinel surveillance sites. Blood sampled from young women (younger than age 25) in their first pregnancies who are not likely to be in ARV treatment can be used to track the transmission of drug-resistant HIV strains. A second strategy for measuring drug resistance is to sample and monitor patients in ARV treatment from the initiation of therapy over a 1-year period. Indicators of drug resistance such as outcome, viral load, and drug adherence can be monitored using this method.

Measuring Broader Impacts

Most participants agreed that in addition to measuring HIV/AIDS impacts of PEPFAR interventions directly, a broader interpretation of impact is also meaningful. Large-scale vertical programs such as PEPFAR can have far-reaching effects—either intended or unintended—beyond HIV/AIDS. In addition, as speaker Compton observed, PEPFAR and other donors are increasingly investing in less narrowly defined interventions that are not so amenable to a conventional evaluation framework of inputs, outputs, outcomes, and impacts. The branching out by donors to broader areas such as gender and nutrition has made impact evaluation increasingly complex. Workshop participants discussed some of the challenges and opportunities in adapting a traditional evaluation framework to measure broad impacts or unintended impacts of PEPFAR interventions. This section describes the measurement of the impact of the following: health systems strengthening, complementary interventions, gender-focused activities, coordination and harmonization, and population-level service delivery.

Measuring Impacts of Health Systems Strengthening

Workshop discussants brainstormed together on how changes in health systems can be measured. These include a broad range of factors related to health care service delivery, such as accessibility, quality, efficiency, and equity of services; management; procurement and distribution systems;

human resource use; policy environment; and infrastructure. Speaker Compton suggested that possible indicators to help track whether reliable and sustainable partner institutions are in place—similar to a system-audit approach that many auditing organizations use to determine whether systems have been established that would enable donors to give money directly to institutions with confidence—could include the following capabilities: to collect information and use it to make good decisions, to plan and budget efficiently, to implement projects effectively and efficiently, and to monitor and evaluate and to collect reliable numbers needed by members of Congress, Parliament, and others. Two case studies were also presented describing indicators and methodologies that can be used in evaluating health systems strengthening.

Evaluating health system–wide impacts of Global Fund interventions. Workshop speaker John Novak, senior monitoring and evaluation adviser of the Office of HIV/AIDS at the U.S. Agency for International Development, presented the experience of evaluating health system–wide impacts of Global Fund interventions carried out in Ethiopia, Malawi, and Benin. The evaluation effort was carried out by the System-Wide Effects of Global Fund (SWEF) network, a collaboration of research institutions seeking to understand how global health initiatives affect the broader health system. A core assumption of the evaluation framework is that programs that massively infuse resources into country health systems can improve or detract from health system accessibility, quality, efficiency, and equity. In Benin, The Global Fund provided 15 percent to 20 percent of the government spending per capita; in Ethiopia and Malawi, it provided 50 percent. Such effects can be intended or unintended. Therefore, any evaluation should go beyond vertical programs and focal diseases to assess effects on the entire health system.

The SWEF evaluation assessed impact of Global Fund interventions on the following four parameters:

- Policy environment (harmonization, alignment, and ownership)
- Human resource use (number, allocation, skills, retention, and motivation of health workers)
- Public–private services and collaborations (number, distribution, and organization of actors; trust and cooperation between sectors)
- Pharmaceutical and commodity procurement and distribution systems

Both quantitative and qualitative methodologies were used in the evaluation. Quantitative facility surveys were conducted in a sampling of health facilities to assess staffing, management, patient referrals, drugs and sup-

plies, lab services, and curative care services. Quantitative provider surveys were used to measure impact on individual providers and facilities receiving funds and to assess training, supervision, motivation, and job satisfaction. In-depth qualitative interviews with important stakeholders were also conducted throughout the entire health system.

Novak stressed the importance of monitoring both positive and negative impacts of interventions, which can help countries address critical issues in the health system. For example, although the SWEF evaluation results showed positive impacts on the health system—such as greater participatory engagement, decentralization, the emergence of new public–private collaborative arrangements, creation of improved incentives and work environment for those working in HIV/AIDS, and harmonization of pricing and cost-recovery approaches—there were also some negative impacts, such as delivery-level constraints as HIV/AIDS drew both human resources and services away from other health areas, and poorly functioning procurement and distribution systems in some countries.

Challenges of using this more descriptive methodological approach include the lack of empirical estimates of impacts, small sample size, short time interval over which change was evaluated, and lack of ability to easily attribute impact.

Evaluating impact of HIV/AIDS interventions on non-HIV primary health care services. Jessica Price, Rwanda country director of Family Health International (FHI), presented results from a study conducted in Rwanda testing the hypothesis that HIV/AIDS interventions strengthened the number of non-HIV primary health care services. Study data were derived from the review of monthly activity reports submitted by health centers to the government of Rwanda. The study compared the quantity of non-HIV health services delivered before and after the introduction of basic HIV care, defined as services including counseling and testing, PMTCT, preventive therapy, and basic upgrades to health center infrastructure. The study assessed 30 FHI partner health centers from 4 provinces and 14 districts in Rwanda, representing 21 faith-based centers and 9 public centers. Hospitals that do not deliver some non-HIV services and health facilities with fewer than 6 months' experience delivering basic HIV care were excluded from the study.

A set of 88 indicators of non-HIV services delivery was tracked, with 22 indicators considered to represent the best range of public health services. These included general services (such as inpatient and outpatient services and lab tests), reproductive health services, and services for children. In addition to monitoring impacts of HIV/AIDS interventions, the study also tracked impacts of two other health programs—primary health care insurance and performance-based financing—and used regression analysis

to isolate the independent effects of HIV/AIDS interventions. The analysis consisted of calculating mean quantities of non-HIV services delivered per primary health center per month between the two time periods, testing for significant differences, and conducting regression analysis to control for experience with other health programs (insurance and performance-based financing) to determine which program, if any, had an independent effect on the observed change.

The HIV programs were shown to have had an independent effect in a number of indicators across a range of areas. These areas included improved coverage for antenatal visits and services, use of health care facilities for maternity services by HIV-positive women, syphilis screening, family planning services, child vaccination and growth-monitoring services, outpatient consultations, and hospitalization services.

Limitations and challenges of the methodology were discussed. In future analyses, evaluation of the impacts of HIV programs should also include hospital settings. Indicators could also be tracked for impacts on other diseases (such as, malaria, TB, and sexually transmitted infections), quality of patient care, costs of HIV-specific services (such as HIV tests) versus non-HIV-specific services (such as infrastructural upgrades like incinerator construction and maintenance of electricity), and client and provider satisfaction. Future studies should also look at larger sample sizes over longer time periods. A random selection of sites should also be considered in future studies, noted speaker Field-Nguer. The fact that all chosen sites were FHI partners may have given them a competitive edge, she noted. If being FHI sites did not confer an edge, then perhaps access to services can be replicated at any site in Rwanda. But if FHI status did confer an edge, then perhaps unique attributes of the partnership can tell us something about how to replicate the impact, she noted. Workshop participant Laura Porter of CDC added that future studies will need to ensure that service delivery improvement is a real effect and not just an artifact of data system improvement.

Measuring Impact of Complementary Interventions

As described in Chapter 2, PEPFAR investments include numerous interventions in programs complementary to more narrowly focused HIV services. These so-called wraparound programs include interventions in areas such as malaria, TB, nutrition education, food security, social security, education, child survival, family planning, reproductive health, medical training, health systems, and potable water.

Workshop speaker Bertozzi described methodologies from two case studies from Mexico in which such complementary interventions were

evaluated: a human-capacity development program for children and a food assistance program.

The Oportunidades program is a Mexican government–sponsored human-capacity development program for Mexico's poorest children. Financial incentives to parents are offered through the program for ensuring children's participation in health, nutrition, and educational services. The Programa del Apoyo Alimentario (PAL) program provided food assistance—either food or cash payments—to small rural communities in Mexico. Impact evaluations of both Oportunidades and PAL were conducted using prospective randomized evaluation, in which later program enrollees were compared to earlier program enrollees. Both health impacts and education impacts were monitored through the evaluations. For Oportunidades, health indicators tracked include use of preventive services (such as well visits and vaccinations), use of curative services, out-of-pocket expenditures, and anemia prevalence. PAL health impacts monitored included height-for-age, weight-for-height, and weight-for-age. Education indicators monitored in the Oportunidades program included grade-level achievement, attendance, early enrollment, and repetition of grades.

The evaluative approach from these studies could potentially be applied to the evaluation of complementary interventions in the PEPFAR program, particularly to health and educational interventions targeting orphans and vulnerable children, noted Bertozzi. Other indicators of "basic capability" child care interventions could include zinc status, sick days, days incapacitated, prevalence of risky and healthy behaviors (such as alcohol use, sexual activity, and exercise), and educational performance.

Bertozzi emphasized the importance of controlling for secular—long-term, noncyclical—trends in impact evaluation. Such trends can sometimes have a large effect independent of the intervention. For example, malnutrition indicators were tracked in the poorest rural communities in Mexico in the 5 years leading up to the start of the PAL program (ENN-1999 versus PAL-2004, the baseline for the PAL intervention). In the absence of any intervention, noted Bertozzi, extraordinary secular trends led to a halving of malnutrition indicators in these communities. Any intervention conducted during this 5-year period would have given the appearance of stimulating a large positive effect when there might have been none at all—or perhaps even a negative effect.

Measuring Impacts of Gender-Focused Activities

Workshop participants discussed some of the challenges and opportunities for evaluating the impacts of gender-focused activities, including those interventions to promote gender equality and women's empowerment. Noting that gender equality and women's empowerment are multidimen-

sional, open, complex, nonlinear, and adaptive systems, speaker Glenzer observed that it is seldom clear what variables are or are not involved. It is a challenge to define what constitutes success and what it looks like on the ground. Glenzer said some of the difficulty in tracking change of gender systems relates to the following characteristics: the large-scale effects of small changes over time, the separation of causes and effects over large spatial and temporal scales, the multiple levels over which change may occur, and the heterogeneity of systems. Speaker Julie Pulerwitz of the Population Council acknowledged the difficulty in implementing rigorously designed evaluations and called for more consensus building about how to operationalize the concept of gender and how to evaluate gender-related activities. Although gender is generally recognized as important, she added, there have been few outcome evaluations and few tools developed on how gender-focused activities affect HIV risk. Few good indicators exist that are useful in understanding social dynamics, and evaluation schemes often underrepresent the perspectives of local people, who are a source of such knowledge, noted Glenzer.

Speaker Pulerwitz described a new method now available for studying the impacts of gender-focused activities and how those impacts can contribute to PEPFAR goals. Pulerwitz directs an operations research program called Horizons at the Population Council that has conducted studies using this method. Pulerwitz shared the study design and tools used for an evaluation of gender-focused programs—group education, community-based behavioral change communication campaigns, and clinical activities—focused on young men in Brazil. A combination of data collection approaches were used, including the following:

- Pre- and postintervention surveys and a 6-month follow-up survey for three groups of young men—two intervention groups and a comparison group, which eventually also received the interventions after a time delay—followed over a year
- In-depth interviews with a subsample of young men and their sexual partners
- Costing analysis and monitoring forms for different activities

An evaluation tool called the Gender Equitable Men's (GEM) scale was used to look at gender norm attitudes and how they changed over time (Barker, 2000; Pulerwitz and Barker, 2008). The scale includes 24 items, including parameters such as home and child care, sexual relationships, health and disease prevention, violence, homophobia, and relations with other men. Certain GEM scale domains are associated with partner violence, level of education, and contraception use. The GEM tool was used to detect significant changes in attitude toward equitable gender norms and

in support of inequitable gender norms in the two intervention groups as compared to the control group. HIV outcomes—condom use with primary partners—were also tested, and one of the intervention groups showed an increase as compared to the comparison group. The study also looked at covariance between changes in attitudes toward norms and changes in condom use; men who were more gender equitable were more likely to report condom use. The in-depth interview component of the analysis unearthed other changes among those in the test groups, including a delay in sexual activity in new relationships.

The evidence generated by the evaluation is supportive of interventions that target gender dynamics and their influence on HIV risk behavior in Brazil, concluded Pulerwitz. She noted that there are ongoing or planned efforts to adapt the GEM tool to other country contexts—India, Ethiopia, Namibia, Uganda, and Tanzania—and to other demographic groups, such as married men. Preliminary findings show that results can be highly country specific. Although a similar trend toward more equitable attitudes has been observed in the work conducted in India, baseline attitudes in that country are much less supportive of equitable gender norms than those in Brazil.

Measuring Coordination and Harmonization

Workshop speaker De Lay spoke of a new opportunity for measuring coordination and harmonization—the alignment of interventions with country-level plans and coordination of efforts among other implementing partners. A new tool, known as the Country Harmonization and Alignment Tool (CHAT), developed by UNAIDS and the World Bank, is now available and could be applied to the standardization of alignment of interventions with country-level plans and coordination of efforts among partners (UNAIDS, 2007a).

The tool has been used to assess harmonization and alignment of the national plan, coordinating mechanism, and M&E plan in six pilot countries, and a launch of the tool is planned in two more countries. The tool has revealed that many national plans are still not credible, not costed appropriately, not prioritized, and not actionable. In addition, the tool has shown that few countries have a central funding channel or single procurement system for the HIV/AIDS response. The tool has also shown that "basket funding," or joint funding by multiple donors, is not normally used. Although donors support the notion of the development of indigenous national M&E capacity, the tool has revealed that in practice donors usually rely on their own M&E systems to collect urgent data when needed.

Measuring Community-Level or Population-Level Service Delivery

Workshop speakers spoke of the challenges of scaling up successful service-delivery interventions for specific populations, such as children, families, communities, and high-risk groups. As workshop speaker Bertozzi observed, sometimes it is difficult to distinguish between a community-level or population-level effect and the effect of an intervention. Tools are needed, noted speakers Kathy Marconi of OGAC and Stoneburner, to measure the effectiveness of interventions in specific populations, including communities, diverse populations, and at-risk or infected populations.

Speaker Field-Nguer announced that a new and important addition to the evaluation toolbox is now available: community-level program information reporting systems (CLPIR) (personal communication, R. Yokoyama, John Snow, Inc., January 18, 2008). CLPIR indicators look strictly at community-level service delivery and help answer questions such as when, how, and where people want testing and treatment.

Attributing Impact

Given the diversity of programs and funders, attributing impact—or relating a particular effect to the work of a specified agent—is a substantial methodological challenge in evaluation, workshop participants said. The World Bank experience shows that because loans or grants are made to governments, speaker Ainsworth said, performance of activities depends heavily on governments, and it is therefore difficult to disentangle the efforts of government and any particular donor from the efforts of all other donors. Even within the programs of a single donor, noted speaker Gootnick, accounting can be complex. Some interventions can be double counted; for example, voluntary counseling and testing is included under both the prevention and care modalities. As PEPFAR moves increasingly toward more harmonized approaches, noted speaker Compton, it will be even more difficult to disentangle effects in an exclusive way.

Many workshop participants agreed that the demand for exclusive attribution by donors may not be constructive. General evaluation of what is and is not working, in contrast, may be desirable, noted workshop moderator Ruth Levine of the Center for Global Development. Speaker Glennerster emphasized that it is preferable to test what works in very specific areas and then judge a program by whether it spends money on interventions whose effectiveness is supported by evidence. All programs are doing many things in-country; they are implementing many different policies. If we want to be effective in focusing resources on what works, we need to identify which interventions have the most impact and which are most cost-effective, she said. Speaker Diaz reinforced this idea, stating that a worthwhile attribu-

tion goal should be to know the effectiveness of certain programs and their coverage in terms of impact measures. A useful attribution exercise, she suggested, might be to determine what level of ART coverage decreases general mortality and what types of prevention activities, in which populations, decrease HIV incidence. Ainsworth added that it is nevertheless useful to analyze the value added of the unique approaches of particular donors.

An important dimension of attribution is the concept of the counterfactual, or the assessment of what would have happened differently had the donor not intervened. Some speakers noted that absence of the donor does not necessarily imply that nothing would have happened. Discussant Jim Sherry of George Washington University observed that one consequence of donor interventions is that the donor occupies a particular space and prevents other organizations from filling it. As speaker Bertozzi pointed out, in the case of South Africa, even if outside institutions did not intervene, given the massive social mobilization potential in the country, dramatic change could have been effected without outside help.

Aggregating Evaluation Results

Several speakers noted that the synthesis or aggregation of evaluation results is a methodological frontier. Workshop participant David Dornisch of the U.S. Government Accountability Office proposed that meta-analysis or synthesis could be used to bring together the results of multiple studies. From the congressional perspective, workshop participant Naomi Seiler from the U.S. House of Representatives Oversight Committee also stated that while prospective evaluation is useful, any type of meta-analysis or synthesis of what is already known about types of interventions, contexts, and populations would be helpful. Discussant Jimmy Kolker of OGAC echoed the need for data synthesis to be relevant to designing or implementing a program.

Workshop discussant Sherry observed that such methods have yet to be developed, however. Sherry predicted that the clustering of country-level assessments and evaluations will likely provide much more information through meta-analysis than one definitive, globally executed impact study. Although there is room for both kinds of evaluations, he noted, there is substantial room for improvement on meta-analysis to look statistically at the results of these studies. Sherry observed that there may be inadequate separation of macro-, micro-, and meta-level evaluation processes, leading to an evaluation either not making sense to policy makers or not being rigorous enough for scientists. Micro-level evaluation tends to be too technical and too situation-specific to be digestible to institutions or useful for interventions. Macro-level evaluation tends to be too soft and too subject to evaluation spin to be digestible or credible. Durable findings are needed

about programs that allow for more sustainable dialogue and learning at the meta-level in terms of evaluation.

Another workshop participant raised a question about the value of performing multiple evaluations. Speaker De Lay commented that although it is sometimes desirable to avoid duplication where it is not needed, sometimes duplication is necessary and multiple perspectives are desirable. For example, validation of existing data by an independent group is often a useful alternative to redoing an entire study.

THEMES COMMON TO EVALUATION METHODOLOGIES AND APPROACHES

This section distills some of the main messages and themes common to the discussions about evaluation methodologies and approaches.

Prioritization

Most evaluations require some type of prioritization to narrow down what is to be measured. Speaker Ainsworth noted that for long-term evaluations, for example, one might select only those issues common to all projects. For a large portfolio of activities, she added, one might select a more narrowly defined set of indicators.

Value of Consultation and Communication

Several speakers emphasized the value of consultation and communication in any evaluation approach. Speakers Compton and Glenzer observed that consultation and communication through the evaluation process are as important in effecting change and course corrections as the data from the evaluation results. It also matters who is consulted, observed speaker Field-Nguer.

Value of a "Learning" Evaluation

Many of the evaluation methodologies described were formative, or "learning" evaluations, designed to help improve institutional performance. As Glenzer noted, evaluation is a long-term learning experience that should unite relevant actors. Speaker Ainsworth added that bringing to bear the findings of past support can inform ongoing programs. Using evaluation to understand the variation in outcomes, or the distribution of outcomes within a population, can help us learn, she said. For example, changes in the average life expectancy or the average change in behavior is not as

interesting as knowing *why* behavior changed in one group of people but not another.

Others emphasized the heuristic value of negative evaluation results. Analysis of failures, observed speaker Field-Nguer, is sometimes more fruitful than success stories. Negative evaluation results should be divulged and shared, one workshop participant urged; if they are not shared, programs lose credibility and waste money. Speaker Glenzer noted that all of CARE's research reports are published on Emory University's website and include some research indicating that CARE is not having long-term impacts on women's empowerment or underlying causes of gender inequality.

The emphasis on learning evaluations contrasts with a more typical systemic bias in the international health community in which actors want to see programs continue, noted workshop discussant Sherry. Therefore, instead of using evaluation for learning, it is used to protect our interests and programs. Sherry underscored the importance of sustaining the institutional learning process. The isolation of evaluation departments in international health systems—analogous to the isolation of smart and reflective people in universities, organized into separate compartments so they have minimal effect on the society around them—is one obstacle to institutional learning, he noted. Decision-making cycles, such as 5-year cycles, reauthorizations, or external audits, drive evaluators into prominence briefly but then fade away. Also observing the existence of different consumers of evaluation, speaker Nils Daulaire of the Global Health Council emphasized the importance of having a single M&E system that satisfies multiple sets of needs. For example, if a customer for evaluation is Congress, then the evaluation will emphasize putting on the best possible spin, but that must be balanced with the use of evaluation on a daily basis to help improve program development and results. One step in achieving a multiuse system is to give evaluators a role in program management and development as opposed to a peripheral role in projects.

Importance of Designing the Evaluation Early

Several speakers emphasized the importance of considering evaluation design early in the implementation process so that the design will be appropriate and so that impacts can be detected early. Speaker Compton urged that evaluations be set up at the beginning of the process, and speaker Bertozzi also spoke about some of the drawbacks of an ex-post evaluation. Speaker Glennerster noted that opportunities to use powerful randomization approaches exist, but they can be used only if the design is included at the beginning of an intervention. Field-Nguer and Bertozzi stressed the importance of baseline assessments, without which the wrong conclusions may sometimes be drawn.

Understanding the Limitations of Models and Data

Workshop participants acknowledged the limitations of data and models used in evaluation. Speaker Pacqué-Margolis emphasized that empirical data are often inadequate, lacking, or inaccurate, and speakers Ainsworth and Compton emphasized that poor data quality at the country level is often a serious problem. Speaker Garnett emphasized the existence of data gaps for measuring efficacy in different epidemiological contexts. Age- and sex-specific empirical data are also lacking, noted discussant Fowler. Ainsworth stressed that incentives need to be created to encourage project staff and governments to establish and maintain monitoring efforts. Not all data are of the same quality, participants said. Speaker Glennerster noted that data based on self-reported behavior might have issues regarding reliability.

Models are powerful tools that can help in evaluation, but they also have limitations. Speaker Glennerster pointed out that models need to be validated with empirical data, and variables need to be added to them to make them more accurate predictors. Speaker Garnett also observed that models are less reliable predictors when the spread of HIV infection becomes epidemic.

Value of Multiple Methodologies

Several presenters noted the value of using multiple methodological approaches in evaluation. Speakers Compton and Ainsworth cautioned against relying exclusively on one evaluation methodology, and speaker Field-Nguer pointed out that multiple methods may yield richer results than one or two methodologies. Field-Nguer also noted that lack of a baseline assessment (as was the case in PEPFAR) may increase the importance of using several methodologies, including qualitative measures. Speaker Glenzer reinforced the point with his comment that centrally planned, mixed-method evaluation designs work best.

At the same time, the use of multiple methods should be strategic, noted workshop speaker Glennerster. She noted that currently organizations often conduct a confused mix of process/output and impact evaluations in too many places. Instead, she recommended conducting good process evaluations everywhere and a moderate number of high-quality impact evaluations focusing on a few key questions.

Value of Randomization

Multiple presenters emphasized the value of randomization tools in the conduct of evaluations. Glennerster pointed out that new methods of randomization are now available that integrate with evaluation with

minimal disruption. In his presentation, Bertozzi also drew on evidence from randomized controlled trials. Speaker Field-Nguer pointed out that nonrandom selection of sites has the potential to limit or weaken a study. Workshop participant De Lay discussed some of the potential problems with impracticality of randomization.

Comparison Across Contexts

Several workshop participants stressed the highly contextual nature of change when comparing across contexts. Evaluations that are centrally coordinated to permit comparison of variables across contexts, while allowing some flexibility in indicator design at the local level, are optimal, suggested speaker Glenzer. Interventions that are successful in one country are not necessarily transferable to another country, noted workshop speaker Stoneburner. Examples provided by Stoneburner and speakers Latkin, Garnett, and Pulerwitz supported this statement. In some cases, factors independent of an explicit program intervention can have an influence on change. In other cases, change in behavior does not always lead to a change in the pattern of the HIV/AIDS epidemic, and changes in the pattern of the epidemic cannot always be translated to a change in behavior. Close engagement of the scientific community in evaluation, urged speaker Latkin, can help to assess the likelihood of transferability of effective programs to other settings.

Appendix A

Agenda

Design Considerations for Evaluating the Impact of PEPFAR
Monday, April 30–Tuesday, May 1, 2007
Institute of Medicine
Keck Building, 500 Fifth Street, N.W., Room 100, Washington, DC

Purpose: To discuss methodological, policy, and practical design considerations from the three main perspectives on accountability:

- "Upward"—Congress
- "Horizontal"—Global Partners/Coordination
- "Downward"—Country Partners/Harmonization

The workshop was moderated by Ruth Levine, director of programs and senior fellow, Center for Global Development.

Portions of the workshop were webcast thanks to the Kaiser Family Foundation and are available at http://www.kaisernetwork.org/healthcast/ iom/30apr07.

Monday, April 30, 2007: Perspectives on Evaluation

9:00–9:10 a.m. Welcome

Michele Orza, ScD
Study Director, PEPFAR Implementation Evaluation
Institute of Medicine (IOM)

9:10–9:15 a.m. Introduction and Framing the Issues

Ruth Levine, PhD
Director of Programs and Senior Fellow
Center for Global Development

9:15–10:15 a.m. Congressional Perspective

Purpose: To understand what Congress wants and needs to learn from impact evaluation; to discuss the evaluation language in The Leadership Act.

Savannah Lengsfelder, MA
Legislative Assistant
U.S. Senate Committee on Foreign Relations, Subcommittee on African Affairs

Christos Tsentas
Senior Legislative Assistant
Office of Representative Barbara Lee

David Gootnick, MD
Director, International Affairs and Trade
U.S. Government Accountability Office

Allen Moore, MBA
Senior Fellow, Global Health Council
Senior Associate, Center for Strategic and International Studies

Jim Sherry, MD, PhD (Discussant)
Subcommittee Member, IOM Committee for the Evaluation of PEPFAR Implementation
Professor and Chair, Department of Global Health
George Washington University

10:15–11:15 a.m. PEPFAR Perspective

Purpose: To understand the U.S. Global AIDS Coordinator's overall strategy for monitoring and evaluation, how impact evaluation fits into it, what program officials need and want to learn from impact evaluation, how they plan to coordinate and harmonize.

Tom Kenyon, MD, MPH
Principal Deputy U.S. Global AIDS Coordinator, Chief Medical Officer
Office of the U.S. Global AIDS Coordinator

Kathy Marconi, PhD, MS
Director of Monitoring, Evaluation, and Strategic Information
Office of the U.S. Global AIDS Coordinator

Shannon Hader, MD
Senior Scientific Advisor
Office of the U.S. Global AIDS Coordinator

11:15–11:30 a.m. Break

11:30 a.m.–1:00 p.m. Perspectives of Global Partners

Purpose: To learn the perspective of other donors with whom the United States should be coordinating with respect to monitoring and evaluation.

Paul R. De Lay, MD
Director, Monitoring and Evaluation
Joint United Nations Programme on HIV/AIDS

Jody Kusek, PhD
Lead Monitoring and Evaluation Specialist, Global HIV/AIDS Program
World Bank

Julia Compton, PhD
Senior Evaluation Manager, Evaluation Department
UK Department for International Development

Ambassador Jimmy Kolker (Discussant)
Deputy U.S. Global AIDS Coordinator
Office of the U.S. Global AIDS Coordinator

Jim Sherry, MD, PhD (Discussant)
Subcommittee Member, IOM Committee for the Evaluation of PEPFAR Implementation
Professor and Chair, Department of Global Health
George Washington University

1:00–1:30 p.m. Lunch Break

1:30–2:15 p.m. Luncheon Speaker and Discussion

"Monitoring and Evaluating Fairness in Scaling-Up ART"
Norman Daniels, PhD
Professor of Ethics and Population Health
Harvard School of Public Health

2:15–4:15 p.m. Perspectives of Implementers and Country Partners

Purpose: To learn the perspective of various partners and stakeholders in the impact evaluation—the focus countries, program implementers, other programs with which the United States should be harmonized with respect to monitoring and evaluation, advocacy groups.

Agnes Binagwaho, MD
Executive Secretary
Rwanda National AIDS Control Commission

Jonathan Mwiindi
Director, Kijabe HIV/AIDS Relief Program, Kenya
HIV/AIDS Program Officer
Ecumenical Pharmaceutical Network

Nils Daulaire, MD, MPH
President and Chief Executive Officer
Global Health Council

Mary Lyn Field-Nguer, MSN, FNP
Director, Global HIV/AIDS Programs/Washington
John Snow, Inc.

Kent Glenzer, PhD
Director, Impact Measurement and Learning Team
CARE USA

Sara Pacqué-Margolis, MPH
Director, Monitoring and Evaluation
Elizabeth Glaser Pediatric AIDS Foundation

4:15–4:30 p.m. Break

4:30–5:45 p.m. Combined Panel

Moderated discussion with all panelists and audience.

Tuesday, May 1, 2007: Evaluation Design Challenges and Solutions

Purpose: To benefit from the experience of people who have evaluated or are currently evaluating programs that have major elements in common with the U.S. Global AIDS Initiative.

9:00–11:00 a.m. Design Lessons Learned from Evaluating Donor Programs (Macro-, Meso-, Micro-Level)

Phillip Nieburg, MD, MPH (Moderator/Discussant)
Senior Associate and Co-Chair
Prevention Committee, HIV/AIDS Task Force
Center for Strategic and International Studies

Jaime Sepúlveda, MD, DrSc
Chair, IOM Committee for the Evaluation of PEPFAR Implementation
Visiting Professor and 2007 Presidential Chair
University of California–San Francisco

Martha Ainsworth, PhD
Lead Economist and Coordinator
Health and Education Evaluation Independent Evaluation Group
World Bank

Rachel Glennerster, PhD
Executive Director
Abdul Latif Jameel Poverty Action Laboratory

Julia Compton, PhD
Senior Evaluation Manager, Evaluation Department
UK Department for International Development

Kent Glenzer, PhD
Director, Impact Measurement and Learning Team
CARE USA

Stefano Bertozzi, MD, PhD
Member, IOM Committee for the Evaluation of PEPFAR Implementation
Member, The Technical Evaluation Reference Group–The Global Fund
Director, Division of Health Economics and Policy
National Institute of Public Health, Mexico

Jim Sherry, MD, PhD (Discussant)
Subcommittee Member, IOM Committee for the Evaluation of PEPFAR Implementation
Professor and Chair, Department of Global Health
School of Public Health and Health Services
George Washington University

11:00–11:15 a.m. Break

11:15 a.m.–1:15 p.m. Methodological Challenges (Meso-, Micro-Level)

Purpose: To discuss critical methodological issues and approaches for addressing them.
Focus: Evaluation of key AIDS-specific outcomes and impacts.

Paul R. De Lay, MD
Director, Monitoring and Evaluation
Joint United Nations Programme on HIV/AIDS

Theresa Diaz, MD, MPH
Branch Chief, Epidemiology and Strategic Information
Global AIDS Program, U.S. Centers for Disease Control and Prevention

Geoff Garnett, PhD
Member, IOM Committee for the Evaluation of PEPFAR Implementation
Professor of Microparasite Epidemiology
Imperial College

William L. Holzemer, RN, PhD, FAAN
Member, IOM Committee for the Evaluation of PEPFAR Implementation
Professor and Associate Dean, International Programs, School of Nursing
University of California–San Francisco

Rand Stoneburner, MD, MPH
Independent Consultant

Agnes Binagwaho, MD (Discussant)
Executive Secretary
Rwanda National AIDS Control Commission

Timothy Fowler, MA (Discussant)
Chief, Health Studies Branch, International Programs
U.S. Bureau of the Census

Rachel Glennerster, PhD
Executive Director
Abdul Latif Jameel Poverty Action Lab

Caroline Ryan, MD, MPH (Discussant)
Chief Technical Officer
Office of the U.S. Global AIDS Coordinator

1:15–2:00 p.m. Lunch Break

2:00–4:00 p.m. More Methodological Lessons/Challenges (Meso-, Micro-Level)

Purpose: To discuss critical methodological issues and approaches for addressing them.
Focus: Evaluation of key outcomes and impacts that are not AIDS specific.

Stefano Bertozzi, MD, PhD
Member, IOM Committee for the Evaluation of PEPFAR Implementation
Director, Division of Health Economics and Policy
National Institute of Public Health, Mexico

William L. Holzemer, RN, PhD, FAAN
Member, IOM Committee for the Evaluation of PEPFAR Implementation
Professor and Associate Dean, International Programs, School of Nursing
University of California–San Francisco

Carl Latkin, MS, PhD
Subcommittee Member, IOM Committee for the Evaluation of PEPFAR Implementation
Professor, Department of Health Policy and Management
Johns Hopkins School of Public Health

John Novak, PhD
Senior Monitoring and Evaluation Advisor, Office of HIV/AIDS
U.S. Agency for International Development

Jessica E. Price, PhD
Country Director, Rwanda
Family Health International

Julie Pulerwitz, ScD
Research Director, Horizons Program
Population Council, seconded from PATH

Martha Ainsworth, PhD (Discussant)
Lead Economist and Coordinator
Health and Education Evaluation Independent Evaluation Group
World Bank

Mary Lyn Field-Nguer, MSN, FNP (Discussant)
Director, Global HIV/AIDS Programs/Washington
John Snow, Inc.

Kent Glenzer, PhD (Discussant)
Director, Impact Measurement and Learning Team
CARE USA

Jonathan Mwiindi (Discussant)
Director, Kijabe HIV/AIDS Relief Program, Kenya
HIV/AIDS Program Officer
Ecumenical Pharmaceutical Network

Mead Over, PhD (Discussant)
Senior Fellow
Center for Global Development

4:00–4:15 p.m. Break

4:15–5:45 p.m. Combined Panel

Moderated discussion with all panelists and audience.
Summary of key issues and the way forward.

Appendix B

Abbreviations and Acronyms

AB	abstinence and faithfulness
ABC	abstinence–faithfulness–condom use
AIDS	acquired immunodeficiency syndrome
ANC	antenatal clinic
ART	antiretroviral therapy
ARV	antiretroviral
AZT	azidothymidine
BED	branched gp41 peptide (a test for HIV-1 incidence)
BUCEN	U.S. Bureau of the Census
CARE	Cooperative for Assistance and Relief Everywhere, Inc.
CD4	cluster of differentiation 4
CDC	U.S. Centers for Disease Control and Prevention
CHAT	Country Harmonization and Alignment Tool
CLPIR	community-level program information reporting systems
DFID	UK Department for International Development
DHS+	Demographic and Health Surveys-Plus
FDA	U.S. Food and Drug Administration
FHI	Family Health International

GEM	Gender Equitable Men's
HIV	human immunodeficiency virus
IOM	Institute of Medicine
M&E	monitoring and evaluation
MTCT	mother-to-child transmission of HIV
NGO	nongovernmental organization
OECD	Organization for Economic Cooperation and Development
OGAC	Office of the U.S. Global AIDS Coordinator
OVC	orphans and vulnerable children
PAL	Programa del Apoyo Alimentario
PEPFAR	President's Emergency Plan for AIDS Relief
PHE	public health evaluation
PLHAs	persons living with HIV/AIDS
PMTCT	prevention of mother-to-child transmission of HIV
SAVVY	sample vital registration with verbal autopsy
SWEF	System-Wide Effects of Global Fund
TB	tuberculosis
TE	targeted evaluation
The Global Fund	The Global Fund to Fight AIDS, Malaria, and Tuberculosis
UNAIDS	Joint United Nations Programme on HIV/AIDS
USAID	U.S. Agency for International Development
VCT	voluntary counseling and testing
WHO	World Health Organization

Appendix C

List of Participants

Martha Ainsworth, World Bank
Silvia Alayon, University of North Carolina–Carolina Population Center
Claudia Allers, John Snow, Inc.
Barbara Aranda-Naranjo, U.S. Department of Health and Human Services
Jeffrey Baldwin-Bott, U.S. Government Accountability Office
Madhusmita Baruah
Michael Bernstein, Office of the U.S. Global AIDS Coordinator
Stefano Bertozzi, Institute of Medicine Committee for the Evaluation of PEPFAR Implementation/National Institute of Public Health, Mexico
Paurvi Bhatt, John Snow, Inc.
Agnes Binagwaho, Rwanda National AIDS Control Commission
Deborah Birx, U.S. Centers for Disease Control and Prevention
Mike Boca, Center for Strategic and International Studies
Paul Bouey, Office of the U.S. Global AIDS Coordinator
Jessica Boyer, U.S. House of Representatives Oversight Committee
Ronald Brinn, Millennium Forum of NGOs
Vanessa Brown, Office of the U.S. Global AIDS Coordinator
Ellen Caldeira, Health Resources and Services Administration
Joanne Capper, U.S. Peace Corps
Yasmin Chandani, John Snow, Inc.
Man Charurat, Institute of Human Virology, University of Maryland
Kathleen Collins
Charlotte Colvin, Elizabeth Glaser Pediatric AIDS Foundation

Julia Compton, UK Department for International Development
Gordon Comstock, The Supply Chain Management System
Kelly Curran, Jhpiego
Norman Daniels, Harvard School of Public Health
Nils Daulaire, Global Health Council
Paul De Lay, Joint United Nations Programme on HIV/AIDS
Jordana De Leon, U.S. Department of State
Theresa Diaz, U.S. Centers for Disease Control and Prevention
David Dornisch, U.S. Government Accountability Office
Patrick Falwell, Center for Strategic and International Studies
Clint Fenning, Office of the U.S. Global AIDS Coordinator
Mary Lyn Field-Nguer, John Snow, Inc.
Kate Fleming, Plan International USA
Karen Foreit, Constella Futures
Timothy Fowler, U.S. Bureau of the Census
Geoff Garnett, Institute of Medicine Committee for the Evaluation of PEPFAR Implementation/Imperial College, United Kingdom
Moira Gaul, Family Research Council
Rachel Glennerster, Abdul Latif Jameel Poverty Action Laboratory, Massachusetts Institute of Technology
Kent Glenzer, CARE USA
David Gootnick, U.S. Government Accountability Office
Jessica Gottlieb, Center for Global Development
Nicole Gray
Sue Griffey, Social and Scientific Systems, Inc.
Giovanna Guerrero, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Shannon Hader, Office of the U.S. Global AIDS Coordinator
Celestin Hakiruwizera
Gray Handley, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Kathleen Handley, U.S. Agency for International Development
Kamden Hayashi, U.S. Peace Corps
David Henek
William Holzemer, Institute of Medicine Committee for the Evaluation of PEPFAR Implementation/University of California–San Francisco
Kathy Jacquart, U.S. Peace Corps
Aranthan Jones II, Office of the Majority Whip
Carmit Keddem, John Snow, Inc.
Tom Kenyon, Office of the U.S. Global AIDS Coordinator
Jimmy Kolker, Office of the U.S. Global AIDS Coordinator
Jody Kusek, World Bank
Anne LaFond, John Snow, Inc.

Carl Latkin, Institute of Medicine Committee for the Evaluation of PEPFAR Implementation/Johns Hopkins School of Public Health
Annie Latour, Office of the U.S. Global AIDS Coordinator
Savannah Lengsfelder, U.S. Senate Committee on Foreign Relations, Subcommittee on African Affairs
Elyse Levine, N. Chapman Associates
Ruth Levine, Center for Global Development
Charles Lule
Temina Madon, National Institutes of Health
Katherine Marconi, Office of the U.S. Global AIDS Coordinator
Veronica Miller, Forum for Collaborative HIV Research
Allen Moore, Global Health Council/Center for Strategic and International Studies
Brittany Moore, U.S. Senate Committee on Health, Education, Labor, and Pensions
Meade Morgan, U.S. Centers for Disease Control and Prevention
Jonathan Mwiindi, Kijabe HIV/AIDS Relief Program, Kenya/Ecumenical Pharmaceutical Network
Phillip Nieburg, Center for Strategic and International Studies
John Novak, U.S. Agency for International Development
Rachel Nugent, Center for Global Development
Nandini Oomman, Center for Global Development
Michele Orza, Institute of Medicine
Mead Over, Center for Global Development
Sara Pacqué-Margolis, Elizabeth Glaser Pediatric AIDS Foundation
Jenny Peterson, Office of the U.S. Global AIDS Coordinator
Laura Porter, U.S. Centers for Disease Control and Prevention
Jessica Price, Family Health International, Rwanda
Julie Pulerwitz, Population Council/PATH
Pamela Rao
Nathan Ricke, World Vision International
Jessica Rose
Sarah Roush, Christian Children's Fund
Caroline Ryan, Office of the U.S. Global AIDS Coordinator
Keith Sabin, U.S. Centers for Disease Control and Prevention
Laura Seaton
Naomi Seiler, U.S. House of Representatives Oversight Committee
Kathy Selvaggio, International Center for Research on Women
Karen Semkow, Social and Scientific Systems, Inc.
Shannon Senefeld, Catholic Relief Services
Jaime Sepúlveda, Institute of Medicine Committee for the Evaluation of PEPFAR Implementation/University of California–San Francisco
Daniel Seyoum

Michelle Sherlock, Office of the U.S. Global AIDS Coordinator
Jim Sherry, Institute of Medicine Subcommittee for the Evaluation of PEPFAR Implementation/George Washington University
Barry Silverman, GH Tech Project
Suam Smits, Office of Senator Richard Durbin
Audrey Solis, U.S. Government Accountability Office
David Stanton, U.S. Agency for International Development
Carl Stecker, Catholic Relief Services
Sara Steinmetz
Kate Stillman, Abt Associates Inc.
Rand Stoneburner, Independent Consultant
Ben Synder, U.S. Agency for International Development
Maureen Thanalappin, Office of the U.S. Global AIDS Coordinator
Christos Tsentas, Office of Representative Barbara Lee
Waimar Tun, Population Council
Festus Ukwuani, U.S. Agency for International Development
Tom Walsh, Office of the U.S. Global AIDS Coordinator
Tom Zingale, U.S. Government Accountability Office

Appendix D

References

Barker, G. 2000. Gender equitable boys in a gender inequitable world: Reflections from qualitative research and program development with young men in Rio de Janeiro, Brazil. *Sexual and Relationship Therapy* 15(3):263–282.

Binagwaho, A. 2007. *Considerations for evaluating the impact of PEPFAR: Rwandan perspective*. Speaker presentation at the Institute of Medicine Workshop on Design Considerations for Evaluating the Impact of PEPFAR, Washington, DC.

Daniels, N. 2005. Fair process in patient selection for antiretroviral treatment in WHO's goal of 3 by 5. *Lancet* 366(9480):169–171.

Daniels, N., and J. Sabin. 2002. *Setting limits fairly: Can we learn to share medical resources?* London, England: Oxford University Press.

Daniels, N., D. W. Light, and R. L. Caplan. 1996. *Benchmarks of fairness for health-care reform*. London, England: Oxford University Press.

Duflo, E., P. Dupas, and M. Kremer. 2006. *Education and HIV/AIDS prevention: Evidence from a randomized evaluation in western Kenya*. World Bank Policy Research Working Paper No. 402. Washington, DC: The World Bank.

Gregson, S., G. P. Garnett, C. A. Nyamukapa, T. B. Hallett, J. J. C. Lewis, P. R. Mason, S. K. Chandiwana, and R. M. Anderson. 2006. HIV decline associated with behavior change in eastern Zimbabwe. *Science* 311(5761):664–666.

Hader, S. 2007. *The U.S. president's emergency plan for AIDS relief: PEPFAR and public health evaluation*. Speaker presentation at the Institute of Medicine Workshop on Design Considerations for Evaluating the Impact of PEPFAR, Washington, DC.

Hallett, T. B., P. J. White, and G. P. Garnett. 2007. Appropriate evaluation of HIV prevention interventions: From experiment to full-scale implementation. *Sexually Transmitted Infections* 83:i55–i60.

Holzemer, W. L., and L. Uys. 2004. Managing AIDS stigma. *Journal of Social Aspects of HIV/AIDS Research Alliance/SAHARA/Human Sciences Research Council* 1(3):165–174.

Holzemer, W. L., L. R. Uys, M. L. Chirwa, M. Greeff, L. N. Makoae, T. W. Kohi, P. S. Dlamini, A. L. Stewart, J. Mullan, R. D. Phetlhu, D. Wantland, and K. Durrheim. 2007. Validation of the HIV/AIDS Stigma Instrument-PLWA (HASI-P). *AIDS Care—Psychological and Socio-Medical Aspects of AIDS/HIV* 19(8):1002–1012.

IOM (Institute of Medicine). 2007. *PEPFAR implementation: Progress and promise*. Committee for the Evaluation of the President's Emergency Plan for AIDS Relief (PEPFAR) Implementation. Washington, DC: The National Academies Press. http://www.nap.edu/catalog.php?record_id=11905 (accessed February 8, 2008).

IPPF (International Planned Parenthood Federation), GNP+ (The Global Network of People Living with HIV/AIDS), ICW (International Community of Women Living with HIV/AIDS), and UNAIDS (Joint United Nations Programme on HIV/AIDS). 2008. *The People Living with HIV Stigma Index user guide*. London, England: IPPF.

IRIN (Integrated Regional Information Networks). 2007. *Mozambique: HIV-infected women blamed and shunned*. Nairobi, Kenya: IRIN, U.N. Office for the Coordination of Humanitarian Affairs. http://www.irinnews.org/Report.aspx?ReportId=71727 (accessed March 5, 2008).

J-PAL (Abdul Latif Jameel Poverty Action Laboratory). 2007. *Policy Briefcase No. 3: Cheap and effective ways to change adolescents' sexual behavior.* Cambridge: Massachusetts Institute of Technology. http://www.povertyactionlab.com/papers/Briefcase%203%20HIV.pdf (accessed March 11, 2008).

Jha, P., L. M. E. Vaz, F. Plummer, N. Nagelkerke, B. Willbond, E. Ngugi, S. Moses, G. John, R. Nduati, K. S. MacDonald, and S. Berkley. 2001. *The evidence base for interventions to prevent HIV infection in low- and middle-income countries*. Working Paper Series WG5:2. http://www.whoindia.org/LinkFiles/Commission_on_Macroeconomic_and_Health_05_02.pdf (accessed February 28, 2008).

Kenyon, T. 2007. *The U.S. president's emergency plan for AIDS relief: The global impact of HIV/AIDS programs*. Speaker presentation at the Institute of Medicine Workshop on Design Considerations for Evaluating the Impact of PEPFAR, Washington, DC.

MOH (Ministry of Health, Uganda) and MACRO. 2006. *Uganda HIV/AIDS Sero-Behavioural Survey 2004–2005*. Calverton, MD: MOH and MACRO. http://www.measuredhs.com/pubs/pdf/AIS2/AIS2.pdf (accessed February 28, 2008).

Over, M., P. Heywood, E. Marseille, I. Gupta, S. Hira, N. Nagelkerke, and A. S. Rao. 2004. *HIV/AIDS treatment and prevention in India: Modeling the costs and consequences*. Washington, DC: The World Bank.

Over, M., A. Revenga, E. Masaki, W. Peerapatanapokin, J. Gold, V. Tangcharoensathien, and S. Thanprasertsuk. 2007. The economics of effective AIDS treatment in Thailand. *AIDS* 21(Suppl 4):S105–S116.

Pulerwitz, J., and G. Barker. 2008. Measuring attitudes toward gender norms among young men in Brazil: Development and psychometric evaluation of the GEM scale. *Men and Masculinities* 10:322–338.

Stoneburner, R. L., and D. Low-Beer. 2004. Population-level HIV declines and behavioral risk avoidance in Uganda. *Science* 304(5671):714–718.

Stoneburner, R. L., D. Low-Beer, G. S. Tembo, T. E. Mertens, and G. Asiimwe-Okiror. 1996. Human immunodeficiency virus infection dynamics in East Africa deduced from surveillance data. *American Journal of Epidemiology* 144:682–695.

Stover, J., and L. Bollinger. 2006. *Goals model for estimating the effects of resource allocation decisions on the achievement of the goals of the HIV/AIDS Strategic Plan, Version 2.91*. Glastonbury, CT: Futures Institute, http://www.healthpolicyinitiative.com/index.cfm?id=softwareDownload&name=GOALS&file=goals.zip&site=HPI (accessed March 6, 2008).

Thornton, R. 2007. *The demand for and impact of learning HIV status: Evidence from a field experiment*. Working paper and forthcoming in *American Economic Review*. Ann Arbor: University of Michigan Population Studies Center. http://www.povertyactionlab.com/papers/Thornton%20HIV%20Testing%20October%202007.pdf (accessed March 11, 2008).

UNAIDS (Joint United Nations Programme on HIV/AIDS). 2007a. *Country Harmonization and Alignment Tool (CHAT): A tool to address harmonization and alignment challenges by assessing strengths and effectiveness of partnerships in the national AIDS response*. Geneva, Switzerland: UNAIDS. http://data.unaids.org/pub/Report/2007/jc1321_chat_en.pdf (accessed February 20, 2008).

UNAIDS. 2007b. *Epidemiologic software and tools*. Geneva, Switzerland: UNAIDS. http://www.unaids.org/en/KnowledgeCentre/HIVData/Epidemiology/epi_software2007.asp (accessed February 28, 2008).

UNAIDS and WHO (World Health Organization). 2007. *AIDS epidemic update: December 2007*. Geneva, Switzerland: UNAIDS and WHO. http://data.unaids.org/pub/EPISlides/2007/2007_epiupdate_en.pdf (accessed February 8, 2008).

WHO. 2006. *Case study. Country-enhanced monitoring and evaluation for antiretroviral therapy scale-up: Analysis and use of strategic information in Botswana*. Geneva, Switzerland: WHO. http://www.who.int/hiv/pub/casestudies/evaluation/en/index.html (accessed February 8, 2008).