

**Are We Learning About What Really Works?
Lost Opportunities and Constraints to Producing
Rigorous Evaluation Designs
of Health Project Impact**

Charles H. Teller

NONIE WORKING PAPER NO. 9

January 2008

What is NONIE?

Nonie is a network of networks for impact evaluation comprised of the DAC Evaluation Network, The United Nations Evaluation Group (UNEG) and the Evaluation Cooperation Group (ECG). Its purpose is to foster a program of impact evaluation activities based on a common understanding of the meaning of impact evaluation and approaches to conducting impact evaluation.

To this end a task team has been constituted and tasked with the following activities:

1. Preparation of impact evaluation guidelines
2. Agreeing collaborative arrangements for undertaking impact evaluation, leading to initiation of the program
3. Developing a platform of resources to support impact evaluation by member organizations

NONIE Working Papers

NONIE working papers present conceptual papers and impact evaluation findings. They may have been published elsewhere, e.g. as government or agency reports, but are included in the NONIE series to increase dissemination. Feedback on papers via the NONIE website is welcome.

**Are We Learning About What Really Works?
Lost Opportunities and Constraints to Producing Rigorous Evaluation Designs
of Health Project Impact**

*Charles H. Teller**
Visiting Scholar
Population Reference Bureau

Abstract

Impact evaluation is often seen as a central building block of results-based management. But in USAID rigorous impact evaluations have been crowded out by the drive to outcome-monitoring in the name of the results agenda. Other constraints on adopting quantitatively well-designed impact evaluation designs have included a lack of the required technical skills amongst those charged with conducting evaluations, the lack of incentives to produce quality studies – including the desire to avoid uncovering weak performance – and hence a lack of political will to expand the impact evaluation program. The paper sets out a proposal for a new evaluation agenda to address these deficiencies.

I. Introduction and the Problem

Impact evaluation should have a central role in evidence-based policy making. But, whilst the USAID Health Bureau has prided itself on taking international leadership in evidence-based results monitoring and comprehensive evaluations, attention to evaluation has declined in recent years.

The earlier technical excellence of the USAID Health Bureau in evaluation methods was demonstrated by the development of the LOG FRAME in the late 1970s (Dunlop, 1981), and later developing tools (TIPS: <http://evalweb.usaid.gov/resources/tipsseries.cfm>) and operational guidance for Agency operations (CDIE, 1996; ADS, 1995). But since 1994 there has been a steep decline in the quantity and quality of evaluation; the number of evaluation studies falling from a peak of 497 a year to a low of 104 in 1998 (Clapp and Blue, 2001). This decline has been attributed to many factors, including Albert Gore's policy change on "reinventing government," a sharp reduction in technical staff, and a change in USAID Guidance from requiring every project to be evaluated to recommending that evaluations *only be done in response to management need*.

Since 2000, there have been several Agency-wide and Bureau specific reviews of evaluation, such as (e.g. Clapp and Blue, 2001, and Weber, 2004), which have demonstrated the loss of institutional learning and best practices. While the country

* Formerly Senior Technical Advisor in Monitoring/Evaluation, Global Health Bureau, USAID, Washington, DC.

USAID missions depended on evaluations, and so their greatest concern was the very limited number of in-depth program evaluations. Moreover, while the partners did many of the evaluations, USAID did not (op cit. p. iii), and those few evaluations supported by the missions were not being submitted to the Development Exchange Clearinghouse (DEC).

Evaluation quality has long been an issue with USAID, both Agency-wide (e.g., Hopstock et al., 1989) and in the technical bureaus (eg., Adamchak and Reynolds, 2004 for Health; Bollen et al, 2005 for Democracy and Governance). Many evaluation reports do not include more than a few paragraphs on method, and many were qualitative and unsystematic: the expatriate ‘fly-in’ assessment where a team comes for 2-3 weeks and bases its findings only on qualitative interviews with key informants and stakeholders. Most reviews of USAID evaluations found weak methods employed, even with external, professional evaluators whose objectivity was often compromised by their desire to please the managers and continue to be hired. This paper documents these trends in recent years, analysing the factors behind them and the steps required to ensure the production of more and better impact studies.

The focus of this study is on the methodological strength and design rigor of evaluations of outcomes, effects, and impacts. White (2006) defines impact as the counterfactual analysis of how an intervention affects final welfare outcomes. In that sense, we want to know if the donor-funded activities are attaining their expected results as set out in the project paper and the results framework or M&E design. However, since donors fund too few real impact evaluations of project attribution and the counterfactual, evaluators have often been limited to considering whether the project achieved its intended outcome in its intervention areas or groups, preferably as compared to control groups. Moreover, in order to inform future programming, evaluation must be transparent and externally credible to decision-makers. However, without the rigor of an impact evaluation these attributes are harder to achieve.

II. Methodology

There are three main approaches taken in this paper:

- (1) A document review of evaluation design, methods, training, and technical leadership in M&E was carried out. These included activities of the flagship USAID/Global Health (GH) evaluation-related projects: MEASURE/Evaluation (Monitoring and Evaluation to Assess and Use Results); Frontiers and Horizons Operations Research; Data for Decision-making, including: State of the Art papers, M&E Reference and Working Groups, Task Forces; International M&E and OR Regional Training and Capacity-building: M&E Working group- survey and needs assessment (2005);and finally, M&E Manuals, Guidelines, Indicators (see website at: www.cpc.edu/measure/evaluation).
- (2) A meta-analysis review of Global Health evaluation documents and their design in the last few years. These included: Strategic Planning and Results/Portfolio

Reviews in 2005 and 2006, which was a stocktaking of the number and types of evaluations per technical office; and a systematic meta-analysis of Global Health evaluations of designs and methodological approaches of 93 GH-funded projects, sub-projects from 2004-2006. Figure 1 shows the revised classification during the content analysis of evaluation designs:

[Figure 1]

- (3) An opinion survey and follow-up interviews with 47 experienced USAID evaluators. These surveys covered ten content areas, soliciting their experience about methodological rigor; the quality of the team; the issues raised; technical, operational, and behavioural constraints; and the use of the findings for policy and program change. There was dissemination of the findings, much discussion, and further in-depth investigation.

III. Findings and Results

Agency-wide and Global Health Planning and Leadership in Evaluation

Given the deterioration in quality and quantity of evaluations in the 1994-2004 period, and the highly critical 2004 assessment of evaluations (Weber, 2004), the USAID Administrator, Andrew Natsios, decided to revitalise evaluation, with the active implementation support of CDIE. Following direct consultations with field missions and central bureaus (Adams, 2005; Kerley, 2005), the Administrator sent a directive to the field outlining the strategy and actions needed, along with a much needed training package (Natsios, 2005). This process was cut short after less than a year by his replacement by a new Federal Director of Foreign Assistance.

Under Ambassador Tobias' new administration, CDIE, with its experienced evaluators, was disbanded, and its mother bureau moved over to the State Department, where all M&E was to be coordinated. The consequent F/State office dropped the revitalisation plans for evaluation, in favour of strict activity and output reporting (Operational Plans) for budgeting purposes (F/State, 2006). Moreover, the quantitative and analytical Results Frameworks and Packages, Strategic Objectives and Intermediates Results, and Performance Monitoring Plans (PMPs) were downplayed, and outcome-oriented annual reports were reduced to annual operational plans for output reporting.

In 2003-4, the Global Health Bureau was entering into a paradigm shift whereby the Emergency Presidential Initiatives in HIV/AIDS (PEFPAR), in malaria (PMI), and Avian Influenza were also shifting away from evaluation and towards strategic information systems and output monitoring. F/State required the development of common indicators under eight common objectives or "elements" and cross-cutting sub-elements. Whilst defining common indicators assists in the alignment of monitoring indicators with agency objectives and the aggregation of results, there is a tension with the need for M&E systems to be context-specific. More importantly, outcome monitoring is not the same as evaluation since it says nothing about attribution.¹

However, the Family Planning/RH and MCH/N office management and technical advisors agreed to maintain their technical leadership in evaluation, and developed several strategies to strengthen M&E. They:

- (1) Supported their own Assessment of Program M&E (Adamchak and Reynolds, 2004) which recommended less monitoring and more rigorous evaluation
- (2) Formed an M&E working group among the USAID partners and collaborating agencies, which recommended more training and sharing of lessons learned (Teller and Pandit, 2004)
- (3) Developed a new conceptual framework and indicators, designed program “pathways” and coordinated annual results reporting of accomplishments.
- (4) Used the showcase MEASURE, and sub-cooperating agencies, to refine the tools and methods necessary to generate, analyse, coordinate, and disseminate information for evaluation, program learning, and capacity building, and to develop a program results system.

Findings of the Meta-analysis of methodological rigor

Here we highlight those findings relevant to the rigor of the design of outcome and impact evaluations, both for projects and sub-project components. In the introduction to this issue, White and Bamberger define the stronger ‘real world’ evaluations as those with a quasi-experimental design, particularly those which have baseline and end-line data and both treatment and control groups (see also Bamberger *et al.*, 2006). The USAID clearinghouse contains only 31 evaluation studies a year for the period 2004-2006 (Figure 2, a total of 93), though as already noted, not all documents are submitted and others are restricted as procurement sensitive.

[Figure 2]

Most of the 93 assessments on evaluation design were pre-post child survival project designs (Figure 3), as required by the USAID Child Survival Grants Program, while other studies had post-hoc end-line designs, which are often unable to deliver reliable impact estimates. The non-child survival evaluations had stronger designs: quasi-experimental with comparison groups (Figure 3). In the DEC classification scheme, most of those classified as assessments or intervention studies were more rigorous than those classified as evaluations (Figure 4).

[Figure 3]

[Figure 4]

Expert survey and in-depth interviews

From the list of Team Leaders, frequent evaluator practitioners, and USAID evaluations advisors, over 60 experts were selected. Most were sent semi-structured surveys and did respond, sometimes in person; over 90 per cent were followed up by telephone or inter-personal interviews

The main findings are:

Evaluation planning and design: there is a lack of adequate project needs assessments prior to baselines; an over-reliance on outputs and milestones, not on project outcomes or attribution; and project designs that are often driven by externalities and “fashions of the month”.

Evaluation Implementation: team members, stakeholders, and counterparts were often inexperienced and untrained in evaluation methods, e.g. former USAID Mission directors hired as Team Leaders despite little hands-on experience in evaluation methods. Also, USAID managers were critical of external evaluators coming in with their own “agendas,” lacking enough knowledge of project context, and seeking to please senior administration officials. We also found a lack of an enabling environment from senior management and an erosion of the culture of evaluation in the field.

Follow-up and Use of Evaluations: a common criticism was the undue influence by key stakeholders over final recommendations. Most frustrating to the evaluation experts was that critical project decisions were made before evaluation finished. Evaluators rarely found out whether their key recommendations were carried out or not.

A key question raised by USAID strategic health planning, other technical, and program offices in preparation for designing a new evaluation agenda, given the shortage of time and resources, is *what and when is it best to evaluate*. Below (in figure 5) is a summary of expert opinion on the subject:

[Figure 5]

IV. Discussion: Constraints to Producing and Using Rigorous Evaluations

There are no-doubt many constraints to designing and implementing strong, rigorous health evaluations, as noted in the newer articles including White and Bamberger in this issue (see also Bamberger et al, 2006; Victora et al, 2007). From my experience as evaluation advisor within the Global Health Bureau, inside USAID country missions, while on secondment to developing country governments, and as a consultant in M&E for health sector NGOs and private sector partners, I would interpret the main constraints recorded from the experts inside and outside USAID/GH into four categories:

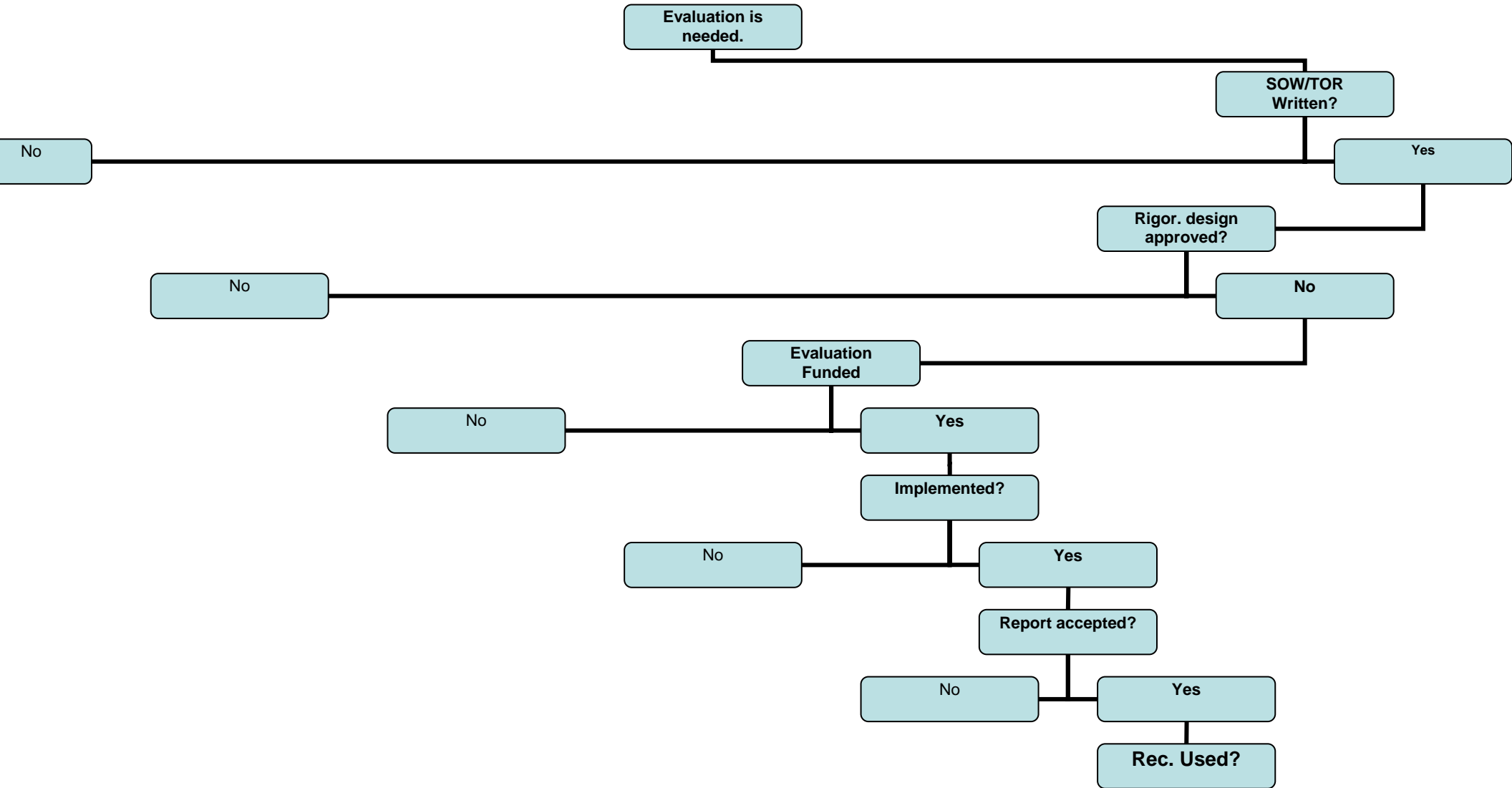
- (1) Technical: Contemporary impact evaluation designs need to be rigorous, requiring multivariate analytical skills. Training opportunities for these skills are limited, and experience in their application is even more limited. Moreover, some of the content of these training courses is not that useful in the real world of health

and development (Reynolds et al, 2006), or in poor and unstable countries. There is concern amongst donors over lack of usefulness of some past evaluations, especially when these have been superficial or carried out at times when they were not needed. Moreover, M&E experts are located more within institutional partners and independent consultants than within the donor.

- (2) Organisational: There is common knowledge of hiding evaluations: many evaluations not made public, either because they are procurement sensitive and then never released, too critical, or poorly done. Organisations tend to self-promote and put a positive ‘spin’ in their own self-interest, advertising competencies for new proposal competition. Moreover, there has been decline in number and continuity of USAID technical advisors and contract officers, and a lack of senior M&E positions in many country missions and Washington GH offices. As in most international organisations, there is a tension over local autonomy of program resource use between central and country USAID offices.
- (3) Political: The Agency’s effort in 2005 to revitalise evaluation demonstrated the inadequate political will to make sufficient resources available for its implementation. What took priority in 2006 was the explicit absorption of USAID into the new State Department policy goal of Transformational Diplomacy and a focus on politically-important countries (e.g., Egypt, Israel, Iraq, Afghanistan, Pakistan). Negative evaluations are feared to play into the hands of the foreign aid critics in Congress and the State Department, thus there is a fear of the visibility of failures and mistakes.
- (4) Behavioural: These restraints include the career ambitions, roles, and interests of individuals and affinity groups to act for their own benefit. Careers can be threatened, given the lack of a reward system for strong evaluations in order to learn from mistakes. There is a lack of motivation by managers, and a lack professional incentives for technical advisors, for rigorous evaluations. Finally, the time pressures of administrative tasks often crowd out the institutional learning tasks.

Figure 6 represents in schematic form the frustrating experience of what has often happened to the proposed or actual designs of rigorous evaluations in the last 4 years. It is estimated that few (less than 20 per cent) of these designs were either approved for funding, or implemented with rigor. All the real world constraints came to bear in creating this major leakage of credibility.

Figure 6 Missed Opportunities for Rigorous Evaluation: A Tree of Leakages from Assessed Need to Unused Recommendations



V. Conclusions and Recommendations for Institutional Learning

This paper has documented the precipitous decline in the number and quality of USAID-wide evaluation designs accessed through the USAID Document Exchange Clearinghouse (DEC) between 1994 and 2006. We used three methods to analyze this decline and the factors behind it:

- (1) Results and Portfolio Review assessment on evaluation of over 100 on-going centrally-funded projects in Global Health in FY 2005-6;
- (2) Meta-analysis of 93 project and sub-project/intervention evaluations done in 2004-06 accessed in the DEC; and
- (3) Opinion survey and interviews of 47 experienced USAID evaluation experts, both internal and external, about their issues and suggestions.

The findings point to the legitimacy of the concern of the senior management about the weaknesses of the Results Review component of the Portfolio Review Process. There are too many missed opportunities as well as institutionalised constraints to learning more about what works and what needs improvement in the GH portfolio. Results reporting can not take the place of impact studies to examine attribution. Thus transparent and evidenced-based policy and strategic planning decisions have been compromised in the process. The constraints to stronger evaluation discussed above- the technical, organisational, political, and behavioural- are formidable, and will require a concerted effort to overcome.

Implications and Recommendations

After the aborted attempt of the Natsios administration to revitalise evaluation, and the F/State focus on short-term operational plans and budget-related output monitoring, the GH Bureau has been considering measures to revitalise evaluation. The demise of CDIE on the one hand, the lack of an autonomous M&E unit in Global Health, and the complicated US Government inter-agency evaluation structures of the Presidential Health Initiatives work against a credible, flexible, and transparent approach to program evaluations.

The development of a new evaluation agenda during FY 08 is an important first step for USAID, though it must be supported by the political will to allocate sufficient priority and resources in support of both central and mission learning needs. This could be considered as moving forwards in a more coordinated manner with other donors on more rigorous impact and program evaluation. A minimum set of recommendations for this new evaluation agenda are suggested below:

- A high level, autonomous (from the technical offices) evaluation group (AEG) or division, including strategic information and analysis, is formed in the Program Office. It should be able to design rigorous impact and outcome evaluations, including cost-effectiveness; complemented by a stable consultant group of independent health evaluation experts
- An evaluation agenda is developed by the AEG in consultation with the technical offices and given a proper place within the new Health Strategy

- Rigorous program evaluations, guided by the Bureau-Wide evaluation, are required every five years before the next project approval cycle; needs assessments, baselines, and mid-term M&E systems are required in most projects
- Capacity-building, professional development, and training, guided by best practices in M&E, are supported for M&E and program design staff– both within USAID and for partners
- A responsive knowledge and documentation centre is reorganised to serve the information, analysis, evaluation, and decision-making needs of the Global Health Bureau. This should be organised in coordination with the Agency-wide centre.

REFERENCES

- Adams, D. (2005) 'Initiative to Revitalize the Evaluation System' Memo to Missions, USAID, July
- Adamchak, S., J. Reynolds and J. Henn (2004) 'Assessment of the M&E in Projects Managed by the BGH, OPRH; LTG Inc. and SSS Inc.', Washington DC, August
- AED (2004) 'Child Survival in Sub-Saharan Africa: Taking Stock, an Overview' Support for Analysis and Research in Africa (SARA) Project, USAID, September
- Bamberger, M, J. Rugh and L. Mabry. (2006) *Real World Evaluation: Working Under Budget, Time, Data and Political Constraints* California: Sage Publications
- Bollen, Kenneth, P. Paxton and R. Morishima (2005) 'Assessing International Evaluations: An Example from USAID's Democracy and Governance Program.' *Am Journal of Evaluation* 26.2
- Bryce, J. C. Victora, et al., (2005) *Ten Methodological Lessons for the Multi-country Evaluation of Integrated Management of Childhood Illness* Oxford: Oxford University Press
- Center for Global Development (2006) *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Savedoff, Levine and Birdsall (eds.) Washington, DC
- Clapp-Wincek, C. and R. Blue. (2001) *Evaluation of Recent USAID Evaluation Experience*. Working Paper No. 320, Washington DC: CDIE, USAID, June
- Ellis, V. and B. Emrey. POPTECH Reports. Selected Sensitive but Unclassified (SBU) Evaluations/Assessments (assembled for Charles Teller, July 20, 2007).
- DFID (2005) 'Guidance on Evaluation and Review for DFID Staff' Evaluation Department, July
- Goldenberg, D. (2003) 'The Mega-evaluation 2002 Evaluation: A Review of Findings and Methodological Lessons' from CARE Final Evaluations, 2001-2002. February
- Hader, S. (2006) 'PEPFAR and Next Steps with Public Health Evaluation' Powerpoint, OGAC
- Kerley, J. and K. Croake. (2005) Report of the Results of the Survey on Evaluation Revitalization, PPC/CDIE/ESPA, October
- Kuo, Kate and R. Sprout, Evaluations: Mission Trends, Agency Updates and Future Directions; Eastern Europe and Eurasia Region, USAID, (powerpoint presentation, Sept. 2007)
- Mabbs-Zeno, C. and C. Teller. Questions that should be answered by the Evaluation Guidance: Feedback from the Technical Office of USAID/GH; SPBO, Dec. 2006 (memo)

MEASURE/Evaluation Project. (2005) 'Results Framework', Annual Report, Carolina Population Center, UNC/Chapel Hill, October

Natsios, A. (2005) 'Actions required to Implement the Initiative to Revitalize Evaluation in the Agency', USAID Directive Cable, June 17

Perry, H. And P. Freeman. (2007) 'How Effective is Community-Based Primary Health Care in Improving the Health of Children? A Preliminary Review of the Evidence' Report to the World Health Organization, UNICEF and the Expert Review Panel. Draft June,

Rossi, P.H., H. E. Freeman and M.W. Lipsey. (1999) *Evaluation: A Systematic Approach* California: Sage Publications

Teller, C., (2006) 'Issues concerning the future of the MEASURE/Evaluation Project' Internal Memo to M/Evaluation Management Team, USAID, September

Teller, C. et al., (2007a) 'Meta Analysis of USAID/GH Evaluations, 2004: Designs, Gaps, Classifications and Towards Rigor' July (powerpoint)

Teller, C., (2007b) 'Summary, Meta-Analysis of USAID Project and Intervention Evaluations' August

Themessi-Huber, M. (2007) 'Evaluability Assessment: An Overview The Evaluator', Spring

USAID (2004) Automated Directive System (ADS), Chapter 203.3.6, Evaluations and Assessments, USAID (updated March 19, 2004).

USAID. Guidelines for Management Reviews and External Evaluations. Evaluation Process Improvement Committee (EPIC) May, 2002

USAID, Africa Bureau. SARA II Evaluation, 1999-2005. R. Simpson and R. Augustin, March, 2005

USAID, Performance and Monitoring TIPS Series.
www.dec.org/partners/evalweb/resources/tipsseries.cfm.

USAID, GH. Guidelines for Management Reviews and Project Evaluations. Draft, July 30, 2007

Victora, CG., J-P Habicht and J. Bryce. (2004) 'Evidence-based Public Health. Moving beyond Randomized Trials' AJP 94

Weber, J. (2004) 'An Evaluation of USAID's Evaluation Function: Recommendations for Reinvigoration the Evaluation Culture within the Agency' Washington DC: PPC, USAID, September

Winfrey, Bill, K. Foreit (2007) Poverty and Equity Considerations in USAID Mission Strategic Plans (unpublished draft)

White, H. (2004) 'Using the MDGs to measure agency performance' in R. Black and H. White (eds.) *Targeting Development: critical perspectives of the Millennium Development Goals*, London, Routledge

White, H. (2006) *Impact Evaluation: the Experience of the Independent Evaluation Group of the World Bank* Washington DC: IEG, World Bank

World Bank (2003) 'Independent Evaluation: Principles, Guidelines and Good Practice', Development Grant Facility Technical Note, November

World Bank (2006) Operations Evaluation Department (OED) and Impact- A Discussion Note.

World Bank (2006) 'Accelerating the Results Agenda: Progress and Next Steps', OPCS, June

World Bank (2004) OED's M&E: Some Tools, Methods and Approaches, 2nd edition

Figure 1 Classification System for Evaluation Designs:	
Revised Classification of USAID Health Project Evaluation Design	
<p>A. Program interventions and project components:</p> <ol style="list-style-type: none"> 1. Post-hoc cross-sectional 2. Baseline & post intervention 3. Baseline & post intervention with comparison/control area/group (QED-light) 	<p>B. Child Survival projects:</p> <ol style="list-style-type: none"> 1. Post-hoc cross-sectional 2. Baseline and post intervention

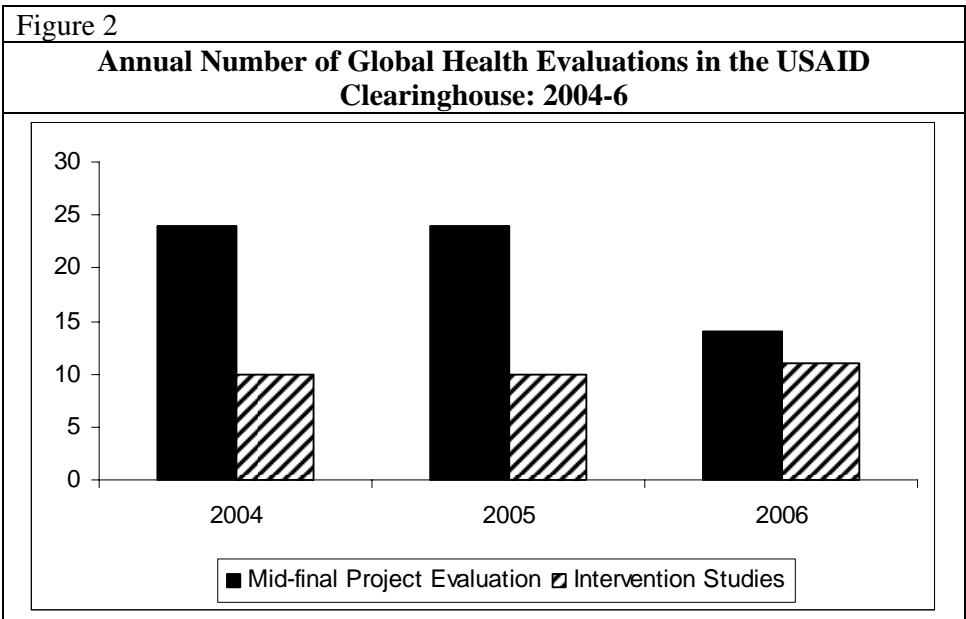


Figure 3
Evaluation Design Classifications: 2004-6

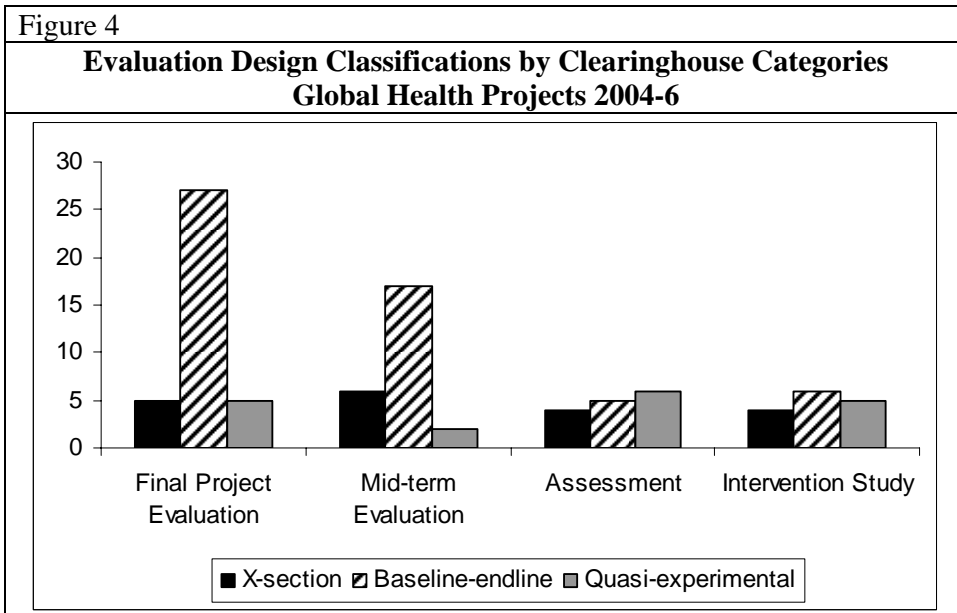
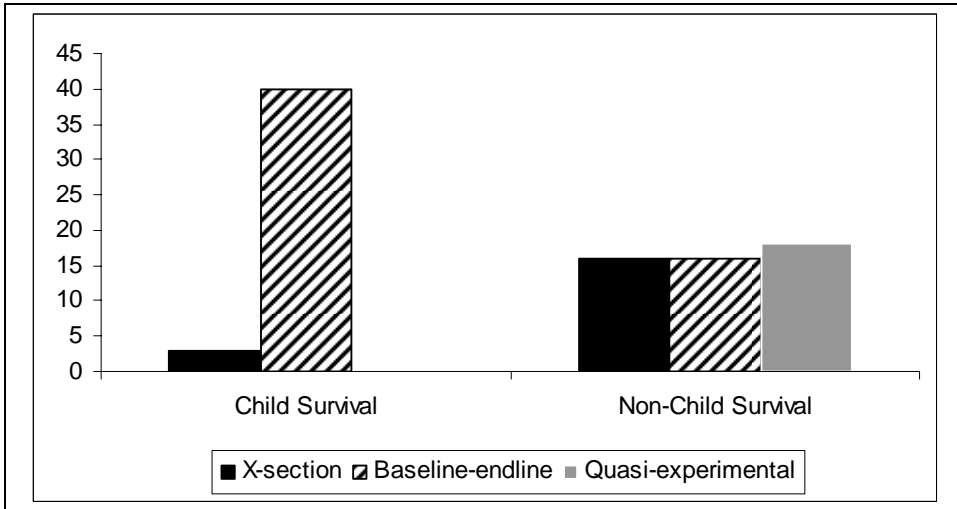


Figure 5
What & When Is It Best To Evaluate?

<p><u>What:</u></p> <ul style="list-style-type: none"> • M&E systems built into project proposal before award decision • Critical interventions need strong evaluations • Innovative activities & interventions whose practical effectiveness not yet rigorously evaluated • Multidisciplinary and contextual factors (the counter-factual; value-added) 	<p><u>When:</u></p> <ul style="list-style-type: none"> • Decisions on replication, extension, scaling up • At end of 4th year of most five-year projects • “Monitor extensively, evaluate selectively” • Major challenges need in-depth analysis into the causes • Demographic and Health Survey (DHS) data show unexpected changes
---	---

¹ Reference is made here to the triple-A requirement for Agency-wide monitoring systems: aggregation (being able to aggregate results across interventions), attribution (being able to link changes in outcomes to the intervention), and alignment (if indicators of project success indicate achievement of the agency’s overall goals); see White (2004).