In: Structural Approaches in Public Health, M. Sommer & R. Parker (eds), Routledge Press 2012 (forthcoming)

Evaluating structural interventions in public health: Challenges, options and global best-practice

Paul Pronyk[1,2,3], Jennifer Schaefer[1], Marie-Andree Somers[1] and Lori Heise[4]

[1] Centre for Global Health and Economic Development, The Earth Institute, Columbia University, USA
[2] Mailman School of Public Health, Columbia University, USA
[3] School of Public Health, University of the Witwatersrand, South Africa
[4] Gender Violence and Health Center, Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, UK

INTRODUCTION

Structural interventions in public health pose a host of challenges to evaluation. They operate through indirect pathways and are often complex and cross-sectoral. Programs may require extended time horizons for health effects to be observed, and their delivery at the level of communities, institutions or populations carries major implications for sampling. Finally, issues of ethics, logistics, and/or political feasibility may limit opportunities for random assignment and the use of experimental designs.

Nonetheless, evaluation is vital – both for strengthening the link between good science and sound policy, as well as ensuring public confidence in how limited resources are deployed. This chapter will explore the range of potential options and considerations when evaluating structural interventions in public health. We will draw upon global experience to explore the merits and limitations of adequacy assessments, plausibility evaluations and randomized experiments, including stepped-wedge designs. We will discuss the challenges of choosing an appropriate comparison group, the importance of impact pathway assessment, and the use of time-series monitoring. Finally, we will highlight the value of mixed-methods approaches, both to assess impact and to document the implementation principles that carry wider application to public health policy and program development.

Here we define structural interventions in public health as interventions that attempt to engage the complex social, economic, and political determinants of health as a way of influencing more downstream outcomes (Blankenship et al., 2000). Such interventions operate at the level of groups or populations and generally attempt to shape an individual's risk of disease through indirect mechanisms. In his seminal work, Rose argued that disease in populations is more a reflection of the mean-level of risk in that society, rather than simply a product of cumulative, independent, individual choices (Rose, 1985). Structural interventions seek to influence this mean-level of risk by engaging up-stream dynamics and conditions, with the aim of shaping norms, behaviors and health outcomes in the population as a whole.

WHY EVALUATE?

There is an ever increasing demand for accountability and evidence-based decision-making in public health. Unfortunately, in the past many structural interventions have been poorly evaluated, evaluated only in retrospect, or not evaluated at all (Savedoff et al., 2006). The paucity of evaluation literature, particularly for complex interventions, means that policy-makers can be forced to make decisions in a vacuum, risking repeating mistakes of the past and failing to benefit from the lessons of successful interventions (Campbell et al., 2000).

There is a clear need to improve the evidence-base for structural interventions in public health. This chapter underscores the many challenges encountered when evaluating structural interventions, highlighting implications for evaluation design. Drawing from a diverse range of disciplinary perspectives, we then profile a range of tools and methods that can be employed to conduct appropriate and rigorous program evaluations.

WHAT MAKES STRUCTURAL INTERVENTIONS DIFFICULT TO EVALUATE?

Policy-makers and program managers in public health are accustomed to weighing, synthesizing and applying new evidence in relatively standardized ways that are heavily influenced by evaluations of discrete, downstream, technical interventions. Evidence from randomized

controlled trials, systematic reviews and cost-effectiveness studies has played an important role in shaping policy and informing practice. However, for numerous reasons, weighing evidence from structural interventions may be different – with such 'gold-standard' evaluations often unavailable or inappropriate (Victora et al., 2004, Shepperd et al., 2009, McKee et al., 1999, Bonell et al., 2006a). We outline a number of the reasons for this below.

Context

Structural interventions are by nature contextual. Strategies to address economic barriers, engage legal systems, or shift cultural norms and power relationships will differ from place to place. Contextual factors influence implementation, affect uptake and utilization, and carry implications for reproducibility and external validity of evidence generated from program evaluations (Campbell et al., 2000). Evaluation designs should therefore employ an appropriate mix of methods to describe and document context, noting the ways this might affect the interpretation of program results.

Understanding contextual factors often requires qualitative research which can be conducted during or prior to an intervention, For example, a pre-intervention observational study assessing the feasibility of introducing schools-based sexual health programs in Tanzania identified many potential barriers including low attendance rates, lack of trust between students and teachers, and limited teacher training (Plummer et al., 2007b). This knowledge allows such barriers to be considered in the subsequent process and impact evaluations (Plummer et al., 2007a, Ross et al., 2007).

Multi-sector focus

The design and evaluation of interventions to address structural determinants of health often requires input from multiple sectors. Despite this, working across disciplines happens too infrequently - with funding mechanisms, the design of academic institutions, and the organization of government departments creating many barriers to such partnerships. New insights and novel innovations have the potential to be important by-products of working

between sectors. In an interdependent world with many complex and interrelated challenges, cross-sectoral perspectives are necessary now more than ever.

Complexity

Complex interventions have multiple interacting components, with the behavior of implementers and the response of program recipients influencing intervention delivery and results (UK Medical Research Council, 2006). Progression from one program phase to another may be non-linear, iterative and adaptive, making strict adherence to an implementation protocol potentially inappropriate (Campbell et al., 2000). This poses obvious challenges to conventional evaluation methods – where standardized exposures, discrete hypothesis testing, and pre-stated effect measurement are generally the norm.

Examples of complex interventions include health systems strengthening programs – such as the Integrated Management of Childhood Illness (IMCI – See **Case Study 1**) which simultaneously addressed improvements in case-management; improvements in health systems; and improvements in family and community practices which aimed to improve child survival (Arifeen et al., 2009, Bryce et al., 2004); or a multi-layered intervention with legal, policy and media components to reduce violence and HIV risk among sex workers in India (Beattie et al., 2010) (see **Case Study 2**).

Timeframes

Because structural interventions act indirectly, longer-time horizons may be required to detect measurable effects on downstream outcomes. It has also been suggested that gains achieved through addressing the structural roots of problems, rather than targeting more immediate risk behaviours, may be better sustained over time. Both issues have implications for the duration of program evaluations. Furthermore, governments and donors are often themselves bound by short-term time horizons, creating pressure for more immediate results.

While longer-term evaluations are required to measure some downstream outcomes and to assess sustainability, there is also ample evidence suggesting that structural interventions have the potential to exert effects over relatively short time horizons. A comprehensive structural intervention to address undernutrition across nine rural sites in sub-Saharan Africa reduced rates of childhood stunting by over 40% in just three years (Remans et al., 2011) (See **Case Study 3**). The Intervention with Microfinance for AIDS & Gender Equity (IMAGE Study) demonstrated major reductions in rates of intimate partner violence and HIV-related risk behavior after just 2 years of exposure to an integrated microfinance and training program (Pronyk et al., 2006, Pronyk et al., 2008b) (See **Case Study 4**). Thus while longer time horizons are advocated for evaluations of structural interventions, shorter-term interim follow-up should also be included in the timing of assessment rounds.

Sampling

As structural interventions are often delivered at the level of institutions or communities, these become the primary unit of assessment. Detecting statistical differences between exposed and unexposed groups is generally more strongly influenced by the total number units sampled rather than the number of individuals within each unit (Hayes and Bennett, 1999, Hayes et al., 2000). This has serious time and resource implications for program evaluations as interventions must be delivered to sufficient numbers of units; with adequate intensity of exposure achieved in each; and sufficient follow-up time to discern measurable effects.

Random allocation

Random assignment of structural interventions is often difficult. It may not be appropriate (such as in the case of a national policy shift or media campaign), ethical (such as a poverty reduction program that cannot be politically withheld from some groups over others) or politically feasible (where governments feel political imperatives are critical in determining who is exposed or unexposed) (Bonell et al., 2006a). Non-random choices of intervention groups have the potential to introduce selection bias – where exposed and unexposed groups may differ or not be sufficiently representative. In addition, it is difficult to conduct blinded assessments or

offer placebos for structural interventions, therefore issues of response bias are important to consider. A more detailed discussion of whether and when to randomize is presented below.

Secular change

Secular change includes the background mix of local and national policies and programs, and other dynamic factors affecting the health outcomes of intervention participants and the wider community, including any comparison group. Secular change can pose major challenges to the evaluation of structural interventions. Factors such as unanticipated policy and program shifts, forces in the media, and/or economic changes all carry the potential to shape social attitudes, behaviors and health outcomes. Unlike a clinical trial, where control groups are generally unexposed to an intervention, evaluations of structural interventions are often confounded by secular change. For example, major long-term evaluations of community-level programs to address cardiovascular disease have yielded disappointing results partially as a result of the rapid pace of social change around healthier diets, greater levels of exercise and reductions in smoking (Susser, 1995). Similar challenges are faced for evaluations of programs to scale-up proven interventions in program-rich environments, where many similar interventions are being implemented by governments or NGOs. In such contexts, reference groups may be exposed to a host of similar interventions to those being assessed, carrying the potential to understate intervention effects (Victora et al., 2010).

Intervention diffusion also has the potential to complicate effects. As opposed to secular change, which affects the population as a whole, diffusion happens when some individuals in the comparison or reference group are exposed to the specific intervention, through migration, or spillover of the intervention services between groups. Such effects are also called 'spillover effects' and 'externalities' in other disciplines. Spillover can be accounted for in the analysis, but it is better addressed in the design phase, ensuring proper site selection and sufficient distance between groups at the outset.

Synergies

As structural interventions often have multiple moving parts, the 'active ingredient' or 'mechanism of action' can be difficult to pin down. In addition, questions of potential synergies may also be important – where the effects of the whole (combined package) may be greater than the sum of the component parts.

There are a number of steps evaluators can take to address these challenges. They include: the use of detailed impact pathway monitoring; employing time-series data; the use of qualitative implementation research; and assessing the degree to which observed effects of combined programs compare to those predicted from evaluations of single interventions. These will be described in more detail in the section that follows.

EVALUATION OPTIONS FOR STRUCTUAL INTERVENTIONS

Robust evaluations of structural interventions generally require a mix of methods to respond to the challenges highlighted above (Bamberger et al., 2010, Leeuw and Vaessen, 2009). Decisions about *what* and *how* to evaluate must be made through careful consideration of *why* an evaluation is being done - with methods reflecting those most appropriate for the questions being posed (Habicht et al., 1999, Victora et al., 2004). In this section, we review a range of complementary methods in the menu of options for evaluating structural interventions, including impact assessment, implementation research and economic analysis.

IMPACT ASSESSMENT

Impact assessments answer questions about whether project goals were achieved and whether these effects were caused by the intervention. There are three main designs that may be suitable for evaluating structural interventions – adequacy, plausibility and probability (Victora et al., 1999).

Adequacy

An *adequacy* assessment compares the performance of the intervention package against a fixed set of goals or targets. The purpose is to demonstrate that the intervention is adequate to meet the designated objectives. The types of targets can be performance indicators such as whether clinics were opened or how many teachers have been trained in a school, or outcome-related indicators such as levels of intervention coverage and disease-specific endpoints. Adequacy assessments may also measure change over time but without any attempt to account for secular change.

Adequacy assessments generally lack a control group, and thus the main limitation is the inability of make statements regarding causality or attribution. For structural interventions, each point in the causal chain is important, as in the absence of adequate program performance, it is unrealistic to expect changes in health outcomes.

Plausibility

A *plausibility* assessment compares observed changes against a non-randomly selected reference group, as a way of accounting for the effects of secular change and confounding. This reference group may be a historical control group (a before and after assessment); or an external control group. The allocation of groups to receive an intervention in a plausibility assessment is generally non-random, therefore factors such as selection bias and confounding cannot be completely excluded.

Strategies to enhance plausibility

There are a number of strategies that can improve the case for attribution – whether exposure to an intervention has caused the observed changes. *Impact pathway mapping* involves applying a pre-defined and theoretically grounded impact pathway that systematically maps proposed mechanisms through which an intervention is expected to lead to changes in outcomes. In a complex and multi-component structural intervention, it is important to track indicators at each point in a causal chain. The size and consistency of changes across a range of activities and outcomes also helps enhance the plausibility that observed effects were a result of an

intervention, while helping to distinguish between interventions that are inherently faulty (failure of intervention concept or theory), those that were simply badly delivered (implementation failure), or those that were truly ineffective (an intervention that was well-conceived, well-implemented, but ultimately not effective) (Rychetnik et al., 2002).

Supporting evidence can also be derived from assessing a *dose-response relationship* which correlates the magnitude of exposure with the level of response. If better outcomes are associated with individuals or groups exposed to a greater intensity of the intervention, this supports the case for attribution. Instrumental variables analysis may be a useful analytical tool in this regard (Greenland, 2000). In this approach, the association between intervention dose (such as number of visits to a health clinic) and health outcomes is adjusted for by using an independent predictor of dosage (such as distance to a health clinic).

Changes observed can also be compared to *national or sub-national trends*. This provides a weaker reference point than a designated comparison group as external data may not be available on the exact outcomes of interest; national surveys may be conducted at different time points; and, they may be subject to uncertain and inconsistent enumerator and respondent bias. However, if intervention participants are outpacing both the designated comparison group as well as regional trends, this provides further support of intervention effectiveness.

Finally, if time series data is available on outcomes of interest – with multiple time points both before and after the start of an intervention – then an *interrupted time series design* can be employed. An 'inflection point' in the time series trend at the start of the intervention provides supportive evidence for the existence of a causal effect (Shadish et al., 2002).

Probability

A *probability* assessment employs randomization. For structural interventions, similar units are designated for participation in an evaluation, with the final selection of exposed and unexposed groups taking place using a random selection process. If a sufficient number similar of units are studied, this design minimizes the effect of *selection bias* (McKee et al., 1999).

When to employ a randomized design

One of the major challenges faced by evaluators of structural interventions is whether to opt for a randomized design. This choice is an important one. Evidence from randomized studies can be summarized in systematic reviews and used to accumulate a body of evidence for a particular intervention. Such information can shape global consensus, inform policy decisions, and influence resource allocation.

However, the random allocation of structural interventions may be difficult for a number of reasons. As described above, randomization may be perceived to be unethical or politically unfeasible. For example, a recent evaluation of a cash-transfer program noted the intervention could not be withheld from poor control groups for extended periods, as originally planned (IFPRI, 2002). Alternatively, intervention funding may have been pre-designated to particular groups based on non-random criteria. Second, for an intervention delivered to very large units (a media campaign in a city), random selection may be logistically complex, and having sufficient numbers of exposed and unexposed groups may not be possible (Mensah et al., 2010). Third, the time required for expected outcomes may be longer than is practical for informing decision-making (Bonell et al., 2006a). Fourth, randomization may not be necessary – such as where an intervention is expected to have a large and immediate effect that is unlikely to be explained by secular changes (UK Medical Research Council, 2006). Fifth, some structural interventions may be so complex, adaptive, or context-specific that the type of impact questions best-answered by randomized experiments may be less relevant. Here the most relevant question might not be "did it work?" but why and how did the various elements of an intervention influence the outcome, and are they likely to work in a similar way in another context? Finally, randomized studies may not be appropriate in settings where secular change is pronounced or in program-rich environments in which relatively untouched control groups are not possible.

In summary, these challenges should be carefully weighed when making choices regarding evaluation design. Randomized evaluations should be strongly considered for structural interventions under the following circumstances (Victora et al., 2010, Bonell et al., 2006a):

- when the aim is to assess the efficacy of unproven interventions;
- where interventions can be delivered in a consistent manner across treatment areas;
- where an intervention is reasonably 'discrete' and operates through pre-specified impact pathways;
- where a relatively large number of units can be randomized, and
- where relatively unexposed control groups can be maintained throughout the assessment period.

Alternative approaches: Stepped-wedge design and regression discontinuity

A variant of the randomized experiment is the *stepped wedge design*. The stepped wedge design may be used to overcome practical or ethical objections to experimental evaluations, especially for potentially effective interventions which cannot be made available to entire populations simultaneously (Brown and Liliford, 2006). This approach involves the advance random selection of intervention clusters, with implementation sequenced according to this random allocation. All clusters are assessed at project inception and at each point when a new group is enrolled, with late adopters serving as comparison groups for early initiators. Eventually, the whole population receives the intervention. The stepped-wedge design has been used in diverse settings including assessing the effect of cash transfer programs on health and nutrition outcomes in Mexico (IFPRI, 2002); examining how housing improvements effect respiratory ailments in England (Somerville et al., 2002), and evaluating the effects of water treatment on community health in South Africa (Bailey and Archer, 2004).

It is also worth noting that in non-health sectors, *regression discontinuity designs* have been employed as an equally rigorous option for evaluating program impacts (Imbens and Lemieux, 2008, Bloom, 2009). In this type of design, assignment to the intervention is based on a cut-off for some external criterion such as income; those above or below the cut-off receive the treatment, while others do not. In public health, for example, services could be offered to the poorest third of the families in a community. Though exposure to the program is not randomized, it is based on a measured criterion that can be modeled in the analysis. Under reasonable assumptions, the causality of this design is as strong as a randomized experiment.

The appeal of this design is that unlike random assignment, the neediest individuals or communities receive the intervention. The limitation of this approach, however, is that it requires 2-5 times the sample size of a randomized experiment.

IMPLEMENTATION RESEARCH

Implementation research, also known as process evaluation, is an essential component in the evaluation of structural interventions. While impact assessments respond to questions of whether an intervention works, implementation research is more exploratory in nature, asking *how* an intervention exerts its effects (Wight and Obasi, 2003). This is particularly important for structural interventions given their multifaceted nature and dependence on social context (Oakley et al., 2006). Such assessments are increasingly recommended for evaluations of complex interventions (Guba and Lincoln, 1989, Oakley et al., 2004), including randomized trials and interventions that are introduced at the level of populations or clusters (Hayes and Bennett, 1999, Campbell et al., 2000, Hawe et al., 2004, Bamberger et al., 2010, Leeuw and Vaessen, 2009).

The primary aim of implementation research is to examine how the process of implementation affects project outcomes (Koepsell et al., 1992, Rychetnik et al., 2002, Wight and Obasi, 2003, Shiell et al., 2008). This has implications for internal and external validity, addresses issues of equity and acceptability, and also reveals unintended consequences, both negative and positive. Despite the importance of process evaluation, especially of complex interventions, the creative use of qualitative methods alongside impact evaluation remains rare (Lewin et al., 2009).

Qualitative research

While the specific methods of implementation research are by nature contextual, a set of best-practice guidelines are emerging to guide implementation science. *Qualitative research* examines the perspectives of participants in the intervention – implementers, project partners and beneficiaries. This may involve key informant interviews with implementers to discuss the

theory and program design, community engagement, local adaptation, and barriers and facilitators to implementation. Interviews with government and other stakeholders document the alignment of the intervention within national policy and programs, and wider implications for integration and scale up. Focus groups with project beneficiaries help unpack the accessibility and acceptability of an intervention as well as generating insights about key ingredients of an intervention's failure or success – which are critical in complex adaptive systems, where seemingly small inputs or omissions can have disproportionate effects on uptake or impact (WHO, 2009).

To be most useful, qualitative investigations should be closely linked to implementation and quantitative assessments, with process evaluation investigators also being involved in the interpretation of the study findings. Too often qualitative research accompanying trials and other impact studies ends up being largely a parallel effort with little influence on study outcomes (Lewin et al., 2009). When reference groups are used, it is also important to use qualitative research to understand and document 'business as usual' or secular change, and the presence of other (similar or different) interventions in these comparison groups.

Performance monitoring

Performance monitoring examines program activities, outputs and outcomes along a range of thematic areas. The emphasis is on systematically documenting the timing, sequence, and uptake of various intervention components in the participating sites. Data are often generated from routinely collected source documents such as health facility forms, attendance or participation registers, or program management reports.

If performance data are collected on a regular basis, this yields time-series data that can be used to help understand when 'tipping or inflection points' took place – whether the introduction of specific investments or activities led to changes in levels of coverage or outcomes. This strategy is often used to enhance plausibility in observational data. Examples include the decline in hospital admissions due to childhood pneumonia that coincided with introducing

pneumococcal vaccination to the routine immunization schedule in USA (Grijalva et al., 2007), or reductions in child mortality that took place in Tanzania alongside increased government spending on efforts to expand coverage of child health interventions (Masanja et al., 2008) .

Context mapping

Investigating *contextual factors* that might influence the delivery, uptake or effects of structural interventions is another important dimension of a comprehensive evaluation. A deeper understanding of contextual drivers may help interpret the presence or absence of program effects. Factors such as policy change, shifting norms or major financial investments may drive wider secular changes and accelerate outcomes, while economic and political shocks may attenuate intervention affects. In addition, better understanding contextual issues may enhance the generalizability of findings (Hawe et al., 2004, Bonell et al., 2006b). For example, key cultural messages and a multi-component media campaign were felt to play a major role in stimulating policy change on domestic violence in South Africa (Usdin et al., 2005). While the content of the intervention may not have external validity beyond the immediate region, there may be processes and lessons around simultaneously engaging stakeholders at multiple levels to evoke policy shifts that are generalizable across a range of settings.

Methods for tracking contextual factors may be difficult to standardize, and may require detailed understandings of a range of secular forces that might influence intervention effects. A diverse set of methods can be employed to facilitate this process and include - but are not limited to - qualitative interviews with key stakeholders, monitoring media reports, tracking budgets and financial flows, satellite imagery or remote sensing. Tools that assess shifting norms, attitudes, levels of service delivery or the presence of external partnerships in comparison sites are also important. Systems for monitoring context are best initiated prospectively, and should document the timing, nature and scope of potential influences, in both intervention and control groups.

ECONOMIC EVALUATION

A final component to the evaluation of structural interventions is assessing economic returns on investment, which can be challenging for a number of reasons. Firstly, valuing the inputs for complex, multi-component structural interventions is often more difficult than for more discrete interventions. Secondly, and most importantly, structural interventions may generate a wide range of health and non-health benefits or outputs, which may extend well beyond the time-frame of the intervention, and may be poorly captured using conventional health economics methods such as cost effectiveness analyses (Jamison et al., 2006).

Cost Effectiveness Analyses (CEA) are used in health economics to compare 'usual practice' with an intervention under consideration. These outcomes use standard measures such as cases prevented, lives saved or life years saved, as well as composite measures such as Disability Adjusted Life Years (DALYs) or Quality Adjusted Life Years (QALYs). CEAs work best when one can specify a health intervention in some detail – such as the dose, frequency and duration of vitamin A, who will be delivering it, and the level at which it will be delivered – as all have associated costs. These costs must then be paired against measurable health endpoints, such as reductions in disease rates or mortality. CEA works less well for interventions that have a range of benefits that may be difficult to quantify in terms of health endpoints.

An alternative approach is cost-benefit analysis (CBA), which involves appraising a programme in terms of costs and benefits to society, with benefits measured in monetary units (Mishan, 1971). This approach may be more appropriate for interventions with multiple effects and provide a basis for valuing non-health benefits as well as 'spill-overs' – where benefits extend beyond programme participants (Mooney, 1994, Ryan, 1995). This is particularly true in relation to changes in the social environment - including improvements in social and economic well-being - which are not captured as immediate health benefits but may nevertheless be important (McKinlay, 1993).

There is growing interest in assessing the value of more diffuse outcomes within the health economics literature, with the community rather than the individual as the unit of interest (Shiell and Hawe, 1996, Jan, 1998). One of the methodological tools used is a *willingness-to-pay (WTP) assessment* (Mishan, 1971), which is based on the assumption that a consumer is the best

judge of the value of the goods and services they consume. There are some concerns about using WTP in the health sector, particularly because willingness to pay tends to be directly related to ability to pay and thus this approach gives lower values to goods and services consumed by people of lower income. Although there are examples in the literature where a WTP approach has been successfully applied in low income settings (Bhatia and Fox-Rushby, 2002, Onwujekwe et al., 2000, Onwujekwe et al., 1998) its validity amongst very poor communities, particularly those which rely heavily on say bartering or subsistence farming, has yet to be firmly established. Further methodological innovation in valuing the costs and returns of structural interventions remains a major priority.

CONCLUSIONS

While there are a number of challenges to evaluating structural interventions, high quality evaluation is vital. Too often, structural interventions have been under-developed and under-researched as their rigorous evaluation seems beyond the scope of established tools and methods. While discrete technical interventions may seem more easily engaged scientifically, addressing complex public health challenges will likely require broader approaches that extend the useful limits of conventional evaluation models.

This chapter has presented a range of complimentary methods to assist the design of evaluations for structural interventions in public health. We have described approaches to impact assessment, implementation research and economic evaluation. Many challenges can be overcome through adequate planning, setting realistic evaluation objectives, optimizing the design, defining clear impact pathways, using an appropriate mix of methods, careful monitoring of secular trends, and drawing upon perspectives and expertise from partners outside the health sector. Taken together, these approaches may foster new insights and a deeper understanding the role of structural drivers in shaping health outcomes.

CASE STUDIES

1. Multi-Country Evaluation of Integrated Management of Childhood Illness (MCE-IMCI)

Intervention overview

In the mid-1990's, WHO and UNICEF launched the Integrated Management of Childhood Illness (IMCI) strategy to improve health and development of children under five, by targeting diarrhoeal disease, pneumonia, malaria and malnutrition, which together accounted for approximately 70% of global under-five deaths. The IMCI strategy includes a combination of structural interventions addressing improvements in case-management; improvements in health systems; and improvements in family and community practices (Arifeen et al., 2009, Bryce et al., 2004).

Evaluation design

Impact assessment: While the individual IMCI interventions had proven efficacy in controlled settings, there was a need to evaluate program effectiveness in 'real-world' settings. The multi-country evaluation (MCE) ran for seven years including feasibility studies in 12 countries and in-depth evaluations in five countries. The impact assessment combined adequacy and plausibility evaluations using non-randomized comparison groups in Tanzania, Peru, Brazil, and Uganda, and a probability evaluation using a cluster-randomized design in Bangladesh (Bryce et al., 2004). The intervention operated in a program-rich environment, with many IMCI components already being scaled nation-wide, making overall interpretation of the results difficult.

Implementation research: Demonstrating the adequacy of the provision, utilization and coverage indicators was deemed essential for interpreting impact results. The evaluators conducted detailed impact pathway mapping, employing a range of implementation research tools. Although no effect on mortality was seen within the timeframe of the Bangladesh study,

positive changes were seen in all input, output and outcome indicators (Arifeen et al., 2009). Conversely, in Peru no associations between IMCI and outpatient utilization, vaccine coverage, mortality or malnutrition were demonstrated and implementation research showed inadequate health systems support for IMCI and low uptake of IMCI training by health-workers (Huicho et al., 2005).

Economic costing: All five in-depth evaluations also included cost effectiveness studies (Bishai et al., 2008, Armstrong Schellenberg et al., 2004). For a more detailed discussion on the methodological lessons learned through design and implementation of the evaluation see Bryce at al. (2004) and Bryce and Victoria (2005).

2. Avahan – The India AIDS Initiative

Intervention overview

Avahan, The India AIDS Initiative is a 10-year intervention that commenced in 2003 with the goal of preventing the expansion of the HIV epidemic in India. The intervention operated in the six states with the majority of HIV cases and focused on high-risk groups such as female sex workers and their clients, men who have sex with men, and injecting drug users (The Gates Foundation, 2008). The initiative addressed downstream risk factors such as STIs and condom use, while simultaneously addressing more upstream risk factors including stigma, violence, legal, environmental, and health infrastructure factors. The initiative was a complex, multi-layered structural intervention that was adapted to suit local needs, involving peer-support activities alongside work with the police, journalists, and government officials.

Evaluation design

Impact assessment: To monitor intervention effects on violence (Beattie et al., 2010), condom-use (Lowndes et al., 2010, Ramesh et al., 2010), and HIV and STI prevalence (Ramesh et al., 2010), evaluators employed primarily an adequacy assessment – conducting cross sectional

surveys among female sex workers (a pre-post design using historical controls). In addition, several strategies were used to enhance plausibility:

- Using temporal data linking observed changes to the timing of program activities (Lowndes et al., 2010)
- Associating the duration of exposure to shifts in key outcomes (Ramesh et al., 2010)
- Documenting the adequacy (whether goals were achieved) of each step along the impact pathway (Chandrasekaran et al., 2008).

Because Avahan focused on high-risk individuals in high-prevalence districts, and because of the potential for diffusion and overlap with existing interventions, it was deemed not practical or ethical to employ randomization or matched comparison groups (Chandrasekaran et al., 2008). Rapid scale-up was considered essential to containing the spread of the epidemic, so a stepped-wedge design was also not possible.

Implementation research: The evaluation framework addressed each step along the impact pathway. Routine program monitoring indicators were collected to assess provision, utilization, coverage and quality of services, before considering questions of impact. Daily tracking of news articles was used to assess levels of stigma and the effects of the community component of the intervention. The use of data-driven management tools to modify and improve the program is considered one of the reasons why successful rapid scale-up has been achievable (Bertozzi et al., 2010).

3. The Millennium Villages Project

Intervention overview

The Millennium Villages (MV) project is a ten year initiative that delivers a package of scientifically-proven interventions with the central aim of achieving the MDGs across diverse sub-Saharan African field sites (Sanchez et al., 2007, The-Earth-Institute and Millennium-Promise, 2010). The project operates in rural areas of the continent where MDG-related progress has been insufficient. Activities are coordinated across multiple sectors (including health, agriculture, environment, business development, education and infrastructure) and adapted to local needs, systems challenges and disease profiles. Inputs are cost-limited, with a modest annual ceiling of $120 per capita across all sectors sustained over a ten year period.

Evaluation design

Impact assessment: To maximize external validity, rural clusters of approximately 40,000 people with high rates of poverty and undernutrition were purposively selected from ten countries to represent over 90% of the agro-ecological zones in sub-Saharan Africa. To measure the adequacy of the intervention in relation to effects on MDG-related outcomes, assessment rounds are conducted at two year intervals. To account for secular change and enhance the plausibility that observed changes are the result of intervention exposure, a controlled design with comparison clusters selected at random from among matched candidates. Finally, national and sub-national MDG outcomes provide an additional reference group. Initial results highlight positive synergies on critical nutrition-related outcomes (Remans et al., 2011).

Implementation research: Performance metrics across all sectors are collected on a monthly to quarterly basis using health facility and school registers, infrastructure mapping, and other available sources of information. Qualitative research includes focus groups and key-informant interviews with intervention recipients, program managers, and key partners on the ground. Detailed intervention timelines are constructed for activities across all sectors. Context mapping takes place in parallel to survey rounds using purpose-built inventories of community-level characteristics such as shocks, changes in service delivery, policy changes and so forth.

Economic costing: An economic assessment examines year-on-year MDG-related spending by sector and by stakeholder. These data will be used to estimate returns on investment, and cost-effectiveness in relation to impacts observed.

4. The Intervention with Microfinance for AIDS & Gender Equity (IMAGE Study)

Intervention overview

IMAGE was a structural intervention for the prevention of HIV and intimate partner violence (IPV). The initiative brought together two components: a poverty focused microfinance program, and a gender and HIV training curriculum. A major focus of the training program was

community mobilization and collective action, which was by nature fluid and adapted to local needs and priorities (Pronyk et al., 2006, Pronyk et al., 2008b).

Evaluation design

Impact assessment: A cluster randomized trial was undertaken using a matched-pair design to examine village level effects. As the total number of clusters was low (n=8), the evaluators considered this a plausibility assessment. Detailed impact pathways examined effects on a range of pathway variables, including dimensions such as economic well-being, empowerment, and social capital, where the size and consistency of changes were profiled in the reporting of results (Kim et al., 2007, Pronyk et al., 2008a). In an attempt to assess the relative effects of the two intervention components, a separate sub-study was conducted among women who received microfinance alone (Kim et al., 2009).

Implementation research: Routine records of microfinance loan performance and attendance at training sessions examined feasibility and exposure to the intervention. Gender trainers employed diaries and 'key-events time lines' to document the narratives of individual loan centers and systematically document community mobilization efforts. Qualitative interviews with program managers, participants, local partners and policy makers examined barriers and facilitators to implementation, as well as opportunities for replication and scale-up (Hargreaves et al., 2010).

Economic costing: A full cost-effectiveness study was done to examine the costs relative to intervention effects on disability-adjusted life years saved from observed reductions in levels of IPV (Jan et al., 2010).

REFERENCE LIST

ARIFEEN, S. E., HOQUE, D. M. E., AKTER, T., RAHMAN, M., HOQUE, M. E., BEGUM, K., CHOWDHURY, E. K., KHAN, R., BLUM, L. S., AHMED, S., HOSSAIN, M. A., SIDDIK, A., BEGUM, N., RAHMAN, Q. S. U., HAQUE, T. M., BILLAH, S. M., ISLAM, M., RUMI, R. A., LAW, E., AL-HELAL, Z. A. M., BAQUI, A. H., SCHELLENBERG, J., ADAM, T., MOULTON, L. H., HABICHT, J. P., SCHERPBIER, R. W., VICTORA, C. G., BRYCE, J. & BLACK, R. E. 2009. Effect of the Integrated Management of Childhood Illness strategy on childhood mortality and nutrition in a rural area in Bangladesh: a cluster randomised trial. *Lancet*, 374**,** 393-403.

ARMSTRONG SCHELLENBERG, J. R. M., ADAM, T., MSHINDA, H., MASANJA, H., KABADI, G., MUKASA, O., JOHN, T. J., CHARLES, S., NATHAN, R., WILCZYNSKA, K., MGALULA, L., MBUYA, C., MSWIA, R., MANZI, F., DE SAVIGNY, D., SCHELLENBERG, D. & VICTORA, C. G. 2004. Effectiveness and cost of facility-based Integrated Management of Childhood Illness (IMCI) in Tanzania. *The Lancet*, 364**,** 1583-1594.

BAILEY, I. W. & ARCHER, L. 2004. The impact of the introduction of treated water on aspects of community health in a rural community in Kwazulu-Natal, South Africa. *Water Science and Technology*, 50**,** 105-110.

BAMBERGER, M., RAO, V. & WOOLCOCK, M. 2010. Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development. *In:* TEAM, T. W. B. D. R. G.-P. A. I. (ed.). The World Bank - Policy Research Working Papers.

BEATTIE, T. S., BHATTACHARJEE, P., RAMESH, B. M., GURNANI, V., ANTHONY, J., ISAC, S., MOHAN, H. L., RAMAKRISHNAN, A., WHEELER, T., BRADLEY, J., BLANCHARD, J. F. & MOSES, S. 2010. Violence against female sex workers in Karnataka state, south India: impact on health, and reductions in violence following an intervention program. *BMC Public Health*, 10**,** 476.

BERTOZZI, S. M., PADIAN, N. & MARTZ, T. E. 2010. Evaluation of HIV prevention programmes: the case of Avahan. *Sex Transm Infect*, 86**,** I4-I5.

BHATIA, M. & FOX-RUSHBY, J. 2002. Willingness to pay for treated mosquito nets in Surat, India: the design and descriptive analysis of a household survey. *Health Policy and Planning*, 17**,** 402-411.

BISHAI, D., MIRCHANDANI, G., PARIYO, G., BURNHAM, G. & BLACK, R. 2008. The cost of quality improvements due to integrated management of childhood illness (IMCI) in Uganda. *Health Economics*, 17**,** 5-19.

BLANKENSHIP, K. M., BRAY, S. J. & MERSON, M. H. 2000. Structural interventions in public health. *Aids*, 14**,** S11-S21.

BLOOM, H. 2009. Modern Regression Discontinuity Analysis. *MDRC Working Papers on Research Methodology*.

BONELL, C., HARGREAVES, J., STRANGE, V., PRONYK, P. & PORTER, J. 2006a. Should structural interventions be evaluated using RCTs? The case of HIV prevention. *Soc Sci Med*, 63**,** 1135-42.

BONELL, C., OAKLEY, A., HARGREAVES, J., STRANGE, V. & REES, R. 2006b. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *British Medical Journal*, 333**,** 346-349.

BROWN, C. A. & LILIFORD, R. J. 2006. The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, 6**,** doi:10.1186/1471-2288-6-54.

BRYCE, J., VICTORA, C. G., HABICHT, J. P., VAUGHAN, J. P. & BLACK, R. E. 2004. The multi-country evaluation of the integrated management of childhood illness strategy: Lessons for the evaluation of public health interventions. *American Journal of Public Health*, 94**,** 406-415.

CAMPBELL, M., FITZPATRICK, R., HAINES, A., KINMONTH, A. L., SANDERCOCK, P., SPIEGELHALTER, D. & TYRER, P. 2000. Framework for design and evaluation of complex interventions to improve health. *British Medical Journal*, 321**,** 694-696.

CHANDRASEKARAN, P., DALLABETTA, G., LOO, V., MILLS, S., SAIDEL, T., ADHIKARY, R., ALARY, M., LOWNDES, C. M., BOILY, M. C., MOORE, J. & PARTNERS, A. E. 2008. Evaluation design for large-scale HIV prevention programmes: the case of Avahan, the India AIDS initiative. *AIDS*, 22**,** S1-S15.

GREENLAND, S. 2000. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*, 29**,** 722-729.

GRIJALVA, C. G., NUORTI, J. P., ARBOGAST, P. G., MARTIN, S. W., EDWARDS, K. M. & GRIFFIN, M. R. 2007. Decline in pneumonia admissions after routine childhood immunisation with pneumococcal conjugate vaccine in the USA: a time-series analysis. *The Lancet*, 369**,** 1179-1186.

GUBA, E. G. & LINCOLN, Y. S. 1989. *Fourth generation evaluation.*, Newbury Park, CA: Sage, Sage.

HABICHT, J., VICTORA, C. & VAUGHAN, J. 1999. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *Int J Epidemiol*, 28**,** 10-18.

HARGREAVES, J. R., HATCHER, A., STRANGE, V., PHETLA, G., BUSZA, J., KIM, J. C., WATTS, C., MORISON, L. A., PORTER, J. D. H., PRONYK, P. M. & BONNEL, C. 2010. Process evaluation of the Intervention with Microfinance for AIDS and Gender Equity (IMAGE) in rural South Africa. *Health Education Research*, 25**,** 27-40.

HAWE, P., SHIELL, A. & RILEY, T. 2004. Complex interventions: how "out of control" can a randomised controlled trial be. *BMJ*, 328**,** 1561-1563.

HAYES, R. D., ALEXANDER, N. D. E., BENNETT, S. & COUSENS, S. N. 2000. Design and analyisis issues in cluster-randomized trials of interventions against infectious diseases. *Statistical Methods in Medical Research*, 9**,** 95-116.

HAYES, R. J. & BENNETT, S. 1999. Simple sample size calculation for cluster randomised trials. *International Journal of Epidemiology*, 28**,** 319-326.

HUICHO, L., DAVILA, M., GONZALES, F., DRASBEK, C., BRYCE, J. & VICTORA, C. G. 2005. Implementation of the Integrated Management of Childhood Illness strategy in Peru and its association with health indicators: an ecological analysis. *Health Policy Plan*, 20 Suppl 1**,** i32-i41.

IFPRI 2002. PROGRESA: breaking the cycle of poverty. Washington, D.C.: International Food Policy Research Institute.

IMBENS, G. & LEMIEUX, T. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142**,** 615-635.

JAMISON, D. T., BREMAN, J. G., MEASHAM, A. R., ALLEYNE, G., CLAESON, M., EVANS, D. B., JHA, P., MILLS, A. & MUSGROVE, P. 2006. Priorities in health: disease control priorities project. Washington: World Bank.

JAN, S. 1998. A holistic approach to the economic evaluation of health programs using institutionalist methodology. *Social Science and Medicine*, 47**,** 1565-1572.

JAN, S., FERRARI, G., WATTS, C. H., HARGREAVES, J. R., KIM, J. C., MORISON, L. A., PHETLA, G., PORTER, J. D. H. & PRONYK, P. M. 2010. Economic evaluation of a combined microfinance and gender training intervention for the prevention of intimate partner violence in rural South Africa. *Health Policy and Planning*, PMID: 20974751.

KIM, J. C., FERRARI, G., ABRAMSKY, T., WATTS, C. H., HARGREAVES, J. R., MORISON, L. A., PHETLA, G., PORTER, J. D. H. & PRONYK, P. M. 2009. Assessing the incremental benefits of combining health and economic interventions: Experience from the IMAGE Study in rural South Africa. *Bulletin of the WHO*, 87**,** 824-832.

KIM, J. C., WATTS, C. H., HARGREAVES, J. R., MORISON, L. A., PORTER, J. D. H., PHETLA, G., BUZSA, J., NDHLOVU, L. & PRONYK, P. M. 2007. Understanding the impact of a microfinance-based intervention on women's empowerment and the reduction of intimate partner violence in the IMAGE Study, South Africa. *American Journal of Public Health*, 97**,** 1794-1802.

KOEPSELL, T. D., WAGNER, E. H., CHEADLE, A. C., PATRICK, D. L., MARTIN, D. C., DIEHR, P. H. & AL., E. 1992. Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Annual Review of Public Health*, 13, 31-57.

LEEUW, F. & VAESSEN, J. 2009. Impact Evaluations and Development: NONIE guidance in impact evaluation. Network of Networks for Impact Evaluation.

LEWIN, S., GLENTON, C. & OXMAN, A. D. 2009. Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. *British Medical Journal*, 339.

LOWNDES, C. M., ALARY, M., VERMA, S., DEMERS, E., BRADLEY, J., JAYACHANDRAN, A. A., RAMESH, B. M., MOSES, S., ADHIKARY, R. & MAINKAR, M. K. 2010. Assessment of intervention outcome in the absence of baseline data: 'reconstruction' of condom use time trends using retrospective analysis of survey data. *Sex Transm Infect*, 86, I49-I55.

MASANJA, H., DE SAVIGNY, P., SCHELLENBERG, J., JOHN, T., MBUYA, C., UPUNDA, G., BOERMA, T., VICTORA, C. G., SMITH, T. & MSHINDA, H. 2008. Child survival gains in Tanzania: analysis of data from demographic and health surveys. *The Lancet*, 371, 1276-1283.

MCKEE, M., BRITTON, A., BLACK, N., MCPHERSON, K., SANDERSON, C. & BAIN, C. 1999. Interpreting the evidence: choosing between randomised and non-randomised studies. *British Medical Journal*, 319, 312-315.

MCKINLAY, J. B. 1993. The promotion of health through planned sociopolitical chance. *Social Science and Medicine*, 36, 109-117.

MENSAH, J., OPPONG, J. R. & SCHMIDT, C. M. 2010. Ghana's National Health Insurance Scheme in the Context of the Health Mdgs: An Empirical Evaluation Using Propensity Score Matching. *Health Economics*, 19, 95-106.

MISHAN, E. J. 1971. Evaluation of life and limb: a theoretical approach. *Journal of Political Economy*, 79, 687-706.

MOONEY, G. 1994. What else do we want from our health services? *Social Science and Medicine*, 39, 151-154.

OAKLEY, A., STRANGE, V., BONELL, C., ALLEN, E. & STEPHENSON, J. 2006. Process evaluation in randomised controlled trials of complex interventions. *British Medical Journal*, 332, 413-416.

OAKLEY, A., STRANGE, V., STEPHENSON, J. M., FORREST, S. & MONTEIRO, H. 2004. Evaluating processes: an example from a randomised controlled trial of sex education: rationale and methods. *Evaluation*.

ONWUJEKWE, O. E., SHU, E., CHIMA, R., ONYIDO, A. & OKONKWO, P. O. 2000. Willingness to pay for the retreatment of mosquito nets with insecticide in four communities of south-eastern Nigeria. *Tropical Medicine and International Health*, 5, 370-376.

ONWUJEKWE, O. E., SHU, E. N., NWAGBO, D., AKPALA, C. O. & OKONKWO, P. O. 1998. Willingness to pay for community-based ivermectin distribution: a study of three onchocerciasis-endemic communities in Nigeria. *Tropical Medicine and International Health*, 3, 802-808.

PLUMMER, M. L., WIGHT, D., OBASI, A. I. N., WAMOYI, J., MSHANA, G., TODD, J., MAZIGE, B. C., MAKOKHA, A., HAYES, R. J. & ROSS, D. A. 2007a. A process evaluation of a school-based adolescent sexual health intervention in rural Tanzania: the MEMA kwa Vijana programme. *Health Education Research*, 22, 500-512.

PLUMMER, M. L., WIGHT, D., WAMOYI, J., NYALALI, K., INGAL, T., MSHANA, G., SHIGONG, Z. S., OBASI, A. I. N. & ROSS, D. A. 2007b. Are schools a good setting for adolescent sexual health promotion in rural Africa? A qualitative assessment from Tanzania. *Health Education Research*, 22, 483-499.

PRONYK, P. M., HARGREAVES, J. R., KIM, J. C., MORISON, L. A., PHETLA, G., WATTS, C., BUSZA, J. A. & PORTER, J. D. H. 2006. Effect of a structural intervention for the prevention of intimate partner violence and HIV in rural South Africa: a cluster randomized trial. *The Lancet*, 368, 1973-1983.

PRONYK, P. M., HARPHAM, T., BUSZA, J., PHETLA, G., MORISON, L. A., HARGREAVES, J. R., KIM, J. C., WATTS, C. H. & PORTER, J. D. H. 2008a. Can social capital be intentionally generated? A randomized trial from rural South Africa. *Social Science and Medicine,* in press.

PRONYK, P. M., KIM, J. C., ABRAMSKY, T., PHETLA, G., HARGREAVES, J. R., MORISON, L. A., WATTS, C. H., BUSZA, J. & PORTER, J. D. H. 2008b. A combined microfinance and training intervention can reduce HIV risk behaviour among young female participants. *AIDS,* 22**,** 1659-1665.

RAMESH, B. M., BEATTIE, T. S., SHAJY, I., WASHINGTON, R., JAGANNATHAN, L., REZA-PAUL, S., BLANCHARD, J. F. & MOSES, S. 2010. Changes in risk behaviours and prevalence of sexually transmitted infections following HIV preventive interventions among female sex workers in five districts in Karnataka state, south India. *Sex Transm Infect,* 86 Suppl 1**,** i117-24.

REMANS, R., PRONYK, P. M., FANZO, J., CHEN, J., PALM, C. A., NEMSER, B., MUNIZ, M., RADUNSKY, A., ABAY, A. H., COULIBALY, M., MENSAH-HOMIAH, J., WAGAH, M., AN, X., MWAURA, C., QUINTANA, E., SOMERS, M., SANCHEZ, P. A., MCARTHUR, J. W., SACHS, S. E. & SACHS, J. D. 2011. A multi-sector intervention to accelerate reductions in child stunting: an observational study from nine sub-Saharan African countries. *American Journal of Clinical Nutrition,* doi: 10.3945/ajcn.111.020099.

ROSE, G. 1985. Sick individuals and sick populations. *International Journal or Epidemiology,* 14**,** 32-38.

ROSS, D. A., CHANGALUCHA, J., OBASI, A. I. N., TODD, J., PLUMMER, M. L., CLEOPHAS-MAZIGE, B., ANEMONA, A., EVERETT, D., WEISS, H. A., MABEY, D. C., GROSSKURTH, H., HAYES, R. J., BALIRA, R., WIGHT, D., GAVYOLE, A., MAKOKHA, M. J., MOSHA, F., TERRIS-PRESTHOLT, F. & PARRY, J. V. 2007. Biological and behavioural impact of an adolescent sexual health intervention in Tanzania: a community-randomized trial. *Aids,* 21**,** 1943-1955.

RYAN, M. 1995. *Economics and the Patient's Utility Function: An Application to Assisted Reproductive Techniques. Ph.D Thesis.* University of Aberdeen.

RYCHETNIK, L., FROMMER, M., HAWE, P. & SHIELL, A. 2002. Criteria for evaluating evidence on public health interventions. *J Epidemiol Community Health,* 56**,** 119-127.

SANCHEZ, P., PALM, C., SACHS, J., DENNING, G., FLOR, R., HARAWA, R., JAMA, B., KIFLEMARIAM, T., KONECKY, B., KOZAR, R., LELERAI, E., MALIK, A., MODI, V., MUTUO, P., NIANG, A., OKOTH, H., PLACE, F., SACHS, S. E., SAID, A., SIRIRI, D., TEKLEHAIMANOT, A., WANG, K., WANGILA, J. & ZAMBA, C. 2007. The African Millennium Villages. *Proceedings of the National Academy of Sciences,* 104**,** 16775-16780.

SAVEDOFF, W., R., L. & BIRDSALL, N. 2006. When will we ever learn? Improving lives through impact evaluation. *Evaluation Gap Working Group.* Washington, DC: Centre for Global Development.

SHADISH, W. R., COOK, T. D. & CAMPBELL, D. T. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference,* Boston, MA, Houghton Mifflin and Company.

SHEPPERD, S., LEWIN, S., STRAUS, S., CLARKE, M., ECCLES, M. P., FITZPATRICK, R., WONG, G. & SHEIKH, A. 2009. Can We Systematically Review Studies That Evaluate Complex Interventions? *PLoS Med,* 6.

SHIELL, A. & HAWE, P. 1996. Health promotion, community development and the tyranny of individualism. *Health Economics,* 5**,** 241-247.

SHIELL, A., HAWE, P. & GOLD, L. 2008. Complex interventions or complex systems?Implications for health economic evaluation. *British Medical Journal,* 336**,** 1281-1283.

SOMERVILLE, M., BASHAM, M., FOY, C., BALLINGER, G., GAY, T., SHUTE, P. & BARTON, A. G. 2002. From local concern to randomised trial: the Watcombe Housing Project. *Health Expectations,* 5**,** 127-135.

SUSSER, M. 1995. The Tribulations of Trials- Intervention in Communities. *American Journal of Public Health,* 85**,** 156-159.

THE EARTH INSTITUTE & MILLENNIUM PROMISE 2010. Harvests of Development in rural Africa: the Millennium Villages after three years. New York: The Earth Institute, Columbia University.

THE GATES FOUNDATION 2008. Avahan—The India AIDS Initiative: The business of HIV prevention at scale. New Delhi, India: Bill and Melinda Gates Foundation.

UK MEDICAL RESEARCH COUNCIL 2006. Developing and evaluating complex interventions: new guidance. London: UK-MRC.

USDIN, S., SCHEEPERS, E., GOLDSTEIN, S. & JAPHET, G. 2005. Achieving social change on gender-based violence: A report on the impact evaluation of Soul City's fourth series. *Social Science and Medicine*, 61, 2434-2445.

VICTORA, C. G., BLACK, R. E., BOERMA, J. T. & BRYCE, J. 2010. Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations. *Lancet*, July 9. 2010.

VICTORA, C. G., HABICHT, J. P. & BRYCE, J. 2004. Evidence-based public health: Moving beyond randomized trials. *American Journal of Public Health*, 94, 400-405.

VICTORA, C. G., HABICHT, J. P. & VAUGHAN, J. P. 1999. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *Int J Epidemiol*, 28, 10-18.

WHO 2009. Systems thinking for helath systems strengthening. Geneva: Alliance for health policy and systems research: World Health Organization.

WIGHT, D. & OBASI, A. 2003. Unpacking the 'black box': the importance of process data to explain outcomes. *In:* STEPHENSON, J., IMRIE, J. & BONELL, C. (eds.) *Effective Sexual Health Interventions: issues in experimental evaluation.* Oxford: Oxford University Press.