WHO Technical Report Series
938

PREVENTING HIV/AIDS IN YOUNG PEOPLE

A SYSTEMATIC REVIEW OF THE EVIDENCE FROM DEVELOPING COUNTRIES

UNAIDS Inter-agency Task Team on Young People

4. The weight of evidence: a method for assessing the strength of evidence on the effectiveness of HIV prevention interventions among young people

David A Ross,[a] Danny Wight,[b] Gary Dowsett,[c] Anne Buvé,[d] & Angela I N Obasi[e]

Objectives To design a method for assessing the strength of evidence on the effectiveness of different interventions to prevent the spread of HIV that will be the basis for the reviews in this series.

Methods The literature on the evaluation of public health interventions was reviewed, and a method was developed in consultation with colleagues involved in this series of reviews and others.

Findings The method involves the following steps. First, define the key types of intervention that policy-makers need to choose between in the population setting under consideration. Second, define the strength of evidence that would be needed to justify widespread implementation of the intervention. Third, develop explicit inclusion and exclusion criteria for the studies under review. Fourth, critically review all eligible studies and their findings, by intervention type. Fifth, summarize the strength of the evidence on the effectiveness of each type of intervention. Sixth, compare the strength of the evidence provided by the studies against the threshold of evidence that would be needed to recommend widespread implementation. Seventh, from this comparison, derive evidence-based recommendations related to the implementation of each type of intervention in the setting or population group.

[a] Infectious Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, England. Correspondence should be sent to Dr Ross (email: david.ross@lshtm.ac.uk).
[b] MRC Social and Public Health Sciences Research Unit, Glasgow, Scotland.
[c] Australian Research Centre in Sex, Health and Society, La Trobe University, Melbourne, Australia.
[d] STD/HIV Research and Intervention Unit, Institute of Tropical Medicine, Antwerp, Belgium.
[e] Liverpool School of Tropical Medicine, Liverpool, England.

Conclusions The method proposed here provides a systematic, rigorous and transparent approach to reviewing evidence on the effectiveness of interventions of different types and in different population settings in order to generate recommendations for policy-makers.

4.1 Introduction

The AIDS epidemic is a major public health emergency, and young people are bearing the main brunt of new infections worldwide. There is an urgent need to work towards a consensus on what should be done in order to meet the internationally accepted goals for the prevention of HIV among young people that were defined at the United Nations General Assembly Special Session on HIV/AIDS (UNGASS) in 2001 (*1*). These global goals are presented and discussed in chapter 1, but in sum they give specific targets for improving access to information, skills and services; reducing vulnerability; and reducing HIV prevalence.

In an area as important as preventing the spread of HIV among young people in developing countries, difficult choices have to be made by policy-makers and programme developers irrespective of whether the evidence that is available to guide these decisions is weak or strong. Although evidence on the effectiveness of interventions will be only one of the factors that policy-makers use when deciding in which programmes to invest, a systematic review of the evidence related to all the options will be more useful to them than piecemeal reviews using different criteria and weights for different types of evidence. As in most areas of social policy, gaining consensus on the relative weights that should be given to different types of evidence has been difficult, but for policy decisions to be rational and transparent, reaching such consensus is crucially important. Furthermore, the lack of any explicit policy or programme is in fact a policy decision. And, finally, because of the complexity of the interventions, the evidence for and against any intervention strategy is likely also to be complex, requiring the synthesis of multiple types of evidence of varying quality and weight. Evaluation researchers should provide evidence that is as valid as possible to policy-makers and ensure that it is synthesized and presented in a way that will make it relevant, accessible and easy to interpret and act on.

While recognizing that there are major obstacles to rational evidence-based decision-making in this field, this chapter aims to indicate a way forward by presenting a structure within which researchers, advisers and policy-makers can assess the strength of the evidence for each of the interventions discussed in subsequent chapters in this series. In this chapter we are concerned with the broad principles involved in assessing the evidence. Later chapters in the series will address how these principles apply to specific interventions. The

criteria that we propose for assessing the evidence draw on recent debates on the relative merits and limitations of randomized controlled trials (*2–7*), and suggestions for approaches to the evaluation of evidence on public health interventions (*6, 8–10*), and to the presentation and systematic review of study results (*11–15*).

4.2 Types of interventions and evidence on effectiveness

Most of the programmes that have been introduced or advocated to reduce the prevalence of HIV among adolescents are complex, often comprising combinations of components, such as:

- in-school teacher-led sex and/or life skills education;
- in-school peer education and/or mentoring or counselling;
- specific interventions (such as peer education) for out-of-school youths (including those who would be expected to be in school but are not), for specific groups of youths (for example, groups affiliated with religious organizations) and for groups at high risk of HIV (such as intravenous drug users, commercial sex workers or men who have sex with men);
- condom promotion and improved access to condoms (for example, through social marketing, health-worker training, providing supplies or reorganizing clinical services);
- youth-friendly health services;
- access to counselling and voluntary HIV testing;
- access to care, support and treatment for people who are HIV positive;
- community development approaches to modifying sexual and social norms;
- mass media approaches to changing social values, norms and behaviours;
- legislative changes.

These components may be targeted at different levels, including the individual (for example, by providing life skills training), the family (for example, by improving intrafamily communication about sexuality) and the community (for example, by providing access to youth-friendly health services, mass media campaigns aimed at changing norms in society regarding gender roles or interventions directed towards men to decrease girls' vulnerability). Furthermore, many of these specific components are, in themselves, complex interventions. To be effective, most would involve bringing about profound social and behavioural changes among both the implementers (for example, who

must respect confidences and understand and empathize with young people's concerns) and the potential target groups; however, evaluating such outcomes is notoriously difficult. Finally, when a programme is made up of several interventions, often with a different emphasis given to each, it is difficult to assess the relative effectiveness of each component.

The key policy questions are:

- which interventions should be selected?
- in which contexts are they appropriate?
- what proportion of the available resources should be allocated to each?

The complexity of the interventions and the inadequacy of evaluations of them mean that policy decisions will often need to be based only on partial or imperfect evidence. Some of the reasons for this imperfect evidence are summarised in Box 4.1.

Box 4.1

Obstacles to obtaining perfect evidence

HIV prevention interventions are complex

- There are numerous interventions and strategies to choose from.
- The content and quality of interventions may differ substantially from one another, and interventions may be implemented in different ways by different people. For example, two life-skills programmes in secondary schools that have different content and theoretical bases and are delivered in different ways are likely to have different impacts.
- The interventions needed to address the five UNGASS goals (see chapter 1) will be social interventions of varying and often substantial degrees of complexity. This will necessarily complicate their evaluation.
    - The mechanisms by which these strategies are meant to work are diverse, complex and poorly understood. In contrast to the biological mechanisms by which therapeutic drugs work, there is far less consensus on the workings of the social world in and through which behavioural interventions operate (*16*).
    - Lack of understanding of the mechanisms raises the added problem that purported intermediate outcomes may or may not be valid. For example, an increase in the skills needed to avoid HIV infection may not necessarily result in a reduction in HIV prevalence among adolescents.

- Specific interventions may be synergistic or even antagonistic, and yet most programmes will combine several intervention strategies, making evaluation of the effects of specific interventions or components within the programme package difficult or impossible to disentangle.

Evaluating interventions targeting young people is difficult

- Measuring the ultimate goal of reducing HIV prevalence among young people requires that substantial numbers of young people are followed up for several years at considerable cost.
- The validity of surrogate outcome measures, such as reported sexual behaviour, may be particularly problematic among young people because of the effects of social desirability biases, age differences between researchers and respondents, etc. (*17*).
- This age group is particularly mobile and therefore difficult to follow through longitudinal research.

Evaluation strategies cannot be standardized

- It is intrinsically easier to evaluate the effectiveness of some interventions (such as those targeted at individuals) than others (such as those targeted at whole communities or nations).
- The timescales in which the various interventions might work vary widely. For example, condom promotion and supply or treatment of other sexually transmitted infections may produce measurable outcomes in a relatively short time, while other approaches, such as changing the socioeconomic status of women, may be expected to have a substantial impact on HIV prevalence among adolescents only in the longer term. Furthermore, some interventions may have longer lasting effects than others.

Evaluation results are not always generalizable

- The impact of an intervention may vary substantially according to the setting in which it is delivered and the broader context. For instance, the effectiveness of a life-skills programme may differ according to the degree of control young women have over their sexuality in that culture.
- Furthermore, the impact of an intervention within a tightly controlled evaluation setting may be different from that within a routine programme.

The contested nature of evidence itself

- Different people accord different weight to different types of evidence. This often reflects their disciplinary background, and sterile debates between “positivists” and “interpretivists” or “relativists” have been at least as common as constructive discussions in this field (*18*).

The need to make decisions using imperfect evidence is the norm when formulating social policy. In fact, within the field of HIV prevention, the evidence available to enable us to make rational policy decisions, and the consensus among researchers and policy-makers, may be greater than it is in many other areas of social policy (*19*). One reason for this relative consensus has been the fact that explicit theoretical models for how interventions are postulated to work exist for most, if not all, of the major interventions that have been proposed for inclusion in programmes to reduce HIV prevalence in adolescents. These models are often imperfect, and empirical evidence for the causal chain within any given model is often weak and sometimes missing, but plausible models usually do exist, often based on social or psychological theory, though also occasionally on biological theory (such as the potential effectiveness of condoms if used correctly).

4.3 Thresholds for strength of evidence needed for widespread implementation

Some types of interventions need stronger evidence than others in order to be recommended for widespread implementation. The strength of the evidence needed depends on their feasibility (including cost), potential for adverse outcomes, acceptability, potential size of effect and potential for other health or social benefits.

- The more **feasible** the intervention, the lower the threshold of evidence needed. Key areas in this domain include the logistics, cost and human resources required for its implementation. The question is: can it be implemented on a large scale in a way that will be sustainable?
- The lower the **potential for adverse outcomes**, the lower the threshold of evidence needed. For example, is there any evidence that the intervention could actually lead to increased HIV incidence or to violations of human rights (*20*) or could it put individuals at an increased risk of domestic violence? Ideally, the assessment of potential adverse outcomes should not be limited to short-term outcomes among the specific individuals targeted but should also include longer-term outcomes within the wider community. For example, in evaluating male circumcision, the assessment should not restrict itself to the impact on the young men who are circumcised. It should also consider the possibility that encouraging male circumcision might lead to more circumcisions being performed in informal non-sterile circumstances, that appearing to endorse "circumcision" might lead to increased female genital cutting, and that it might increase sexual risk-taking because those circumcised may think they are "immune" from HIV and other sexually transmitted infections.

- The more **acceptable** the intervention, the lower the threshold of evidence of effectiveness required. The intervention's acceptability needs to be assessed not only among the target group but also among implementers, politicians, donors, religious and other community leaders, and within the wider community. A controversial intervention will require stronger evidence than a well accepted intervention simply because of the greater reluctance that policy-makers will have to introduce it because of the risk of opposition or protest from key stakeholders. For example, in most contexts policy-makers are likely to be reluctant to introduce active condom promotion and provision within primary schools and more likely to allow the provision of basic information about what HIV is and how it is spread.

- The greater the **potential size of the effect**, the lower the required evidence threshold. Not surprisingly, given the complexity described above, most interventions do not have empirical evidence of their impact on key outcomes such as HIV prevalence. In the absence of this, however, it might be possible to make a plausible assessment of maximum potential impact based on theoretical grounds, process evaluation data or data on intermediate outcomes. Policy-makers might be more willing to gamble on an intervention that has the potential to bring about a major beneficial impact (as long as its cost and potential for adverse outcomes are low and its acceptability and potential sustainability are high) than on another intervention that may have only a marginally beneficial impact. A related issue is the time required to achieve a measurable effect: the longer the time needed, the higher the evidence threshold.

- Some interventions, such as increased access to schooling for girls, may receive additional justification because they are associated with **other health or social benefits**. If so, policy-makers might reasonably have a lower threshold for the strength of evidence of the intervention's impact on HIV risk.

Subsequent papers in this series review the evidence on interventions in five different "settings": schools, health services, geographically defined communities, specific population groups at high risk of HIV infection, and interventions delivered through the mass media. The grid in Box 4.2 has been used in the "settings" papers in this series to decide what threshold of evidence a particular type of intervention requires in order for it to be recommended for widespread implementation in developing countries. The decision on the strength of evidence needed for widespread implementation should be taken prior to considering the actual evidence that is available for a particular type of intervention. The examples in Box 4.3 illustrate why some interventions require stronger evidence (that is, have a higher evidence threshold) than others.

Box 4.2

Threshold of evidence needed to recommend widespread implementation in a developing country, based on five key attributes of the intervention

Threshold of evidence needed	Attributes of the intervention				
	Feasible	Low risk of adverse outcomes	Acceptable (to target population, practitioners, gatekeepers)	Large potential effect size	Other health or social benefits
Low	√	√	√	D	D
Medium	D	√	√	D	D
High	X	X	X	X	X

Key
√ = necessary
D = desirable
X = not necessary

Box 4.3

Examples of interventions and thresholds of evidence needed to recommend widespread implementation

Intervention	Feasible	Low risk of adverse outcomes	Acceptable (to target population, practitioners, gatekeepers)	Large potential effect size	Other health or social benefits	Threshold of evidence
Information[a]	Yes	Yes	Yes	Probably not	Yes (contraception may be used, prevalence of STIs[c] may be reduced)	Low
Condoms in schools[b]	Yes	Perceived potential to increase promiscuity (although there is evidence that this rarely, if ever, occurs)	No (especially among teachers and gatekeepers)	Yes	Yes (provides contraception, prevalence of STIs may be reduced)	High

[a] For example, provision of basic information on the cause of AIDS, how it can be transmitted, and how it can be avoided, especially early in an epidemic.
[b] For example, promotion of the use of condoms and provision of condoms within schools.
[c] STIs = sexually transmitted infections.

For each type of intervention, the recommendations relating to whether a particular intervention should be implemented depend on the pre-defined threshold of evidence needed and the degree to which the evidence meets that threshold. Four kinds of recommendation are found in the papers: widespread implementation now (categorized as "Go"), widespread implementation with careful evaluation in terms both of outcomes and processes ("Ready"), implementation within specific evaluation studies but not yet in large-scale routine intervention programmes ("Steady"), or do not implement because there is strong evidence of a lack of effectiveness or there is evidence of harmful effects ("Do not go") (see also chapter 1). In this series of papers, the guidelines in Box 4.4 were developed to assist authors in reaching decisions about which recommendation should be made for each intervention.

4.4 What information do policy-makers need?

Ideally, detailed and clear information is needed on all of the following aspects of any intervention under consideration:

- a **detailed description** of the characteristics of the most promising approaches or strategies for implementing a particular intervention, including its content, delivery setting, intensity of implementation (for example, the number of hours of training or education involved) and the human, financial and other resource requirements;
- the **theoretical mechanism** by which the intervention is postulated to lead to a reduction in HIV prevalence in young people. Ideally, as well as there being a plausible mechanism, there should also be empirical evidence that the intervention actually works through this mechanism and evidence that relevant changes can occur through this mechanism. As will be discussed in the next section, this evidence need not necessarily come from the specific field of HIV prevention in young people or even HIV prevention at all. It could equally well be drawn from evaluations of interventions using the same mechanism to achieve other outcomes. For example, evidence of the effectiveness of mass media as a mechanism for influencing behaviour could come from interventions related to, for instance, drug abuse, healthier eating, safer driving or the use of seat-belts (*18, 19*);
- the **feasibility and cost** of its implementation, including its sustainability and acceptability to different stakeholders. For instance, there is little value in implementing an intervention that would be too expensive to disseminate widely, would require skills or knowledge that the implementers do not have or could be trained in readily, or that is resisted by the professionals that are meant to implement it. Clearly, taking practitioners' views into account is likely to be critical in assessing feasibility;

Box 4.4

Guidelines for making recommendations of "Steady", "Ready", "Go" or "Do not go"

Minimum criteria to recommend	Interventions needing a high threshold of evidence to be reached				Interventions needing at least a low threshold of evidence to be reached			
	Go	Ready	Steady	Do not go	Go	Ready	Steady	Do not go
Quality of intervention								
Identified mechanism of action	√	NN	NN[a]	NA	√	NN	NN[a]	NA
Experiential base	√	√	√	√	√	√	√	√
Adequate intensity, duration and completeness	√	√	NN[a]	√	√	√	NN[a]	√
Quality of evidence for positive outcomes								
Careful pilot or informed judgement	NA	NA	√	(Neg)	NA	NA	√	(Neg)
Evidence of associations	NA	NA	NN[a]	(Neg)	NA	√	NN[a]	(Neg)
Plausibility evidence[b]	NA	√	NN[a]	Neg	√	NN[a]	NN	(Neg)
Probability evidence[c]	√	NN[a]	NN	Neg	NN	NN	NN	Neg
Evidence								
Probable size of positive effect[d]	≥M	≥S[a]	NA	Neg	≥S[a]	≥S[a]	≥S	Neg
Positive effect is in the cultural context being proposed	√	NN[a]	NN[a]	Neg	√	NN[a]	NN[a]	Neg
Consistency of findings in > 1 study	√	√	NN[a]	Neg	√	√	NN[a]	Neg

[a] Areas where further research should concentrate.

[b] Cases in which other potential explanations have been largely discounted.

[c] Evidence from randomized controlled trials.

[d] This is based on the statistical effect size plus the "reach" of the intervention.

√ = necessary condition

NN = not a necessary condition

NA = not applicable

Neg = negative (harmful) effect is sufficient condition for "Do not go"

(Neg) = sufficient condition to recommend "Do not go" if lack of effectiveness or harmful effects found in several studies for this type of intervention

≥S = probable size of beneficial effect is at least small

≥M = probable size of beneficial effect is at least moderate.

- in evaluating the strength of evidence provided by a particular study, it is essential to have detailed evidence on the actual **process** of delivery of the intervention that establishes the extent and quality of delivery as well as evidence on intermediate indicators that support the theoretical mechanism of the intervention (*16*). For example, for an intervention based on in-school teaching sessions, process information might include data on the number and quality of sessions taught, attendance rates at these sessions and a qualitative assessment by the participants of the sessions' usefulness, appropriateness and relevance. The evidence collected by implementers or practitioners in their daily work can be valuable in offering insights into the daily operations of an intervention and into the kinds of evidence practitioners draw on in their work. Evidence and evaluations at the level of daily practice or through "learning by doing" are often needed to frame future policy. Yet evaluators sometimes do not take into account the key fact that – to be effective when it is scaled-up from a pilot project to the national scale – interventions are likely to need further modifications. Additionally, issues such as political commitment, feasibility, cost and acceptability to implementers and gatekeepers will increase in importance;
- the degree to which the intervention's effectiveness is dependent on the specific **context** in which it is being implemented, for instance the setting, the local and national sociocultural contexts and the specific time period or specific group involved. Information on the context will elucidate factors that may have been necessary preconditions for the intervention to have had the effects observed. Conversely, such evidence will help policymakers decide on its likely generalizability to other settings or populations. If an intervention has been shown to be highly effective in multiple different, but relevant, contexts, this increases the likelihood that it may also be effective in a new context (*21*);
- the **effectiveness** of the intervention in achieving each of the five key UNGASS goals (*1*) described in the introductory chapter using appropriate outcomes. These goals are:
    - Goal 1 – provide appropriate information to young people and evidence of improvements in their resulting knowledge.
    - Goal 2 – provide appropriate skills training to young people and evidence of their ability to demonstrate these skills, and, if possible, evidence that they have actually used these skills to decrease their risk of becoming infected.
    - Goal 3 – provide appropriate skills-based training, equipment and supplies to health-workers and evidence of this resulting in increased delivery of effective, high quality health services to young people. In this context, the health services that are particularly important include

providing advice, counselling on the sexual health (and other concerns) of young people, condoms, treatment for sexually transmitted infections, HIV counselling and testing, and family planning. Clean needles, and other medical instruments and uninfected blood products are essential (see chapter 6).

- Goal 4 – provide evidence of decreased vulnerability to HIV among young people, such as changes in the attitudes and behaviours of adult community members, fewer girls having to resort to "survival sex", and reductions in HIV prevalence among young people's potential sexual partners.
- Goal 5 – provide evidence of a reduction in HIV prevalence among young people that can be attributed to the intervention.

Policy-makers will also need to know many other things: the scale, trends and likely future course of the epidemic in their region, country or district and within specific subgroups of the population (for example, among young people as a whole – that is, those aged 10–24 years, adolescents – those aged 10–19 years – and youths – aged 15–24 years, married and unmarried young people, rural and urban young people, injecting drug users, commercial sex workers and men who have sex with men). Furthermore, policy-makers are likely to put much more weight on some outcomes, such as a decrease in incidence or prevalence of HIV, than on other outcomes, such as those related to the global goals on knowledge, skills, services and vulnerability.

4.5 Assessing the quality of an intervention

Results from a high quality evaluation of a poor quality intervention (that is, an intervention that is badly conceived or badly implemented) merit less weight than those from a high quality evaluation of a good quality intervention. For example, only low weight should be given to the outcome results of a rigorous evaluation of an intervention in schools in which only 20% of the sessions were actually taught. On the other hand, a process evaluation that seeks to explain why this intervention was not delivered effectively in this particular context might be of great value for future attempts to develop an effective delivery strategy for this intervention.

Some criteria that may be used to assess the quality and appropriateness of an intervention are listed below.

- **Relevance:** How relevant is the intervention to HIV prevention among young people? Are the main objectives relevant? Is the intervention relevant to this context? For example, in contexts where most HIV infection is transmitted through injecting drug use, an intervention that ignores this mode of transmission will be of only limited relevance.

- **Experiential base:** To what extent was the intervention developed in the light of existing experience with similar interventions either by drawing on the literature or practitioners' experience?
- **Theoretical basis:** Is there an explicit and plausible theoretical mechanism by which the intervention is postulated to contribute to a reduction in HIV prevalence among young people? Added weight should be given to this criterion if there is evidence that a particular mechanism has worked in other contexts or for other outcomes. For example, if the intervention involves peer education, what is the evidence that peer education has worked in other contexts, such as among older adults or in high-income countries, or for other outcomes, for example programmes directed at preventing domestic violence or decreasing the consumption of alcohol, tobacco and other drugs?
- **Careful pilot testing:** Has the intervention undergone successful pilot testing in the relevant target group? Has it been appropriately evaluated and modified?
- **Feasibility:** Is the intervention logistically viable, acceptable to the relevant stakeholders, and can it be widely disseminated and sustained given existing and projected funds and human resources?
- **Quality and completeness of implementation:** Has the intervention been implemented to a high standard?

Other chapters in this series will address the extent to which specific interventions to achieve the global UNGASS goals meet these criteria.

4.6 Types of evidence and their relative weight

There is a wide array of types of evidence that can be used to guide policy. These range from informed judgements based on experience without any objective evidence of impact on the indicators of the five UNGASS goals through to evidence that is based on more rigorous qualitative and quantitative evaluations of the processes, implementation and outcomes of interventions. One can distinguish between criteria by which to assess the methodological quality or soundness of evidence in its own right and criteria by which more or less weight might be given to findings from different types of evaluation research of equally high quality.

4.6.1 *Assessing the methodological quality of evidence*

The criteria for good evaluation evidence are largely the same as those for research evidence in general, and they can be found in numerous research

methodology textbooks (*22, 23*). Some of the main criteria that apply to both qualitative and quantitative research are summarized below.

- **Transparency:** How clear are all aspects of the research design, the theoretical framework for the study and the literature base? Are the aims and objectives explicit? Is there a clear description of the data collection methods and how the data were analysed? Is the completeness of the data clear (such as, refusals to participate, partial participation, losses to follow-up)? Are possible biases of the researchers made explicit?
- **Representativeness of the data:** Can the findings be assumed to apply to the whole population or group that they are purported to apply to?
- **Data presentation:** Are sufficient data included to mediate between the data and the interpretation?
- **Analysis:** Does the analysis take account of all relevant data?
- **Validity:** Is there an objective assessment of the internal and external validity of the indicators used?
- **Plausibility:** Is a plausible argument made as to why alternative potential explanations for the findings are unlikely or at least less likely than that the findings were due to the intervention itself?

4.6.2 *Criteria for attaching weight to different kinds of evidence*

Given interventions (section 4.5) and evaluations of equally high quality (section 4.6), policy-makers should place different values or weights on different types of evidence. Criteria that can be used to assess the weight that should be placed on evidence include the repeatability of the findings in similar and/or different contexts. Also, evidence based on multiple evaluations with consistent findings should receive more weight than evidence from a single evaluation. If a programme is to be implemented in a similar context to where the evaluations were done, the key issue will be the repeatability of the results from evaluations done in that context. However, for a new or different context, the key issues will be the repeatability of the results from evaluations done in as many different contexts as possible.

In terms of evidence related to the impact of interventions on health and social outcomes, a useful framework for categorising evidence from summative evaluations has been proposed by Victora, Habicht and colleagues (*8, 9*). This framework proposes three levels of evidence:

- adequacy evidence. (This is the term used by Victora, Habicht and colleagues (*8, 9*) though "supportive" might be a better term). For this level

of evidence, all that needs to be shown is that an intervention was implemented and the expected changes occurred;

- plausibility evidence. Here, in addition to the expected changes occurring, it needs to be shown that the effects related to the programme were greater than could be explained by any other external influences;
- probability evidence. In addition to plausibility evidence, it needs to be shown there was only a small statistical probability that the programme's observed effects would have occurred by chance. This type of evidence can come only from randomized controlled trials.

It is important to note that plausibility evaluations must include an adequacy component, and that probability evaluations benefit from assessing adequacy and plausibility at the same time (for example, through careful process evaluation and through comparisons of the effects among those who actually received the intervention, sometimes known as "compliers", versus those in the group that was allocated to the intervention but did not receive it, sometimes known as "non-compliers").

This hierarchical typology of evaluation evidence is demonstrated by the example given in Box 4.5. In this framework, evidence may come either from experimental studies (randomized controlled trials or quasi-experimental studies) or from observational studies, such as cross-sectional, case–control or cohort studies. Quasi-experimental studies are those in which individuals or groups are deliberately and prospectively allocated to intervention or comparison groups, but this allocation is not done randomly. The advantages and disadvantages of randomized controlled trials for evaluating behavioural interventions have been discussed in detail elsewhere (*2–9, 24*). Assuming that both the design and implementation of the intervention and evaluation are of high quality, and that there is evidence of ethical practice and generalizability, the hierarchy of evidence used in this series of papers will be as follows. The greatest weight will be put on evidence from "probability evaluations" (that is, randomized controlled trials) that potentially provide very strong evidence. Next in the hierarchy will be quasi-experimental evaluations that have one or more contemporaneous comparison groups that potentially provide strong evidence. These will be followed by before-and-after or time-series evaluation studies in individuals or groups of individuals (all of whom receive the intervention) that potentially provide weak-to-moderate evidence depending on the degree to which other potential causes of any observed effects have been ruled out. The least weight will be given to reports of anecdotal or experiential evidence or informed judgement, which potentially provide very weak to weak evidence depending on the degree to which other potential causes of any observed effects have been ruled out.

Box 4.5

Examples of evaluations providing supportive (adequacy), plausibility, and probability evidence

Example: A condom promotion programme is initiated among young people throughout a developing country using mass media and social marketing approaches.

Key outcome evaluated: Use of condoms

Supportive (adequacy) evaluation: The number of condoms recorded as having been distributed to young people and the proportion of sexually active young people who reported having used a condom during their last sexual intercourse were substantial after the introduction of the intervention.

Plausibility evaluation: In addition to supportive (adequacy) evidence, there is well documented evidence that both the condoms distributed and the proportion of young people using condoms were substantially greater than before the programme was launched. This could be demonstrated from before–after or time-series studies. There is also evidence that the impact was proportional to the intensity of the intervention in various geographical areas or among various population groups, and the impact was substantially greater in areas that received the intervention when compared with areas that did not. This is best demonstrated through quasi-experimental methods using a control group that is similar to the intervention group. Finally, there is documented evidence that no other activities or background (secular) changes could explain the effects seen. It is therefore plausible that the programme was responsible for the increases that were observed.

Probability evaluation: In addition to plausibility evidence, a sufficient number of individuals (or, where relevant, clusters of individuals) were randomly allocated to receive the new programme. For example, the programme was phased in and during the initial evaluation phase 12 regions were randomly allocated to receive the programme immediately, while the other 12 received the programme after the probability evaluation (randomized controlled trial) period.

Two important caveats should be emphasized. First, there is an important distinction between evaluations of efficacy and effectiveness. Efficacy studies aim to measure the impact of an intervention when delivered in a manner that is as close as possible to the ideal; effectiveness studies measure impact when the intervention is delivered through routine real-life channels. Usually the efficacy of an intervention will be greater than its effectiveness when implemented on a large scale through routine channels. The second caveat is the importance of considering context (for example, delivery setting, culture, country or timing) in evaluating all such evidence. The fact that there may be strong evidence from a well conducted plausibility or probability evaluation that a particular intervention has the intended effects, does not necessarily mean that it will bring similar benefits if implemented in a different context. For instance, bar-based HIV peer education for gay men was effective in the United States of America in the early 1990s (*25*) but not in Scotland a decade later, probably because of different cultural norms and the fact that the intervention was implemented at a different stage in the epidemic (*26*). It is important, therefore, to include at least an adequacy or plausibility evaluation when an intervention that has been found to be efficacious or effective in one context is implemented in a substantially different context.

For simplicity, in this series the policy recommendations ("Do not go", "Steady", "Ready" or "Go") will be made for developing countries as a whole. However, policy-makers should review these in the light of local contexts to ensure that the interventions are important and relevant to their context. For example, the priority given to interventions among intravenous drug users will obviously depend on the frequency of intravenous drug use; equally well the likely effectiveness of interventions in schools will depend on, among other things, the proportion of young people in the relevant age group who attend school.

Information that comes from informed judgment – that is, the considered assessments, decisions and opinions of experienced practitioners or key informants – constitutes a different kind of evidence. This might ordinarily be thought of as being less objective in scientific terms and may not always fit directly into the three-part model discussed above. Nevertheless, such evidence can offer important guidance to policy-makers in the absence of evidence of adequacy, plausibility or probability from formal evaluations of a programme's impact. It can also be useful for triangulation with other data to provide extra certainty in terms of indicators of a programme's operations and effects when aspects of programmes are not documented in ways that are easily accessed by other methods of data collection. Informed judgement and expert-generated evidence (sometimes called evidence of best practice) can be gathered by methods such as interviewing key informants and Delphi scans (*27*). Each of these methods may have different levels of theoretical

sophistication and methodological rigour depending on the design and resources deployed. Such techniques gain strength with repetition over time (for example, repeated interviews with key informants used as a part of a process evaluation) and can be particularly valuable in contextualizing interventions.

Ideally all the different types and sources of evidence for and against the likely effectiveness of an intervention should be appraised in the assessment of whether that intervention should be adopted. The decision should be taken after careful appraisal of the quality of the intervention (see 4.5), the quality of each piece of evidence (see 4.6.1), assignment of weights based on both the evidence threshold for that type of intervention (see 4.3) and the strength of the evidence available (see 4.6.2). To some extent this follows Tones' argument for the use of a "judicial review" in deciding on interventions (*28*) but, unlike us, he avoids weighting one kind of evidence over another and simply calls for triangulation.

Common situations faced by policy-makers include that of having different types of evidence for different interventions or a situation in which the evidence for one intervention is more comprehensive than that for another. In these situations it will be important to carefully assess the evidence and be explicit about what weight is assigned to the different types of evidence. This can be illustrated by comparing the evidence available for the "Stepping Stones" community-wide intervention in the Gambia (*29*) with the evidence available from a recent trial of a complex package of interventions largely targeted directly at young people within the "MEMA kwa Vijana" project in the United Republic of Tanzania (*30*). Put simplistically, there is now a lot of evidence from relatively small-scale programmes that the Stepping Stones approach is feasible to implement (if expensive per person involved) and is associated with changes in knowledge, reported attitudes and reported sexual behaviours (*31*). However, no evaluation has yet reported on the impact of the Stepping Stones community-wide interventions on HIV incidence or prevalence or on other biological markers of sexual behaviour. The relatively large-scale MEMA kwa Vijana project showed that the package of interventions tested within this rigorous community-based randomized trial resulted in substantial improvements in knowledge, reported attitudes and some reported sexual behaviours. However, this trial also evaluated the impact on HIV and other biological markers of sexual risk behaviour, and showed that, at least within a 3-year follow-up period, there was no consistent impact on these biological outcomes (*32, 33*). Policy-makers may be tempted to choose the Stepping Stones intervention because there is no discouraging, short-term biological outcome data but this would be illogical.

Policy-makers must beware of equating evidence from high quality, rigorous evaluations with evidence from less rigorous evaluations. Although a recent systematic review comparing effect sizes in randomized and non-randomized studies did not find a consistent difference (*34*), for interventions of equal quality and effectiveness, the less rigorous the evaluation the more likely it is to give encouraging results (*35, 36*). This has been demonstrated in a review of pregnancy prevention among adolescents (*37*) and presents a real threat to evidence-based decision-making when rigorous evaluations are not available.

4.7 Conclusion

This chapter has proposed a method for reviewing evidence on the effectiveness of interventions that aim to contribute towards achieving the global goals related to HIV prevention among young people in developing countries (improving access to information, skills and services; reducing vulnerability; and reducing HIV prevalence). This method has been used in the five chapters that follow, each of which reviews the evidence for the effectiveness of interventions in a key prevention setting or population group (in schools, health services, geographically-defined communities or groups at high risk of HIV, and through the mass media).

The method involves the following key steps.

1. Define the key types of intervention that policy-makers need to choose between in the population setting under consideration (for example, schools).
2. Define the strength of evidence that would be needed to justify the widespread implementation of this type of intervention ("the evidence threshold").
3. Describe explicit inclusion and exclusion criteria for the studies that will reviewed.
4. Critically review all studies that meet the inclusion criteria and their findings, by type of intervention. This review should include a critical appraisal of:
    - the quality of the intervention. In particular, is it feasible and does it have a clearly identified mechanism by which it operates in order for it to be effective?
    - the data on the process of implementation. Is there evidence that intermediate outcomes predicted by the theoretical mechanism of action are achievable?
    - the context. Is the context in which the evaluation evidence was generated relevant to the context in which the intervention is now proposed?

- the quality. What was the quality of the outcome evaluations, and what were their findings?

5. Summarize the strength of the evidence for the effectiveness of each type of intervention in making progress towards each of the global goals.
6. Compare the strength of the evidence provided by the studies against the threshold of evidence needed to recommend widespread implementation.
7. From this comparison, derive evidence-based recommendations related to implementation of each type of intervention in this setting or population group, putting each type of intervention into one of the "Do not go", "Steady", "Ready" or "Go" categories.

It is important not to be misguided by positive results from poor research. We argue that it is preferable to roll-out a well evaluated programme with good evidence of modest impact than to roll-out a poorly evaluated programme with weak evidence of a larger impact.

The evidence on which we have to make decisions about interventions to prevent the spread of HIV is extremely complex, being about different kinds of interventions, most of which are themselves complex, and arising from diverse evaluation methods. Furthermore, the evidence is imperfect, particularly due to the scarcity of rigorous evaluations of outcomes. Another complication in assessing the evidence is that the very existence of evidence for some interventions and not for others does not occur for reasons that are neutral, but is often the result of past policy preferences, the intrinsic ease of conducting either the intervention or its evaluation, or because the intervention has been seen as controversial. In spite of these difficulties, policy-makers must strive to apply rigour and logic to the selection of intervention strategies, resisting political and other pressures that fly in the face of the evidence.

Finally, even if the evidence that is available leads policy-makers to invest in a particular intervention, this does not mean that there is sufficient evidence about that intervention. Rather, we should always be building on and refining the evidence in the course of implementing public health interventions. Furthermore, evaluators should strive to provide as much detail as possible about the actual intervention, process indicators of the coverage and quality of its implementation, and its costs and effectiveness.

Acknowledgements

We thank Jane Ferguson, Bruce Dick and Ties Boerma of WHO; Mark Petticrew of the MRC Social and Public Health Sciences Unit, Glasgow; Cesar Victora of the University of Pelotas, Brazil; and the participants at

several meetings including Talloires (May 2004), Gex (March 2005) and Bogis-de-Chavannes (June 2005) for their stimulating comments and suggestions on earlier versions of this paper.

References

1. United Nations. *Declaration of commitment on HIV/AIDS.* United Nations General Assembly Special Session on HIV/AIDS. Geneva, United Nations, 2001 (also available from http://www.un.org/ga/aids/coverage/FinalDeclarationHIVAIDS.html).
2. Kippax S, van de Ven P. An epidemic of orthodoxy? Design and methodology in the evaluation of the effectiveness of HIV health promotion. *Critical Public Health*, 1998, 8:371-386.
3. Kippax S. Sexual health interventions are unsuitable for experimental evaluation. In: Stephenson JM, Imrie J, Bonell C, eds. *Effective sexual health interventions: issues in experimental evaluation*. Oxford, Oxford University Press, 2003:17-34.
4. Klein R. From evidence-based medicine to evidence-based policy? *Journal of Health Services Research and Policy*, 2000, 5:65-66.
5. Heller RF, Page J. A population perspective to evidence based medicine: "evidence for population health". *Journal of Epidemiology and Community Health*, 2002, 56:45-47.
6. Rychetnik L et al. Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology and Community Health*, 2002, 56:119-127.
7. Ross DA, Wight D. The role of randomized controlled trials in assessing sexual health interventions. In: Stephenson JM, Imrie J, Bonell C, eds. *Effective sexual health interventions: issues in experimental evaluation*. Oxford, Oxford University Press, 2003: 35-48.
8. Victora C, Habicht J-P, Bryce J. Evidence-based public health: moving beyond randomized trials. *American Journal of Public Health*, 2004, 94:400-405.
9. Habicht JP, Victora CG, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *International Journal of Epidemiology*, 1999, 28:10-18.
10. Briss PA et al. Developing an evidence-based guide to community preventive services: methods. *American Journal of Preventive Medicine*, 2000, 18 Suppl 1:S35-43.
11. Moher D et al. The CONSORT statement: revised recommendations for improving the quality of parallel group randomized trials. *Lancet*, 2001, 357:1191-1194.
12. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ,* 2004, 328:702- 708.
13. Campbell C et al. Framework for design and evaluation of complex interventions to improve health. *BMJ,* 2000, 321:694-696.
14. Des Jarlais DC et al. Improving the reporting quality of non-randomized evaluations: the TREND statement. *American Journal of Public Health*, 2004, 94:361-366.

15. Cochrane Collaboration. *Preparing, maintaining and promoting the accessibility of systematic reviews of the effects of health care interventions* (http://www.update-software.com/Cochrane).

16. Wight D, Obasi A. 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, 2003.

17. Plummer ML et al. "A bit more truthful": the validity of adolescent sexual behaviour data collected in rural northern Tanzania using five methods. *Sexually Transmitted Infections*, 2004, 80 Suppl II:S49-56.

18. Pawson R, Tilley N. *Realistic evaluation*. London, Sage, 1997.

19. Pawson R. Evidence-based policy: The promise of 'realist synthesis'. *Evaluation,* 2002, 8:340-358.

20. Human Rights Watch. The less they know, the better: abstinence-only HIV/AIDS programs in Uganda. *Human Rights Watch*, 2005, 17(4):1-79.

21. Victora CG et al. Context matters: interpreting impact findings in child survival evaluations. *Health Policy and Planning*, 2005, 20 Suppl i:S18-31.

22. Bowling A. *Research methods in health*. Maidenhead, Open University Press, 2002.

23. Bryman A. *Social research methods*. Oxford, Oxford University Press, 2004.

24. Stephenson JM, Imrie J, Bonell C, eds. *Effective sexual health interventions: issues in experimental evaluation*. Oxford, Oxford University Press, 2003.

25. Kelly JA et al. Outcomes of a randomized controlled community-level HIV prevention intervention: effects on behaviour amongst at-risk gay men in small US cities. *Lancet*, 1997, 350:1500-1505.

26. Flowers P et al. Does bar-based, peer-led sexual health promotion have a community-level effect amongst gay men in Scotland? *International Journal of STD and AIDS*, 2002, 13;102-108.

27. Sarantakos S. *Social research*. 3rd edition. Melbourne, Palgrave Macmillan, 2004.

28. Tones K. Beyond the RCT: a case for "judicial review". *Health Education Research Theory and Practice,* 1997, 12:i-iv.

29. Welbourn A. Gender, sex and HIV: how to address issues that no-one wants to hear about. In: *Tant qu'on a la Santé*. Geneva, DDC, 1999:195-227. (Chapter in English.)

30. Hayes RJ et al. The *MEMA kwa Vijana* project: design of a community-randomised trial of an innovative adolescent sexual health intervention in rural Tanzania. *Contemporary Clinical Trials*, 2005, 26:430-442.

31. Paine K et al. 'Before we were sleeping, now we are awake': preliminary evaluation of the *Stepping Stones* sexual health programme in the Gambia. *African Journal of AIDS Research*, 2002, 1:39-50.

32. Ross DA et al. MEMA kwa Vijana, a randomised controlled trial of an adolescent sexual and reproductive health intervention programme in rural Mwanza, Tanzania. 3. Results: knowledge, attitudes and behaviour. In: *International Society for Sexually Transmitted Diseases Research 15th biennial conference*. Ottawa, ISSTDR, 2003. (Abstract No. 0698.).

33. Changalucha J et al. MEMA kwa Vijana, a randomised controlled trial of an adolescent sexual and reproductive health intervention programme in rural Mwanza, Tanzania. 4. Results: biomedical outcomes. In: *International Society for Sexually Transmitted Diseases Research 15th biennial conference*. Ottawa, ISSTDR, 2003:253. (Abstract No. 0699.)

34. MacLehose RR et al. A systematic review of comparisons of effect sizes derived from randomized and non-randomized studies. *Health Technology Assessment (Winchester)*, 2000, 4:1-154.

35. Rossi P. The iron law of evaluation and other metallic rules. *Research in Social Problems and Public Policy*, 1987, 4:3-20.

36. Schulz K et al. Empirical evidence of bias dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association*, 1995, 273:408-412.

37. Guyatt GH et al. Randomized trials versus observational studies in adolescent pregnancy prevention. *Journal of Clinical Epidemiology*, 2000, 53:167-174.