Chapter 1
Ensuring effective evaluation of tobacco control interventions

Introduction

This volume is concerned with methods for evaluating the evidence for the effects of policy initiatives. By policies we mean the enacted decisions of governments and their consequences on the environment (legal, social and physical) in which tobacco use takes place or on tobacco use directly; that is, specific instances of the policy's manifestations (interventions). This means evaluating the effects of laws, regulations, taxes, administrative decisions, programmes and efforts to publicise or disseminate discrete interventions such as smoking cessation aids. It includes evaluation of policies that have the explicit goal of tobacco control, as well as policies that affect tobacco use incidentally, although our focus is primarily on the former. The Working Group (WG) is primarily interested in evaluating interventions that are designed to have effects at a population level, especially those enacted at a national level, but the principles apply to many subnational- and even local-level policies. While the focus of the WG is on how to assess policy consequences of governments, the evaluation framework we have developed could equally apply to the disseminated programmes of non-governmental agencies.

This chapter provides an introduction to the importance of having well-evaluated, population-level tobacco control interventions and of having a framework for achieving them. It outlines criteria used to evaluate constructs and measures, and how these relate to strategies for most effectively gathering information to evaluate the effectiveness of interventions, the mechanisms by which they work, and the conditions that moderate their effects.

Cigarette smoking is not only the most prevalent form of tobacco use, it is also among the most harmful, as it kills one in two long term users prematurely. In the 20th century, cigarette smoking caused an estimated 100 million deaths worldwide. Most of these deaths were in developed countries of the world where cigarette smoking first became popular in the 1920s to 1940s. This resulted in an epidemic of smoking-induced cancer, heart disease, and chronic obstructive pulmonary disease (COPD) deaths. In 2000, tobacco use was responsible for approximately 4.83 million deaths, evenly divided between the industrialised and non-industrialised worlds (Ezzati & Lopez, 2003). If current trends continue, it will cause some 10 million deaths each year by 2030, with around 70% in low-resource countries (Peto & Lopez, 2001; Ezzati & Lopez, 2004). This projected shift is due, in part, to increasing population size and increased smoking in low-resource countries, but it is also partly due to greater success in controlling smoking in many higher-resource countries. In the 21st century, if current usage patterns persist, smoking will cause approximately 1000 million deaths, a tenfold increase over the previous century (Gajalakshmi *et al.*, 2000). A substantial fraction of these expected deaths could be averted by efforts to discourage tobacco use and to assist those addicted to tobacco to quit (IARC, 2007a).

Most countries have ratified the World Health Organization's (WHO) Framework Convention for Tobacco Control (FCTC). It is the first piece of international law emanating from the WHO. Its objective is:

"…*to protect present and future generations from the devastating health, social, environmental and economic consequences of tobacco consumption and exposure to tobacco smoke by providing a*

framework for tobacco control measures to be implemented by the Parties at the national, regional and international levels in order to reduce continually and substantially the prevalence of tobacco use and exposure to tobacco smoke." (Article 3) (WHO, 2003).

To achieve this objective, the WHO FCTC calls for a comprehensive range of measures, specifically:

- Price and tax measures to reduce demand (Article 6)
- Protection from exposure to tobacco smoke (Article 8)
- Regulation of the contents of tobacco products (Article 9)
- Regulation of tobacco product disclosures (Article 10)
- Controls on packaging and labelling of tobacco products (Article 11)
- Programmes of education, communication, training and public awareness (Article 12)
- Bans on tobacco advertising, promotion and sponsorship (Article 13)
- Programmes to promote and assist tobacco cessation and prevent and treat tobacco dependence (Article 14)
- Elimination of illicit trade in tobacco products (Article 15)
- Measures to prevent sale of and promotion of tobacco to young people (Article 16)
- Provision of support for alternative crops to tobacco (Article 17)

In addition, Part VII of the WHO FCTC, on "Scientific and Technical Cooperation and Communication of Information" spells out a framework for research, surveillance and technical cooperation to facilitate the achievement of the policy goals.

Article 20, "Research, surveillance and exchange of information", calls for "The parties [to] undertake to develop and promote national research and to coordinate research programmes at the regional and international levels in the field of tobacco control." The article, among other things, calls for the development and promotion of national research efforts, national systems of surveillance of tobacco consumption and related social, economic and health indicators; coordination of activities so that data can be compared across countries; exchange of publicly available scientific, technical, socio-economic, commercial and legal information, as well as information regarding practices of the tobacco industry; and that the financial and institutional resources be put in place to allow this to happen.

Article 22, "Cooperation in the scientific, technical, and legal fields and provision of related expertise", expands on Article 20 with regard to such things as providing developing countries with technical and material support and training, and identifying methods for tobacco control, including comprehensive treatment for nicotine addiction.

The WHO FCTC will likely result in the proliferation of policies and associated programmes designed to reduce tobacco use. These will include but not be restricted to those mandated or recommended by the Convention. Ensuring the right mix of policies requires an understanding of the determinants of tobacco use and of how tobacco harms health.

Tobacco use is determined by multiple factors, and attempts to control the epidemic require changes in societies as well as individuals (see Figure 1.1). Analysis of the factors that influence tobacco use should encompass smokers, those vulnerable to uptake, tobacco products, those who produce and sell tobacco products, and governments who determine the parameters of use. The role of cultural and economic diversity should also be considered. Further, we need to understand how both the determinants of use and actual use and/or exposures are affected by interventions.

Policies and the disseminated programmes that result from policy decisions are of particular interest because of their potential to affect large numbers of people, in some cases entire populations. As a result, it is important to be able to show that they achieve their objectives and do so in a cost-effective way, with any incidental effects ideally having net benefits. Evaluation allows the most effective interventions to be maintained (and perhaps improved further) while less effective interventions are either improved or abandoned.

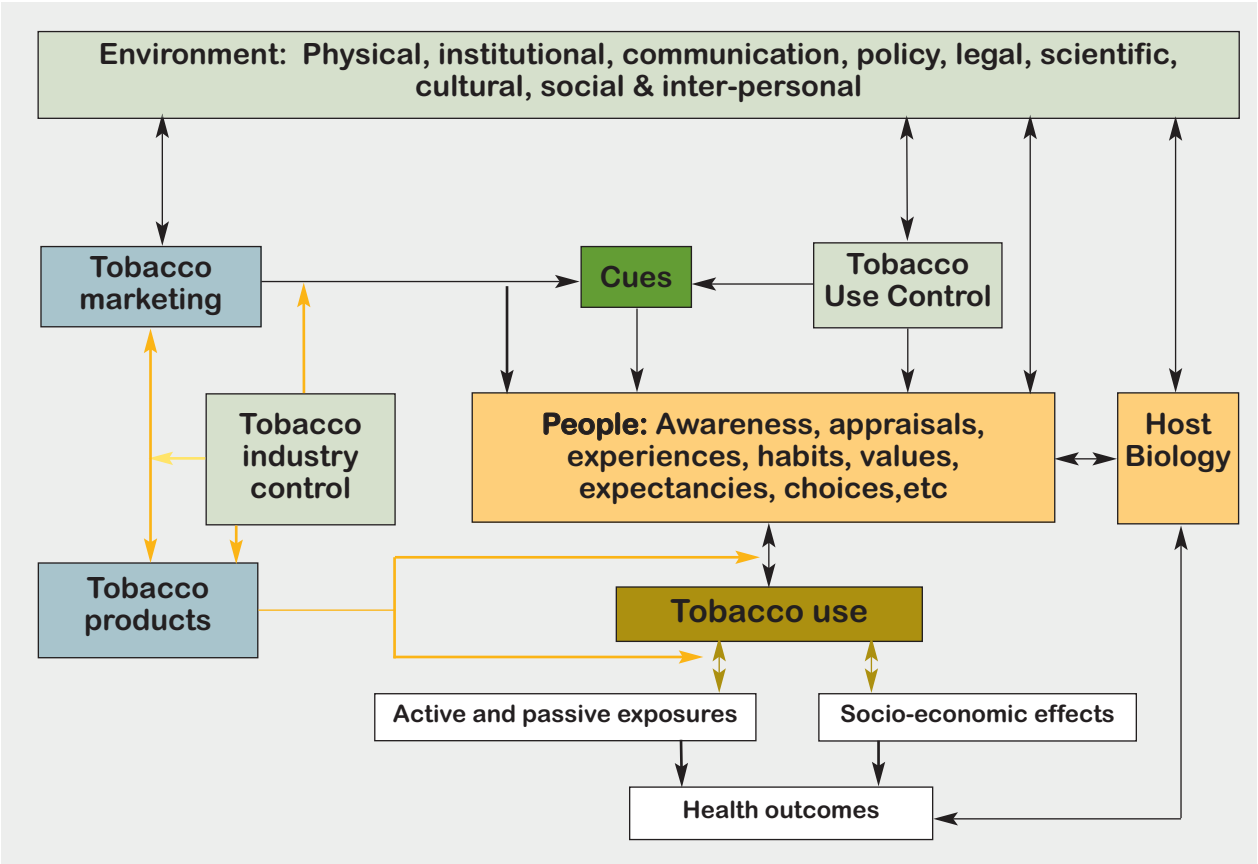

Figure 1.1 Major influences on tobacco use and its consequences
Used with permission of Ron Borland

Tobacco and health

The amount of harm created by tobacco use in a given population is a function of the toxicity of the products, the site(s) of exposure, the toxins taken in, the period over which exposures occur, and the distribution of those exposures in the population (IARC, 2004, 2007b). The harms from tobacco use are mainly from long-term use, which is made more likely by the addictive nature of the product. Calculation of the potential harms that tobacco products cause should consider the composition of what is ingested and how the products are designed to be used. Thus for combusted tobacco products, the focus needs to be on the smoke, rather than on the unburned product, although the composition of the unburned product is relevant to the extent that it influences the composition and/or density of the smoke. Mode of ingestion is often ignored; however, some chemicals are more toxic when absorbed through the lungs than through the mouth lining or stomach because the lungs are more sensitive. The evidence that exclusive cigar or pipe smokers have notably less health risk than cigarette smokers (Doll, 2004) is probably because these smokers tend to only take the smoke into their mouths.

Decades of research on the health effects of tobacco have identified numerous diseases causally related to tobacco use, including several sites of cancer (including lung, oral cavity, esophagus, larynx, stomach and pancreas), major vascular diseases

(including ischemic heart disease, peripheral vascular disease and cerebrovascular disease), major respiratory diseases (including chronic obstructive pulmonary disease, tuberculosis, and pneumonia), reproductive effects and reduced bone health. Epidemiological methods have been applied to estimate how much of these diseases in different populations with different tobacco use histories is due to tobacco (Peto *et al*., 1992).

While prolonged exposures are responsible for most fatal consequences of tobacco use, there is increasing evidence of adverse short-term effects, seen most clearly in the rapidly reversible impacts of passive smoke exposures on non-smokers (Raitakari *et al*., 1999; Wong *et al*., 1999; Wakefield *et al*., 2003a). There is no safe level of exposure to tobacco smoke. Risks of cardiovascular problems are largely reversible, and effects seem to asymptote at lower doses than those related to cancers and chronic lung conditions (e.g. emphysema), where the dose-response curve is more linear across typical exposure patterns (Law & Wald, 2003; Pechacek & Babb, 2004). The addictive nature of tobacco makes it likely that people who begin to use it will not be able to stop before the negative effects associated with long-term harm start to occur.

Nicotine is the main psychoactive ingredient of tobacco and the source of its addictiveness, but is otherwise a minor contributor to the harm (Murray *et al*., 1996; Benowitz, 1999). Most of the harm is due to other constituents in tobacco and tobacco smoke (IARC, 2004). Thus nicotine only indirectly contributes to most of the harms, by leading to prolonged use of dirty delivery systems, especially cigarettes.

The epidemiology is clear. The health risks of smoking are far greater than those associated with smokeless tobacco use. The health risk of each kind of smokeless tobacco varies significantly as a function of their toxicity. For smoked products, the likely variability in toxicity does not seem to translate into clear differences in health risks. To date, cigarettes with levels of toxins reduced by enough to be plausibly less harmful are not used by smokers, so are irrelevant to tobacco control efforts.

Some harms, particularly minor harms and those related to cardiovascular disease, are reversible on quitting smoking. While quitting can improve health, cutting down on consumption does not seem to (Hecht *et al*., 2004; Tverdal & Bjartveit, 2006). This may be in part because, for some illnesses much of the harm occurs at relatively small doses, and partly because smokers who reduce the number of cigarettes they smoke, often smoke the remaining cigarettes harder, ingesting more toxins per cigarette, thus reducing or eliminating the potential benefits of smoking less (National Cancer Institute, 2001). There has been some success in reducing the toxicity of smokeless tobacco products. Changing from smoked to smokeless products (particularly the toxin-reduced forms) can reduce harm, but does not eliminate it (Critchley & Unal, 2003; Foulds *et al*., 2003; Roth *et al.,* 2005; Henley *et al*., 2007). Reducing or eliminating smoked tobacco use is a higher priority for health than reducing smokeless tobacco use. Research is needed to determine whether smokeless tobacco might play a role in this or whether nicotine replacement products and other cessation aids are all that is needed.

Patterns of tobacco use

Tobacco is a plant containing the psychoactive and addictive drug nicotine. It has a long history of use and has been used in a wide variety of forms. The two main forms of tobacco use are by smoking and by chewing or parking wads of tobacco in the mouth and allowing the active ingredients to be absorbed (smokeless use). In the 20th century, the use of cigarettes came to dominate both the smoked and overall markets in nearly all countries. It is also the product that has been the focus of most of the research. In most countries factory-made cigarettes dominate the market; however “roll your own” cigarettes have enjoyed a resurgence in some countries. In other countries, most notably India, people consume a diverse range of tobacco products, both smoked and smokeless. Among smoked products, the “bidi” (tobacco hand-rolled in a leaf) is the predominant form used in the Indian sub-continent. Use of water pipes is common, particu-

larly in the Middle East. Cigars occupy a position as a 'luxury' tobacco product, but use is generally low. All forms of smoked tobacco are extremely dangerous to health, and there has been no major progress towards creating less toxic versions of these products that are sufficiently acceptable to consumers to be successfully marketed. Smokeless tobacco is not used in many parts of the world, but use is significant in other parts, with the products used ranging widely in places like India (e.g. gutka, use with betel quid, nicotine toothpaste), but is limited to one main form in others (e.g. snuff (powdered tobacco) either in loose or prepackaged, small tea-bag-like portions). Use of smokeless tobacco is increasing in some places (e.g. Sweden) (Foulds *et al.*, 2003). Non-cigarette tobacco use is under-researched in comparison to cigarette use.

The proportion of the population who use tobacco varies greatly from around 20% to around 60% (Shafey *et al.*, 2003). In many countries, few women smoke, often accompanied by high smoking rates in males (e.g. in Asia). By contrast, in most developed countries female smoking rates are typically only a few percentage points below that of males. There has been some predictability in these patterns of use, leading to Lopez, Collishaw & Piha's (1994) four-stage model of the tobacco epidemic, with developed countries first to experience it. In this model, female smoking initially lags behind male smoking, with female rates eventually rising. The experience of countries like Singapore and Thailand, which have so far successfully prevented female uptake, suggest that the Lopez *et al.* model does not describe a necessary progression, but that the epidemic may be able to be largely averted in some sub-populations, most notably women, when effective tobacco control policies are implemented.

Over the last 20–30 years, smoking prevalence has fallen markedly in some countries. This is well documented for some industrialised countries (Gilmore, 2000; Giovino, 2002; White *et al.*, 2003). One country, Bhutan, has banned the sale of tobacco products to its citizens. However, in some other countries, rates of tobacco use may have increased. The great diversity both between countries and within countries over time creates huge challenges and opportunities for scientific understanding. One challenge, for example, concerns preventing women from smoking in societies where few currently smoke. This challenge needs to be taken up in ways that are not contrary to the greater emancipation of women in those societies. In developed countries, e.g. in North America and Western Europe, the tobacco industry skilfully used female emancipation as a strategy for linking smoking to images of the modern woman. The slogan "You've come a long way baby" from the notoriously successful Virginia Slims advertising campaign typifies this strategy (US Department of Health and Human Services, 2001).

The most comprehensive change in tobacco control has been in attitudes and rules about smoking in enclosed public places and workplaces. As late as 20 years ago, smoking was effectively ubiquitous in most countries, with smoking allowed virtually everywhere (except where there was a danger of fires or damage to equipment). In some countries, this environment has transformed; several countries (starting with Ireland and Norway) now prohibit smoking in all public places and workplaces, and other countries are following rapidly.

The social acceptance of smoking is declining in most places where it has been studied. This decline seems to be related to the length of time the society has taken to regard the problem as serious, and to progress in the implementation of smoke-free places. In Thailand, for example, equivalent levels of smokers see their habit as non-normative (i.e., that society disapproves) as in Western countries such as Australia, Canada, the UK and the USA, all of which have decades of strong action. By contrast, even though personal disapproval of smoking is high in neighbouring Malaysia, which has only recently taken up the issue systematically, smokers are far less likely to perceive societal disapproval (ITC South East Asia project, unpublished data).

However, it is not just trends in tobacco use and tobacco-related knowledge that are likely to affect efforts to control tobacco use. Broader societal issues may also play a key role. The rapid

emergence of China and other countries as economic powerhouses is likely to affect tobacco use, at least in those countries, as more and more people have money to spend on consumer products like tobacco that are marketed to appeal to "modern" sensibilities. Worldwide concerns about the environment, including the issue of global warming, and the rise of religious fundamentalism in some countries are also likely to have effects, but it is beyond the scope of this volume to speculate as to what these effects might be. However, unless efforts are made to understand how tobacco control fits into broader social changes that are sweeping the world, important determinants of use may be missed, with the resultant reduction in the capacity to identify and implement policies and programmes that work.

In thinking about the potential health benefits of interventions, it is important to consider both their potency and their timing (see Figure 1.2). While the understanding of their potency is focal to this volume, it needs to be remembered that the sooner action is taken, the more lives will be saved. Every year of delay adds millions to the eventual burden of lives lost. Enough is known to act in a comprehensive manner now. The evaluation effort is primarily about helping us refine those interventions, to ensure they are delivered in ways that maximise their effects, and only secondarily, to the development of new more effective interventions.

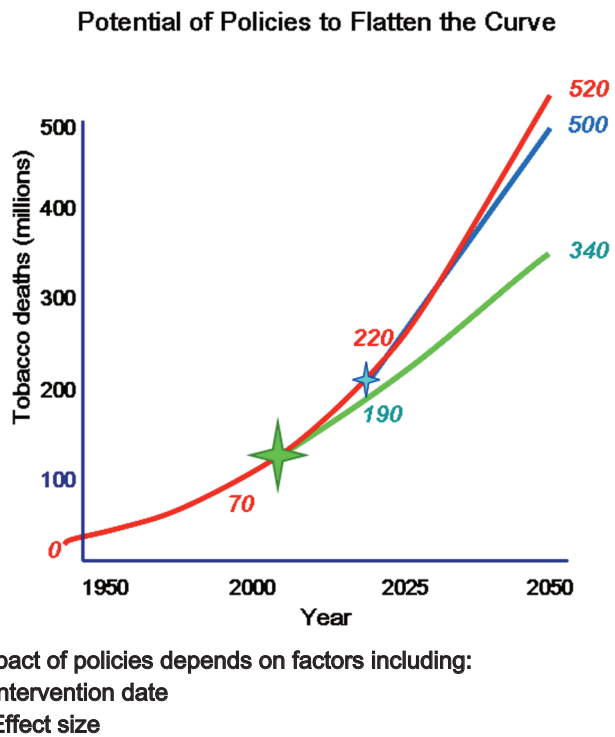

Figure 1.2 Projected impact of population-level tobacco control interventions on estimated cumulative tobacco deaths

Estimated cumulative tobacco deaths 1950-2050 showing the effects of different intervention strategies. In red baseline, in blue if proportion of young adults taking up smoking halves by 2020 and in green, if adult consumption halves by 2020

Adapted from Jha & Chaloupka (1999), The World Bank

Where does this volume fit within Tobacco Control?

This Handbook is not intended to be a one-stop resource for all tobacco control evaluation needs. It is designed to present a framework for evaluation directed at policy effects and to provide strategies and measures that are specific to tobacco control, rather than try to replicate material that is general to all forms of evaluation.

In analysing the potential contribution of research to policy evaluation, it is useful to outline the various roles it can play. Applied science proceeds through a series of iterative stages once a problem has been identified (in this case tobacco as a cause of health harm): elaboration of a theory or theories as to the cause of the problem and of possible solutions, observation and description of the problem informed by the theory, understanding causal mechanisms, intervention

development, intervention deployment and evaluation, and re-evaluation of the problem. From this, there might be the need for new or revised solutions, which may require refinement of the theory or development of a new one. Research can play a number of important roles in the process of developing and disseminating the most effective policy interventions. It can be used to:

1. help in the development of new interventions;
2. help make the case for an intervention being adopted;
3. fine-tune an intervention before implementation to meet local needs (formative evaluation);
4. document the quality of implementation (process evaluation);
5. assess the effectiveness of component parts, or of the intervention under ideal circumstances;
6. evaluate the effects of the intervention as implemented, both intended and incidental;
7. determine the cost-effectiveness of the intervention; and
8. assess the cumulative effects of changes in outcomes on health.

Of these, only number 6 is of focal interest here. All of the others are important, but to have covered them all would have made the volume too broad and too long. We also do not consider evaluation of the efficacy of discrete interventions that can readily be tested in randomised trials; e.g. smoking cessation aids. The Cochrane Collaboration (www.cochrane.org, for reviews) provides regularly updated reviews of evidence in these areas. However, we are concerned with the evaluation of effects of these interventions when applied to populations.

The focus of this volume is the evaluation of tobacco control policies in the short to medium term. We concluded that for policies directed at tobacco use, tobacco use was the outcome of interest, rather than on the subsequent health effects. Clearly, as we move forward, we will want to evaluate the summative effects of all the efforts to reduce tobacco use, and the consequential health outcomes. For a few jurisdictions that have had active tobacco control programmes for decades, this process is already underway (Thun & Jemal, 2006). However, the reality is that for most countries, we will never know exactly how many tobacco-caused deaths are being averted, because there is insufficient data on how many such deaths are currently occurring. The global estimates referred to earlier are a result of careful extrapolation from those countries where good data is available and from studies that have been able to estimate the fraction of deaths from various causes that are due to tobacco. The methods for doing this are beyond our remit, as are ways to model the potential impacts of interventions on smoking prevalence or on the burden of disease (e.g. Levy *et al.*, 2006).

The typical evaluation research study can be thought of as having five components:

1. A research design
2. The choice of constructs and measures to assess them (predictors and outcomes)
3. A sampling strategy
4. Study implementation
5. Data analysis

Of these, we only focus on the first two, although some attention is given to issues of sampling, particularly of the value of having representative samples as a core part of the research design. We do not consider data analysis as the tools here are largely generic and are covered in the main computer analysis packages, including the emerging techniques of GEE models (Generalized Estimating Equations) (Hanley *et al.*, 2003).

This Handbook was not written with the needs of those conducting evaluations at a community level in mind. However, much within it is likely to be relevant, at least at a conceptual level. The cumulative approach adopted means that for evaluations of interventions that have been shown to be effective in comparable situations, the need for intense evaluation will be less, as the evaluation can rely on indicators validated in previous work. However, for novel interventions, the more powerful methods outlined here should still be used wherever the resources allow. The US Centers for Disease Control (CDC) has published a useful guide to the evaluation of more local programmes (MacDonald *et al.*, 2001). A major difference between that guide and the present volume is the capacity to use national surveys and data collections in ways that are not

usually possible for local initiatives. That said, to evaluate local initiatives country-level data can be used as a control, with complementary data collected from the community to assess the intervention effects.

Policy areas not emphasised in this volume

There are a number of tobacco control policy domains that are either not included, or not emphasised. This is not because the WG believes that they are not important, but because it sought to keep the size of the volume manageable. Policy domains not focussed on include some that are designed to affect tobacco use directly, such as sales to minors, restrictions on sales outlets, and school-based prevention. Others are directed more at the tobacco industry, or parts of it, and include prevention of illicit trade, industry subsidisation, controls on access of for-profit companies into the market (and the role of government monopolies), and agricultural policies that affect leaf production.

The most significant area we have not focussed attention on in the volume is the lack of detailed attention to population-level prevention policies. There is a large body of evidence on the effectiveness of school-based education programmes (Thomas & Perera, 2006). The current evidence shows that, taken in isolation of other societal efforts, the impact of school-based programmes is generally weak, and there exists the potential for poorly thought-through programmes to actually be counterproductive. Most of the research on the effects of prevention programmes in schools is from industrialised countries. School programmes are plausibly of more importance in non-industrialised countries, where school is a conduit for new knowledge into the community in a way it no longer is in industrialised countries. The difficulty of developing successful prevention education comes at least in part from the problem that raising the issue engenders interest and thus curiosity about the products. Doing this in a way that overcomes the potential threat of curiosity leading to increased experimentation, and that has a net negative effect on use, has proven difficult. This may explain the interest of some tobacco companies in promoting such strategies. To the extent that educational programmes are translated into the mass media, strategies for evaluating them are covered in Section 5.6 on Measuring the Impact of Anti-Tobacco Public Communication Campaigns.

Another prevention strategy we do not address the evaluation of is policies to prohibit sales of tobacco products to minors, and to enforce these laws by using young people attempting purchases. Such programmes can result in a decline in the proportion of such attempts that result in sales, but the evidence that this actually reduces youth smoking is not strong (Stead & Lancaster, 2000).

In the broad area of tobacco industry control, there is some consideration of illicit trade in the section on sources of production and trade (Section 4.2) and in the section on tax policies (Section 5.1). Neglected areas include restrictions on the number or type of outlets in which products are sold. There are few examples of attempts to restrict the number or type of outlet selling tobacco. However, it seems inevitable that in the future some jurisdictions will try to restrict access to all smokers, not just youth.

We also do not address the evaluation of policies that restrict for-profit companies from operating in the market. Some countries have actual or virtual state monopolies on the sale or production of tobacco products. Several countries have been forced to abandon these monopolies by the World Trade Organisation. It has been argued that non-profit control of the industry should make tobacco control efforts easier (Borland, 2003), but there is little work evaluating either the move to free markets or the potential of restricting the markets. In both these areas, research is needed to evaluate possible options and to estimate likely effects.

A critique of current approaches to evaluation

To achieve maximally effective tobacco control requires the development and ongoing refinement of a viable set of

methods for integrating research and evaluation in the implementation of tobacco control interventions. The population health challenge is to use scientific methods to ensure that systems are set up to understand the effects of the policy initiatives in such a way as to allow their evolution into the most effective ways of controlling the epidemic of tobacco use and related harms. Evaluation researchers in tobacco control, like professionals in other areas of population health, have been concerned for some time about limitations in the evaluation framework used.

The current dominant model of intervention evaluation for improving population health involves extrapolation from the use of randomised controlled trials (RCTs) of clinical (most typically, pharmaceutical) therapies. It is based on the desire to identify the active therapeutic agent or agents within any intervention. This model is important and extremely successful for testing the efficacy and often effectiveness of discrete interventions offered at the individual (and even small group) levels, particularly where double blinding is possible. This is where neither researcher nor participant know who is getting the therapeutic agent under evaluation and who is getting either a placebo or the existing best-practice intervention. RCTs produce considerable certainty about causes. However, reliance on RCTs is not always possible or appropriate for the evaluation of policy impact in the population for a number of related reasons. First, implemented policies cannot be randomised and analogue studies, where randomisation can occur, may lack critical elements of policy interventions (e.g. authority of law, or it being applied to all in the community). Second, over-reliance on RCTs, which focus on the detection of intervention effects, can lead to a neglect of theory, which is critical for generalising from results to related areas, and for understanding the mechanisms by which interventions work. Third, RCTs are not able to answer questions about the relative effectiveness of interventions across different populations. Fourthly, when RCTs are compromised, in terms of deviation from the double-blinded ideal, they are less powerful, and may be less strong than alternative methods with different validity limitations. Finally, focussing on RCTs to provide answers to questions can result in a neglect of other evaluation techniques, which although not as inferentially strong as RCTs, may have complementary strengths. It is important to understand the conditions under which RCTs are limited and what the implications are for inference.

Limitations of RCTs

Determining whether a discrete intervention works involves answering three questions, which sometimes can only be answered separately: the questions of efficacy, effectiveness, and dissemination (Flay *et al.*, 2005). First is the efficacy question: Can this intervention work? That is, when implemented in a controlled and optimal way, does it work? Here the double-blinded randomised controlled trial (RCT) is the gold standard, where possible. Second is the question of effectiveness: does it actually work when implemented under real-world conditions, and with what degree of variation? Third is the question of dissemination: Is the intervention used by enough of the population who would benefit from it to have an impact? An effective intervention that few are prepared to offer or few are prepared to use is of little benefit. One must also consider the extent to which the intervention is similarly attractive for all with the problem. When only a subset of the population benefits, any barriers to selective adoption or influence should be examined. As we move from addressing questions of efficacy, through effectiveness, to dissemination issues, it becomes increasingly difficult to fit the conditions for RCTs, even for clinical interventions.

RCTs involve a number of (usually implicit) assumptions. First, RCTs assume that the measurement required for the evaluation does not affect the integrity of the intervention. Second, it is presumed that the interventions can be evaluated in isolation of environmental factors, including the society's response to

the intervention and to other cultural trends; i.e., that the effectiveness of the intervention can be determined prior to its widespread implementation. Third, it is assumed that any impact of personal choice over whether to have the intervention can be separated from the core therapeutic effect. Fourth, it is assumed that the intervention is uniformly effective for all who are eligible to be given it. None of these assumptions are tenable for policy interventions and disseminated programmes.

The assumption that a given dose of an intervention is assumed to have an equivalent effect on all who have the condition it is intended to treat is problematic even with many pharmaceuticals. The solution to this problem has been to treat each identified population as novel and to require separate RCTs. This might work for major distinct differences, but when there are many possible populations to consider, the strategy becomes cumbersome and costly. More efficient strategies are required.

RCTs are similarly a cumbersome method for evaluating interventions that vary continuously, as they involve creating discrete categories for randomisation. This means there is, for example, poor quality information on optimal dosage, both amount per dose and duration of use. This makes RCTs a particularly cumbersome method for evaluating interventions where the dose of an intervention can vary considerably.

Finally, there is no capacity to consider closely related — indeed, functionally equivalent — interventions as a class, and develop different criteria for evaluating new versions of essentially the same intervention. For example, different executions of a cognitive-behavioural cessation treatment or even the various forms of Nicotine Replacement Therapy (NRT) get treated as independent interventions. In the case of NRT, all variants have had to go through the same process of testing through independent randomised trials before they were able to be marketed.

Population interventions tend to be different in observable ways wherever they are implemented. Information-based interventions are dependent on language, and the language used must vary by culture, not just linguistic group. Language must be kept up-to-date to make it contemporary, and thus attract interest (and sometimes increase) comprehension. People-based interventions invariably differ. Policy-related interventions encompass those major aspects of the system that allow them to operate, not just the core requirements. It is not reasonable to assume that population-based interventions have their effects independent of anything the person does or thinks, unlike most pharmaceutical interventions. Like virtually all psychological and social interventions, as well as some pharmaceutical and other ones, the effectiveness of policy interventions is critically dependent on how the individual responds to them. For clinical interventions, the frame is quite different. Their questions are framed: If the appropriate system is put in place to ensure the person with the illness uses the intervention properly (or is given it properly), then can we demonstrate a benefit? The question the WG ask is quite different and much broader: Can a system be put in place that will make the intervention work, and how can that system be optimised under different conditions?

Where limitations exist on the internal validity of RCTs for making the inferences of interest, the strategy of using meta-analyses of similar studies to draw inferences is similarly problematic. Alternative means are required to control for these threats to inference. It is only in the context of being able to assume generality, having few enough interventions to assume each is an independent case, and having the capacity to test interventions in isolation of their context, that the model of RCTs as the keystone of evaluation is possible.

The allure of having a simple model based on RCTs to allow definitive inferences about the effects of interventions treated in isolation seems to have distracted us from considering the potential utility of other approaches. In particular, the RCT-focussed framework tends to neglect the role of theory and of the potential contribution of combined studies with different kinds of limitations.

The contribution of theory is undervalued in tobacco control and in public health more generally. We agree with the noted psychologist Kurt Lewin: "There is nothing so practical as a good theory." Some in the social sciences take theory to refer to the existing, sometimes demonstrably limited social science models, and take the theories from other areas (typically from the biological sciences) to be accepted fact, rather than theoretical models; e.g. of how a chemical will affect behaviour. Theory is thought of in an encompassing sense of the accumulation of our understanding of how things work, not merely the original ideas. Theory provides the mechanism to systematically use existing knowledge to understand likely future effects. The aim should be to develop consistent sets of ideas (theories) that describe and predict actual outcomes. A hunch or a past empirical finding is an unarticulated theory of what will happen in the future. Unless articulated, these implicit theories cannot be subject to proper scrutiny. If they turn out to predict outcomes, there is no capacity to work out why without first articulating them.

Theories specify mechanisms or mediating pathways of effects, allowing these pathways to be tested. They also can specify conditions under which interventions will work (i.e. moderate intervention impact). One can test whether these factors affect outcomes, and thus be better placed to develop the suite of interventions needed to provide maximal help to all, or to produce the desired structural or cultural changes. No single theory can encompass the complexity of controlling tobacco use; however, more can be done to consider how theories that deal with different aspects of the problem interrelate, including different timescales for change (e.g. behaviour change versus change in cultural norms and practices). The set of theories used should be compatible with each other, even if the nature of the interrelationships is not fully articulated.

The most important implication of considering theory is that it allows explicit linkage of tobacco control to relevant existing knowledge. A focus on evaluating interventions in isolation tends to distract from what is known, specifically:

- Information campaigns can increase knowledge about tobacco.
- Knowledge can change beliefs and attitudes.
- Beliefs and attitudes can affect tobacco use.
- Advertising can change behaviour independent of conscious awareness of the influence.
- There are programmes and aids that can help people quit using tobacco.
- There are ways that the toxicity of products can be reduced.
- Price rises affect levels of consumption of tobacco products.
- Poorly designed and/or executed communications can have boomerang effects.

This knowledge is part of a foundation that is sometimes forgotten. The question we are really asking is: Under what conditions can the desired effects be optimised? This includes concern about the form of the intervention, the ways it is delivered, and various characteristics of the populations to whom it is provided.

A new evaluation framework, one that is less reliant on the RCT, is required. It should have a systems perspective; use the best possible methods, including RCTs where appropriate; allow a more central role for theory, to allow more efficient consideration of possible variation in effects across populations; and provide a more efficient means of understanding effects of dosing and other aspects of implementation.

One approach to evaluation that is popular among public health practitioners, but that has less credibility with researchers, is that of programme evaluation (e.g. Patton, 1997). These models have grown in areas where there are no simple relationships between programmes and sought policy outcomes, yet there is a need to demonstrate progress. Thus the focus of these models of program evaluation is often on determining intermediate effects when it is difficult to demonstrate effects on the main outcome goals. We believe that there is value in extending these models to consideration of outcomes as well. The essence of these approaches is to test the theory behind the programme, sometimes also

called the "programme logic", to assess whether the various aspects of a programme can be shown to contribute to the achievement of its goals (MacDonald *et al.*, 2001). The WG has adopted the idea of using logic models as a core element of the framework we have developed. We found that doing so increased conceptual clarity and provided a useful organising frame for thinking about the policies and a more coherent way to organise the chapters and sections.

Framework for tobacco control evaluation

The role of evaluation is to determine the effects of interventions, determine under what circumstances these effects occur, and help identify ways to make the interventions more effective. To do this involves determining how the interventions work, and diagnosing any problems that either prevent them from working as desired or diminish their impact, in particular any differences of effects within the target population (equity issues). In doing this one should consider the totality of effects, both intended and incidental. To do effective evaluation we need to consider what effects might occur (theory), and design studies that allow detection of effects in the variables of interest (description) and making of valid causal inferences about the contribution of the intervention to the observed changes in outcomes.

Theory

Evaluation must begin with a theoretical evaluation of how an intervention might work. Often there will be one clear theoretical mechanism, generally provided as part of the justification of having the policy, but sometimes alternative modes of effect might be postulated. This is particularly the case when the head of power (constitutional source of capacity to legislate/regulate) under which policies are enacted is limited. Thus policies to protect workers from exposure to passive smoking cannot explicitly consider the possible benefits of smoke-free places for reducing cigarette consumption or for enhancing quitting. Good evaluation requires consideration of all potentially important outcomes, not just those used to justify or provide a legal basis for the policy.

Evaluation is enhanced by showing the mechanisms of the effects, not just restricting itself to determination of effect size. This is critical in population research because most of the outcomes we are interested in are potentially determined by multiple factors; thus it helps demonstrate a contribution from the focal interventions as distinct from other interventions happening at the same time. Thus, the theory needs to spell out the mediational model of how an intervention might work. Mediational models allow us to test each step along a proposed causal chain from intervention exposure to behaviour (see Figure 1.3). If some relationships are not as predicted, the intervention may not be working, at least in the way it was intended to work. In cases where the intervention is known to be potent, evaluation of mediators may only need to proceed as far as assessing uptake/exposure. However, where the potency is unproven, testing the intervention's impact through to the desired outcomes (e.g. smoking cessation) becomes necessary. In an area like tobacco control where the main outcomes of interest (e.g. smoking cessation, pre-

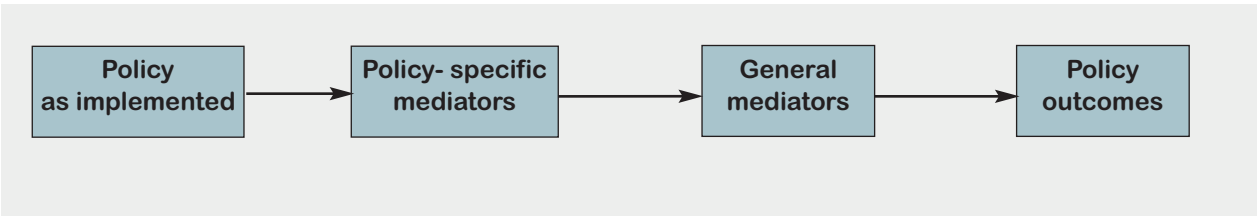

Figure 1.3 A generalised model of mediation

vention of uptake) are determined by multiple factors, mediational models can also help establish the relative contribution of specific interventions. Testing mediational models can also enhance understanding of basic mechanisms and facilitate the development of new and improved interventions.

Other theoretically important factors are those that may moderate the relationship between the intervention and outcomes. That is, what conditions affect the efficacy of the intervention, or how does its effectiveness vary by identifiable sub-groups. Where one finds or theorizes moderator effects, it is important to understand where they occur along the proposed mediational pathways, or indeed whether different mediational pathways exist for different groups or situations (see Figure 1.4). For example, if an intervention is not seen to be relevant to or targeted at a group, this group may not respond to it. Here, making the intervention relevant might be all that is needed to remove the moderating effect. A good example of this is advertisements whose spokespeople are old, which are typically not seen as relevant to young people (the converse is less likely to be true). Something as simple as choice of actor can create moderator effects, which under other conditions would not be present (or be so small as to be ignored).

Incidental effects must also be considered. Sometimes it can be useful to separate these out from the intended effects (see Figure 1.5). Incidental effects can occur for a range of reasons; some may be theoretically expected, while others may not. Some can occur as a result of counter-actions of sections of the tobacco industry to reduce the threats of policies to their profitability. These effects can be incorporated within the more general model (Figure 1.4) as all such effects can be either due to reactions to the policy, or to independent other factors (and thus should be treated as moderators).

Description

The relevant theory tells us which constructs to measure. Evaluation requires a good description of the problem and its context, and of how these are changing. This involves finding appropriate measures of the constructs of interest and of collecting data using the appropriate measures. The goal here is to provide population estimates of what people do and think, focusing on key outcomes. It involves collecting data in four principal domains: 1) who uses tobacco, what they use, how much, and where and when they use it, as well as any relevant knowledge, beliefs and attitudes (including those of ex-smokers and non-smokers); 2) tobacco industry behaviour, including characteristics of their products; 3) tobacco control activities to which people are exposed; and 4) aspects of the broader environment that might affect tobacco use or tobacco harm outcomes (cultural norms, controls on activities like alcohol consumption that are linked to tobacco use). High-quality data collections, such as regular cross-sectional surveys, are essential to describing the nature of the problem and the

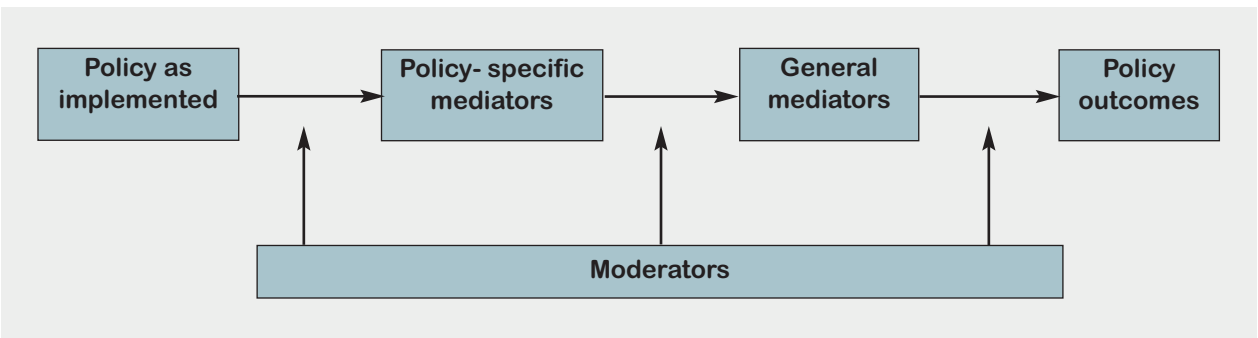

Figure 1.4 A generalised model of mediation, making allowance for moderator effects

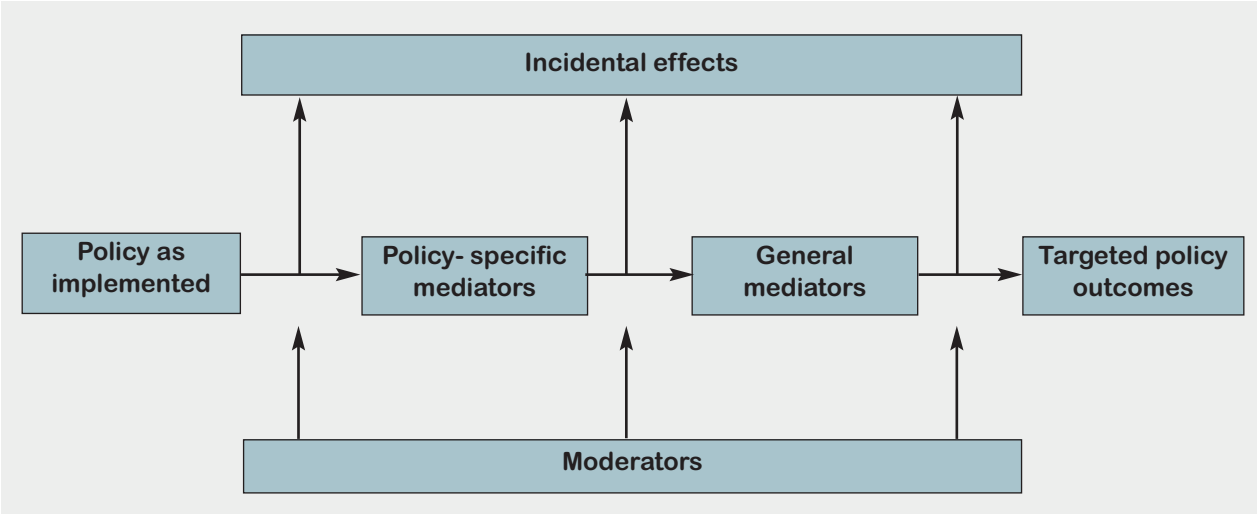

Figure 1.5 A generalised model of mediation, making allowance for both moderator and unintended or incidental effects

quantification of trends in tobacco use and in key determinants of use. In tobacco control, because the tobacco industry or sections of it might be motivated to moderate the effects of policies, it is important to conduct surveillance of possible counteractions to policies. More generally, possible incidental effects of policies should always be considered and measured where appropriate.

There are five broad types of outcomes that relate to individuals: improvements in knowledge, changes to attitudes and related normative beliefs, changes to behaviour patterns, changes in exposures, and health outcomes (particularly acute ones that can be detected soon after a policy is implemented). Interventions typically change the environmental conditions that affect and thus sustain these outcomes. Mechanisms for behaviour change can be through rules and restrictions, making available alternatives or substitutes, and/or providing relevant resources and/or skills. The mediational pathways vary both for outcomes and policies. For example, mediational pathways to knowledge acquisition are shorter than ones to smoking cessation.

Inference

The core of good evaluation is designing studies to detect changes in outcomes that might be attributable to a specific intervention, and putting in place measures to rule out alternative explanations. These alternative explanations are of three types: those related to systematic errors of measurement (bias), those related to alternative mechanisms of effect (confounding), and chance effects. Bias occurs where the measures used to assess the constructs of interest actually measure something different (usually a closely related construct) or are contaminated by some systematic error (e.g. social desirability can affect responses about beliefs and intentions). Confounding occurs when the association with the outcome of interest appears stronger or weaker than it truly is as a result of an uncontrolled association between the intervention and other mechanisms of effect (e.g. a different policy intervention). The contribution of chance is a function of naturally occurring variability in outcomes of interest, and its impact is controlled for by ensuring adequate sample sizes.

The quality of evidence from any single study is a joint function of the study design and of the quality of the measures used: that is, their reliability and validity. Where optimal research designs

are not available, one must focus on the relative strengths of different designs. It is not enough to conduct meta-analyses of the individually strongest studies. A diversity of research designs (and associated measures) with complementary strengths, should be combined, and that information combined in ways that increase the validity of inferences. Demonstration of similar effects with different methods and/or measures increases confidence in the reality of effects and of the plausible causal mechanisms.

Evaluation as a dynamic process

The evaluation of policy interventions occurs after they are instituted, as they first must be implemented somewhere before it is known how they actually work. Because the authority of government policy or law may affect compliance, it is not possible to confidently generalise from the results of analogue studies conducted prior to implementation. This means one cannot in principle be certain of the effectiveness of interventions before they are implemented; hence, lack of evidence needs to be used with caution as a reason for delaying needed policy change. Scientific methods can be used to help us minimise our risk of error, but they can never eliminate it completely. Science should not inhibit action when there is a need for action, but rather act to maximise the chances of success and minimise the risks of wasting resources. This involves a model in which science plays a role of evaluating interventions once they are disseminated, not just restricting its activity to evaluating interventions before they are disseminated. It is a science of evidence in action, not just of evidence preceding action. One aim of this volume is to provide the conceptual framework and some of the tools to allow more effective evaluation of implemented policies and programmes. It is designed to complement the often (necessarily) limited evaluation of interventions that occurs before they are implemented.

There is the possibility that empirical work will show the theoretical model used to develop and or evaluate the intervention to be flawed: either incorrect in some of its assertions (including inclusion of factors that have little or no influence), or incomplete by ignoring important factors. It is only by specifying models that one can systematically work to make them better.

A model of evaluation is required that is designed for the dynamic, ever-changing world in which we live. The potential of the world's diversity must be viewed as a tool to aid in understanding, not an obstacle to be overcome. Each action of government is an attempt to influence outcomes in ways consistent with policy goals, which, hopefully, aim to improve the health and well-being of the community. Similarly, the actions of tobacco companies are also designed to affect smoking, in this case in ways that enhance shareholder value, which is why they are almost invariably directed at increasing or at least maintaining use. Even the best thought-through interventions sometimes fail to work as expected, and policies that work in one context sometimes stop working when the context changes. Because neither past experience nor theory can be relied upon to always deliver the best solution to our problems, methods must be established to check when and how things are working. This is what modern evaluation is about. A framework for effectively evaluating policy interventions is essential.

Such a model places less stringent tests on demonstrating that something has equivalent effects in a new context or when delivered in a new form (where there is no reason to expect changes in efficacy) than it does for evaluation of truly novel interventions or their implementation under conditions where differences in effects is plausible. However, it still calls for stronger evaluation methods when evidence accumulated to question an assumption of equivalence. Thus it provides an explicit link between the roles of ongoing auditing of programmes to ensure continued effectiveness and more intensive evaluation activity when there are concerns. As these decisions are based around clearly articulated theories, the framework is open to scrutiny and should allow the most cost-effective possible evaluation by demanding plausible reasons

before testing for differences in effects.

Characteristics of interventions

Typically, policy interventions are designed to have sustained effects, but in some cases this may require designing ongoing programmes to ensure that this happens. Further, there may be short-term onset effects. For example, there is evidence that warning labels on cigarette packs have an onset effect as well as a sustained effect (Hammond *et al.,* 2007a). We need evaluation methods that can differentiate onset effects from sustained effects, and which also can help us understand the conditions under which both kinds of effects are maximised.

It is necessary to understand what, if anything, is required to sustain potential enduring effects: that is, what endures without further intervention and what requires regular updating, or a sustained presence. For example, anti-smoking mass media campaigns have a short-term impact on quitting (Snyder, 2001). It seems important to maintain cues in the environment to remind people of information for that information to have a maximal impact. The form of some kinds of interventions may also need to change over time if the effects of the intervention are to be sustained. This applies particularly to communication-based interventions. What is seen as up-to-date, and thus of most relevance for communication, changes quite rapidly in some communities. Similarly, across cultures, intervention may need to be framed differently to ensure cultural relevance. Under some circumstances it can be useful to conceptually separate the core concepts in an intervention from the mode of communication used to convey them. Thus evaluation might focus on the cultural relevance of the intervention or on its underlying potency, or both. Analogous to the way societies and/or people change, interventions need to change to maintain their relevance. This calls for an equivalent model to that of how to create new immunizations for emerging strains of influenza. Here, the rate of change in the problem is too rapid for RCTs to be practical, and quite different methods are used.

Changes to interventions may also be required as a society progresses through the adoption of an innovation cycle for adopting new sets of values and behavioural options for tobacco use. Take, for example, encouraging the adoption of smoke-free homes. This happens first in the face of social disapproval, or at least lack of understanding. An entity instituting a ban will often be asked to justify it, and some might see it as unreasonable. However, as such bans become more common, there comes a tipping point where smoke-free environments become the norm. Since justification is no longer necessary, smokers often just do not smoke when indoors, and those without such bans feel a need to justify their positions. Before the tipping point, even quite intense interventions may have limited impact (as has been the case for implementing smoke-free homes (Hovell *et al.*, 2000)), while after it people may be readily able to change without help (as evidenced by rapid adoption of the practice in some countries (e.g. Borland *et al.*, 1999)). Where things change, the rate of change must be considered as well. When it is more rapid than the time for the institutionalisation of interventions through traditional testing of efficacy and so on, then new methods must be adopted to allow interventions to be changed in train with the changing context. This is one of the reasons why it is important to pre-test the messages used for cultural relevance, even for proven interventions when applied in new contexts. This is also why it is important to conduct ongoing evaluation of disseminated interventions.

How policy interventions that target behaviour work

Evaluations of population-level interventions are typically interested in determining the overall effect of the intervention. As a consequence, it is not so much about asking whether an intervention of this kind can work, but of asking under what circumstances does it work and how to optimise those conditions to get maximal impact. This involves consideration of the reach of the intervention (sometimes no more than awareness),

the ways people respond to it and its underlying potency or efficacy.

There are three key aspects of interventions from the perspective of the individual: awareness of, acceptance of, and actions taken in response to policies. Evaluation must deal with all three. The first aspect is determining the extent to which the target population is aware of the intervention, which is a function of its implementation, dissemination, and surrounding publicity about it. Awareness is generally a prerequisite of policy effects, except in those rare cases where the policy creates environmental conditions that can have direct conditioned effects; i.e. independent of conscious awareness.

The second aspect is documenting attitudes towards the intervention by the target population, as this can affect their responses to it. Policies that are unpopular are more likely to be resisted, and forms of assistance that seen as inappropriate to the person's needs are unlikely to be adopted. Thus, a smoker who objects to smoke-free rules is more likely to ignore the rules or to seek convenient alternatives, while a smoker who approves and sees this as an opportunity to gain greater control over their smoking, may not only comply, but use the opportunity to either quit altogether or reduce their consumption. A price increase will only cause smokers to try to quit if they see the increased price as making smoking no longer worth the cost. Alternatively they could smoke more of each cigarette to maintain the value, or shift to a cheaper brand, or seek out sources of cheaper cigarettes, or even re-interpret smoking as something more exclusive and thus desirable. Like awareness, acceptance can only really be evaluated at a population level, although it is typically the acceptance of each individual that is critical. In some collectivist cultures, the views of community leaders are also critical, as they determine what it is acceptable to think and do. These roles are in addition to the roles of leaders in all cultures as policy makers.

The third aspect is the evaluation of the actions that result: that is, the consequences or outcomes of the intervention in terms of both intended and unintended incidental effects. This is a function of both the actions taken by the individual and the potency of the intervention. While traditional intervention evaluation restricts its focus on outcomes among those who are encouraged to use the interventions, for policy interventions this is not a useful restriction; one must consider the total impact on the population, including those who are unaffected. Outcomes should be considered as a joint function of the potency of the interventions, the ways they are used or responded to (a function of attitudes to them), and the degree of exposure to them.

The theories behind tobacco control

A critical step in developing an evaluation framework is having a coherent theory or set of theories as to what tobacco control is about. This should extend beyond the list of tasks identified in the WHO FCTC to an analysis of how the various domains of intervention are theorised to contribute to the overall goal. The nature of the relationship between tobacco use and harm must be sufficiently understood to know what behavioural aims are appropriate. Such an analysis should consider the broad scope of potential impacts, not just those that are part of the rationale for implementing any particular policy initiative. For example, the impact of smoke-free places, introduced to protect non-smokers, also have beneficial effects on smokers and do not appear to have some of the adverse effects on economic activity that some had feared (Scollo *et al.*, 2003). Detailed analysis of the conceptual foundations of specific interventions is provided in the relevant sections later in this volume. Here the WG addresses a few broader issues.

A broad schematic overview of key influences on tobacco use and tobacco-related harm is provided in Figure 1.1. This figure makes it clear that policy and socio-cultural influences have indirect effects on use and that the most proximal determinants of use are the product; cues in the environment; characteristics of people, including cognitions about the products; and the person's biology (both conditioned and innate). Further, the behaviour and the product jointly determine exposures, which, in interaction with existing biology,

determine harm (see Figure 1.6). The role of a systematic science of tobacco control is to analyse and clarify the components of this system and their interrelationships over time, with the aim of introducing interventions that will minimise the harms. Figure 1.6 is a generic model for this. It is possible to elaborate this figure to include other impacts of policies (see Figure 1.7). With generic models of this kind, areas that require greater attention can be expanded upon and boxes where things are more straightforward can be combined.

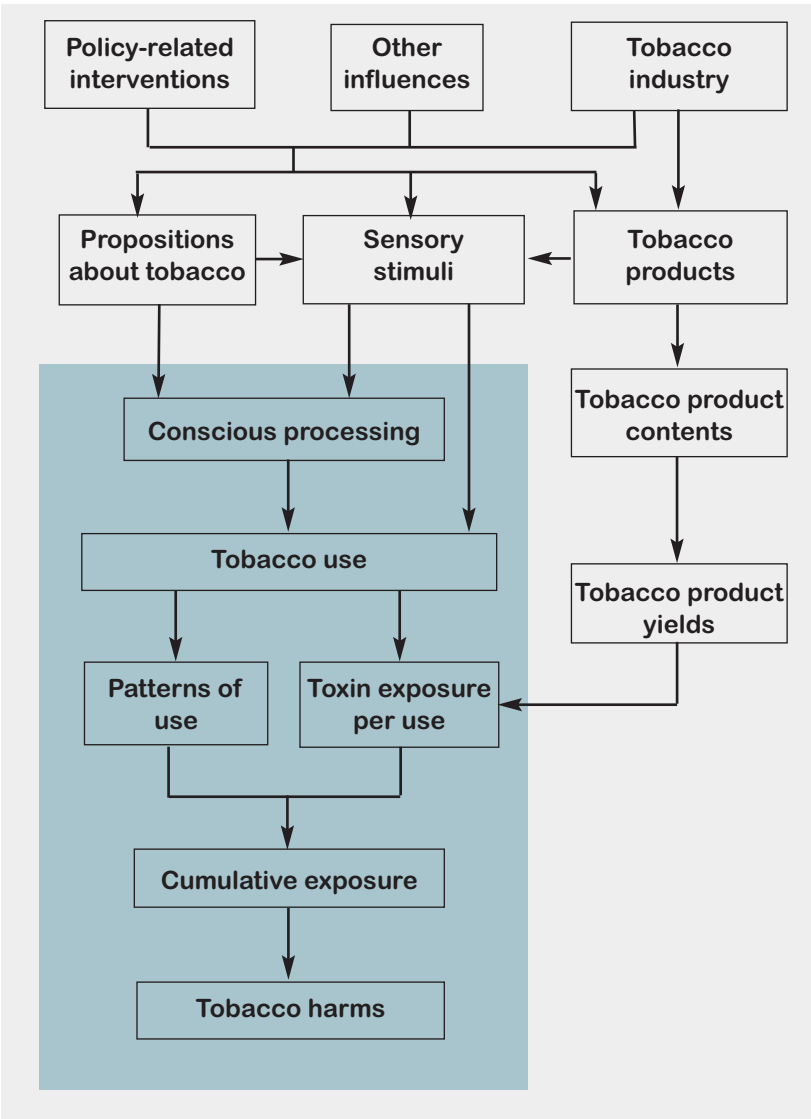

Figure 1.6 Schematic diagram of main pathways by which policies affect tobacco use, tobacco exposures and tobacco harms

Tobacco control efforts can be focussed on users and potential users of tobacco products (e.g. changing knowledge and beliefs), or they can be designed to directly reduce use (e.g. price and availability controls), or to reduce use indirectly by changing the environment to increase cues to inhibit use (e.g. warning labels on packs), or reduce cues to use (e.g. by constraining tobacco companies' marketing practices), or by changing the nature of the tobacco products on the market (see Figure 1.8). Efforts can also be directed at reducing the toxicity of tobacco products (targeting the industry), and at reducing the exposures of non-smokers (targeting tobacco users). To intervene in any of these ways with either people or companies requires a good understanding (theory) of how the factors producing unwanted effects operate and how the intervention will affect those operations. It is beyond the scope of this volume to spell out such a complex theory, although in each section, relevant elements are canvassed.

Tobacco industry controls

Tobacco industry controls are about targeting the 4 Ps of marketing: Product, Price, Place (or availability) and Promotion; to which a fifth P can be added, Packaging; and, unrelated to marketing, the imposition of specific obligations to provide information (for example, warning material) regardless of its impact on the marketability of the products. This is

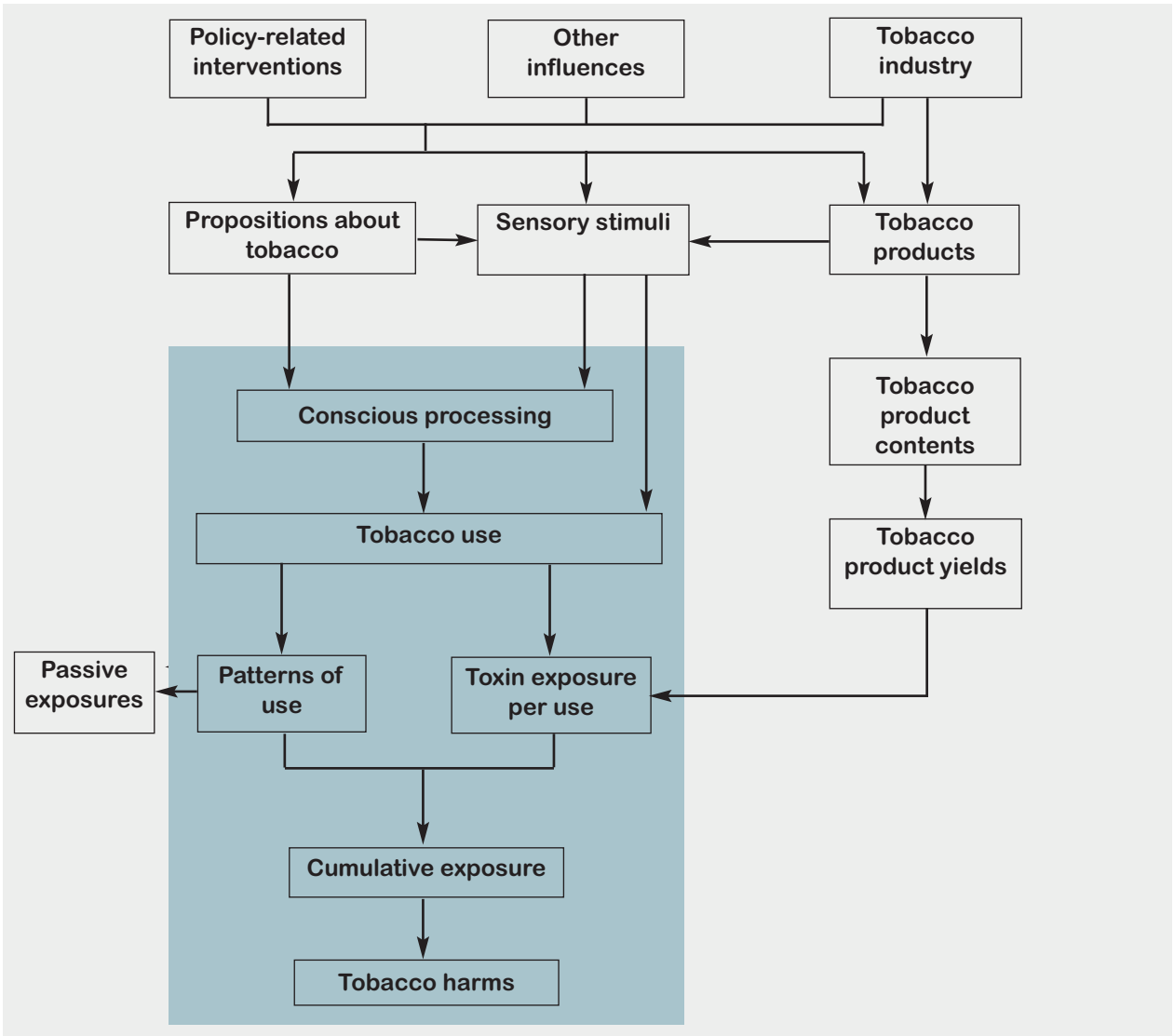

Figure 1.7 Model from Figure 1.6 expanded to illustrate where effects other than on tobacco use fit in

achieved through a mix of laws and agreements, generally targeted at manufacturers or distributors, but in other cases, at other points in the supply chain (e.g. retailers). Evaluation of tobacco industry controls also requires an analysis of possible industry actions to counter the intended effects, or to otherwise minimise adverse effects on their business.

Product controls (see Section 5.3) include rules about what types of products can be sold (e.g. smokeless tobacco is banned from sale in some jurisdiction), levels of constituents or emissions (e.g. upper limits on tar, nicotine and carbon monoxide as measured by ISO standard testing; restrictions on additives/ ingredients), or on engineering features (e.g. mandating reduced ignition propensity cigarettes, filters). The aims of

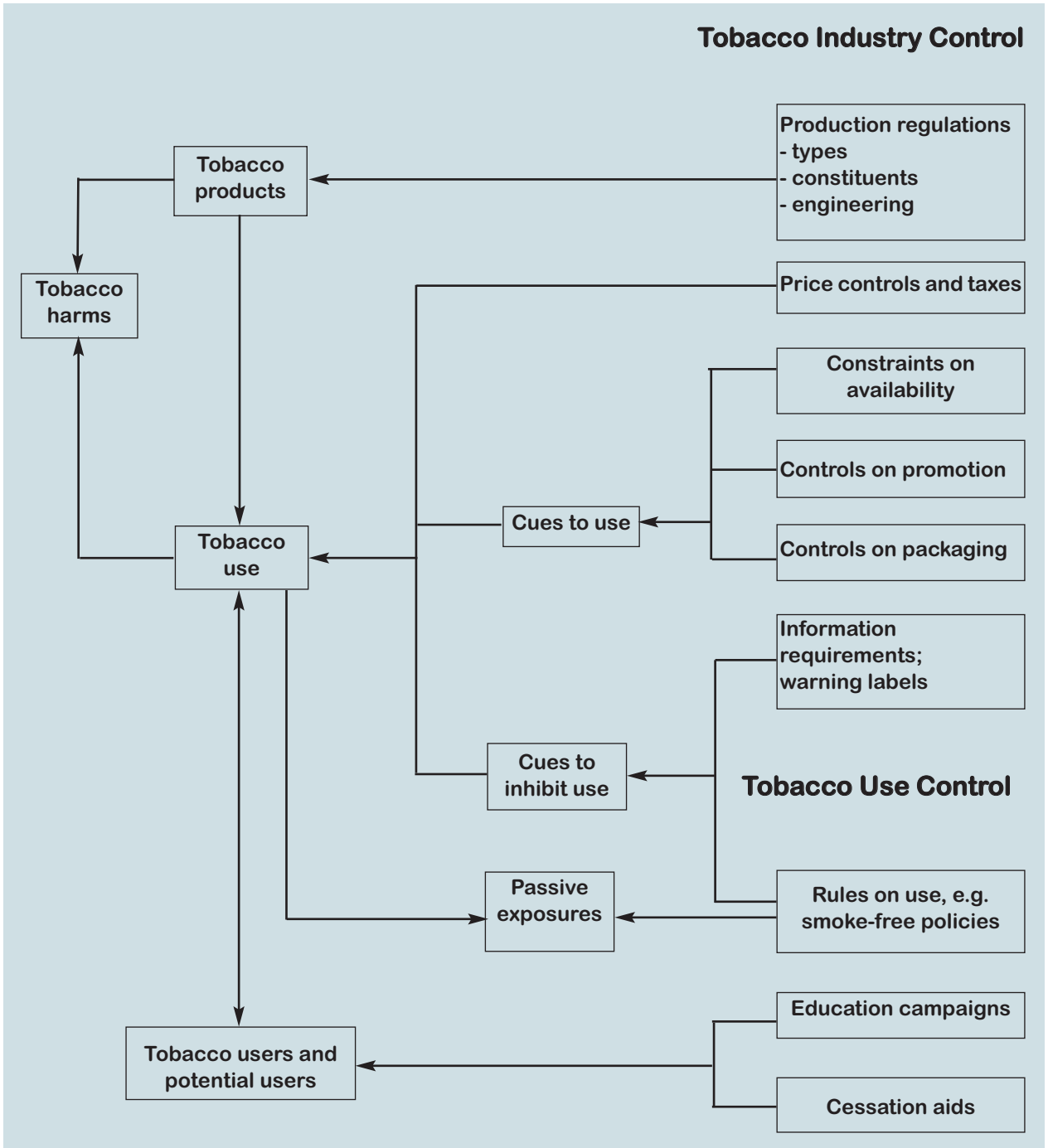

Figure 1.8 Schematic overview of tobacco control interventions and how they relate to tobacco products, users and potential users

product rules vary from preventing new forms of tobacco (to a market) becoming established (e.g. bans on smokeless), to reducing their appeal (e.g. bans on flavourings), both of which are designed to reduce use, and rules to reduce the harmfulness of the products (e.g. constituent limits), which can also have direct effects on the harm caused.

Price controls (see Section 5.1) includes efforts to damper demand through increasing prices (e.g. taxation of various forms), which can have direct effects on use, as well as strategies to prevent price-related marketing (e.g. setting minimum and/or maximum prices to prevent discounting and other forms of price-related marketing).

Place or availability controls refer to efforts to reduce the availability of the products and include restrictions on the number or types of outlets, and to whom they can be sold (e.g. age limits and bans on vending machines). Many of the existing rules have been put in place to discourage use by young people, but restrictions could also be used to reduce impulsive purchases and/or to discourage use in certain venues (e.g. bans on sales in bars).

Packaging controls include rules about what can be on the pack (e.g. use of terms like "Light" and "Mild"; see Section 5.5). It also includes rules that prohibit sale of single cigarettes and establish a minimum pack size to stop use of packs with small numbers of cigarettes, which are known to appeal primarily to young people (Wilson *et al.*, 1987; Assunta & Chapman, 2004a; Prokhorov *et al.*, 2006). The effects of such policies may operate through reducing cues to use, or by making the product less attractive, reduce the value of using such products.

Controls on promotion (see Section 5.4) are the most prominent form of control on the industry. They are essentially about reducing cues to use, but in doing so, might also reduce the appeal of the products. Controls include bans on paid advertising, sponsorships, and product placement, and encompass restrictions on packaging (including controls on the use of trademarks, e.g. generic packaging). Because tobacco is sold in a competitive market, some signs differentiating products as belonging to a manufacturer/marketer are necessary. Even in places when brand displays and advertising is banned at point of sale, a generic sign saying that tobacco is sold is allowed. This promotes availability. Tobacco retailers can also promote products to customers by word of mouth.

The final type of rules is independent of attempts to control marketing, and is about what form and content are required for warnings. The content may include facts about the adverse effects of tobacco use, benefits of quitting, and information about toxin levels (see Section 5.5). Here the aim is to discourage use or at least ensure that any continuing or new use occurs in the context of some information about the risks; that is, it provides cues to inhibit use. Warning and other risk-related information can be required on packets, at the point of sale, on any permitted advertisements, or in conjunction with any depiction of trademarks or commercial mention of products.

Tobacco industry controls are often about reducing cues to use tobacco, while tobacco use control efforts and information provision requirements directed at industry are about increasing cues to discourage use. For cues to use, the effect on behaviour is often conditioned such that they will stimulate tobacco use unless actively resisted. By contrast, cues to inhibit use are more likely to operate via conscious processing.

Evaluation of tobacco industry control is first about assessing compliance with the rules. This is unlikely to be an issue where the rules are to control obvious activities of small numbers of companies (e.g. compliance with labelling requirements), but can be an issue where there is more potential for avoidance (e.g. many potential actors or where the action is not so obvious; e.g. payment/avoidance of taxes). Evaluation is next about determining the effects of the rules. What is involved here varies as a function of whether the rules mandate some actions (e.g. warning labels, higher prices) or whether they mandate removing something (e.g. promotional cues to smoke) that would otherwise be there. In the former case, issues of reactions to the change need to be evaluated. In the latter, the extent

of previous response to the cues (or other things) removed must be known before the impact of their removal can be effectively evaluated. As noted above, it is necessary to monitor and evaluate any industry actions that might occur to reduce the impact of the rules on their businesses.

Tobacco use control

Tobacco use interventions are those targeted at tobacco users or potential users directly. They include rules about use, attempts to provide messages aimed at providing information and changing attitudes and beliefs, and programmes to deliver interventions that can facilitate appropriate behaviour change, or in the case of prevention interventions, effectively inoculate against uptake of any of addiction-level use.

Rules about use include policies to make various places smoke-free (see Section 5.2). Smoke-free rules are generally designed to protect non-smokers, although in doing so they have effects on smokers that need to be understood. Rules could also be about which products could be used, and by whom. However, where there are restrictions on use of products or who can use them, they are usually also codified as rules against selling such products (e.g. smokeless tobacco) or selling to particular individuals (e.g. minors), so these are best considered under industry control even when the parallel restrictions are imposed on individuals as well.

Provision of messages essentially relates to mass media campaigns, where the intent is to expose as many people as possible to the campaign (see Section 5.6). This may include campaigns to promote programmes. Campaigns are designed to inform people and to make the issue emotionally salient enough to stimulate appropriate action. One of the enduring challenges of tobacco control is that because the main adverse effects of smoking are not evident until after a long lag time, smokers do not experience any significant sense of the harm they are doing, and thus tend to underestimate its harmfulness (Slovic, 1998). There are extra issues to consider in the evaluation of prevention campaigns. Focussing on an issue increases awareness of it and may increase interest, which if unchecked could lead to increased experimental use. Designing prevention campaigns or programmes in ways that overcome this increased interest requires thought. There is evidence that some prevention campaigns, especially those emanating from tobacco companies, can have adverse effects (Wakefield *et al.,* 2006), presumably through the increased interest in the issue they engender.

Programmes to disseminate interventions include rules regulating cessation medications, provision of services, and subsidies to products or services (see Section 5.7). The kinds of products/services vary, including self-help resources, stop-smoking pharmaceuticals and coaching or advice programmes of various types. As noted earlier, this volume is not concerned with evaluating the efficacy of these products or services themselves, but on evaluation of their community-wide dissemination and use. Beyond this, there is interest in considering the effects of the existence of cessation services on the broader community. There is some evidence that awareness of the availability of quit-smoking programmes can stimulate quitting activity even among those who do not use the services (Ossip-Klein *et al.*, 1991).

Evaluation of tobacco use interventions should consider both their intended effects and incidental effects. They need to be informed by a sophisticated understanding of psychological principles, and where there are competing psychological processes involved, it is important to put in place measures of all relevant processes. Where additional effects to those sought are known (or hypothesised) they can become further justifications for action (or inaction, if they are or might be undesirable).

Use of logic models

Achieving a comprehensive approach to tobacco control requires adoption of a range of different strategies, underpinned by differing constructs and theories. It is important to spell out the relevant concepts to consider in each area in which a policy intervention might be

planned. The WG has adopted the strategy of encouraging the use of logic models or flow charts to spell out the main constructs that need to be measured for each type of policy. The criterion we adopted was to divide an area to the point where the causal pathways were sufficiently different to make dealing with the various possibilities difficult within the one frame. The WG used Figures 1.4 and 1.5 as generic models, but as will be seen, found the need to modify them considerably for some policy areas. We accept that as knowledge about how some of these interventions work accumulates, new distinctions may become necessary, which could lead to further subdivisions of intervention type. Further, in some cases, distinctions may be shown to be of lesser importance, allowing some of the existing boxes to be combined. It is only once a coherent theoretical model of the domain has been established that determining the constructs to measure becomes possible.

Measurement issues

Measurement is critical to evaluation. To measure the concepts of interest, these concepts must first be defined in ways that make them amenable to measurement. These definitions constitute the constructs. Constructs can be operationalised in many ways. This operationalisation must come from a clear consideration of the concepts and thus of the underlying theory. Because constructs are defined in terms of the theory and not directly in relationship to what measures them, error is localised in the imperfect relationship between the underlying construct and the measures used to assess it. Many of the concepts that need to be measured are not directly observable, or, where they are, they sometimes stretch the capacity of the respondent to recall or otherwise come up with a valid answer (e.g. remembering quit attempts months or years ago). As a result, most measures are subject to a range of possible biases as indicators of their target constructs. Exceptions are characteristics such as sex and date of birth, which in most cultures at least can be reported very reliably (although not in all). One of the great challenges of measurement is that the measures that are most easily obtained are often not ideal operationalisations of the constructs of interest. For self-reported data, most things people report are used as indicators of behaviour patterns or of underlying beliefs, behaviour patterns and/or understanding, not as simple answers to the question. The lack of direct measures also occurs for many physical measures. For example, cotinine levels are sometimes used to assess intake of nicotine or extent of smoking. However, because people differ both in size and in rate of nicotine metabolism, cotinine is a biased measure of intake or exposure at an individual level, although it can be a good estimator at a population level. The problem of the measure that is available not being a direct measure of the construct of interest may be greater when existing data are used, as compromises are commonly made in the interests of being able to use what is at hand. These data were often collected for quite different purposes to those of focal interest, and thus the measures used are often of related constructs, not the exact ones being studied. Dependent on the study, evaluators may be forced to use measures of constructs with different limitations. They need a language to help them talk about the quality of measures in relationship to the constructs they are using the measures to assess. Unfortunately there is no consistent language for talking about these distinctions, and the WG were unable to develop one for this volume. The WG views the development of such a language as critical to reducing the potential for conceptual confusion that can occur from failing to consider the limits of specific measures to actually measure the constructs evaluators are interested in measuring.

Determining what to measure

Choice of potential measures begins with an elaboration of the theory or theories as to how the intervention might work, including the range of expected outcomes and potentially mediating (or intermediate) and moderating variables (effect modifiers), as well as incidental effects. It might also

consider questions like: "What outcomes will lead to health gains?" and "What might influence policy adoption and/or continuation?" Evaluators should also consider whether the same outcomes are relevant to all cultures. For example, in Islamic countries and others where alcohol use is prohibited or not socially significant, consideration of smoking policies in bars is of little interest. Also the relevance of some issues can change as a function of a society's status in regards to tobacco control efforts. For example, support for and reports of smoke-free hospitals are now so high in many countries, it is no longer necessary to ask. However, in countries where passive smoking has not become an issue, asking about smoke-free hospitals may be critical to assessing emerging community concern. This analysis identifies the concepts that it would be desirable to measure.

Next, evaluators need to consider how they want to operationalize the concepts as constructs. This needs to be done in a way that ensures that the constructs are structurally independent of related constructs they might want to relate them to in causal pathways. Further, they need to consider whether the construct can always be measured in the same way. Physical measures typically measure the same thing regardless of context, but answers to questions may not. For example, the direction of social desirability biases might switch as smoking becomes less socially normative. For any given study, they must assess how well the constructs of interest can be measured. Where adequate measures do not exist, there will unavoidably be gaps in the modelling. Sometimes these gaps can be covered, at least in part, by using sets of measures of related constructs.

In Chapters 4 and 5 of this Handbook the WG provides guidance on measures that might be used in various evaluation contexts. For any domain of interest we attempt to characterise constructs that might be measured as one of:

1. *Core constructs:* those that should be included whenever this domain is being studied. These will include key outcomes along with major theorized mediators and moderators. Not having measures of any of these is likely to compromise the study, or at least limit the range of inferences that can be drawn.
2. Important *complementary constructs*, to use for detailed investigation of a domain.
3. *Other measures or indicators* that may add some limited or uncertain value, but which we cannot recommend (for or against), or only recommend in limited circumstances.
4. *Not recommended*: these only need to be specified for commonly used measures that have been shown to have no utility.

The quality or validity of the measures used for each construct also must be considered. *Validity* of measures refers to the extent to which they actually assess the construct they are designed to. This can be assessed through the relationship between the measure and a gold-standard measure (criterion validity), or by showing that the measure related to other theoretically related constructs as hypothesized (convergent validity). One form of convergent validity is predictive validity, where the measure is shown to predict outcomes as theorised. A valid measure of one construct is unlikely to be an equally valid measure of even a closely related construct. Also, the validity of a measure may vary as a function of how it is being used. Thus reports of awareness of environmental cues are not a valid measure of the extent to which any single individual is exposed (because of differences in sensitivity), but may be a valid measure of overall community exposure (as the individual errors are assumed to cancel out across the population). Validity also only relates to the contexts in which it is established. As the context changes the validity of a measure may vary. For example, self-reported age is generally a valid measure of how old somebody is. This is so in cultures where birthdays (anniversaries) are important occasions, but may be less so in cultures where people take no notice of birthdays. Also the validity of measures varies directly with the precision required of the measures: measures that may be valid for detecting large-scale effects might not be adequate for detecting small effects.

The WG uses the following broad categories to provide an

indication of the quality of measures:

- *Gold standard measure.* Established valid measure of a construct of interest that is better than alternatives in all ways.
- *Clearly validated outcome or predictor*. There is evidence that this is a good way of measuring the construct, in at least some specifiable contexts. Limits to validity should be noted.
- *Evidence of utility.* There exists some validity data, but it is not strong. It might be one of a range of alternatives with no clear way of differentiating between them. These should only be chosen when no better measure is available.
- *Face validity*. This involves an analysis of the extent to which the question taps the construct, and may be all that is available for single item self-report measures.

Where possible, we also provide an indication of the sensitivity of the construct to measurement error. For example, how robust is a question to differences in wording? Or indeed, might wording or contextualizing statements need to differ by context and/or by characteristics of the respondent? For example, some questions need to change for use with current smokers as compared to ex-smokers; e.g. "How confident are you that you will be able to stay quit, if/when you try (The last qualifying phase is not needed for ex-smokers)?" Users of this manual should keep in mind that the quality of a measure may be dependent on the type of study in which it is collected and the use to be made of it. The assessments made here assume the measures are made in appropriate circumstances.

Types of data used in evaluation

The type of data needed for evaluation varies, and in some cases it can be found in existing data collections, although sometimes measured in ways that are less than ideal for the new purposes to which it is going to be put. In some cases, measures of the variables of interest are available from more than one source. In these cases, decisions need to be made as to which sources of information are most useful. Issues to consider here are validity, practicality of collection, and the extent to which the data can be related to specific individuals. However, in most cases, the necessary information will need to be collected, giving the researcher greater control over the ways in which the relevant constructs are measured. Some of the main types of data and major ways of collecting it are outlined below.

1. Documentation of policies.
Critical to any form of evaluation is documenting the nature of the intervention. Documentation of policy can occur at two levels: the espoused intent or formal policy (something that is typically documented), and the actual program of activity that is put in place to implement it (which is usually more difficult to document). Policy documents should be collated and coded in ways that allow appropriate comparisons to be made. There is now an international repository of information about the content of national tobacco control policies (See Section 4.1), making this task easier, at least for national-level policies. Some countries collect this information for sub-national policies, but in most cases, the information will need to be collected from each jurisdiction. Where there are many such sets of rules (e.g. of workplaces, local governments), it is usually more convenient to either obtain samples of policies, or to use respondents in population studies to report on the rules that apply to them. Clearly, this latter form is subject to the problem that ordinary people often do not know about rules, and where they do not, may respond in terms of what they remember. For example, when asked if there are bans on smoking in their workplace, some will know the formal rules and respond appropriately, whereas others may know the rules but respond in terms of what actually happens (e.g. if there is a rule, but it is ignored, they will report that there is no rule, interpreting the question to mean, "Can people smoke?"). Others will only be able to answer in terms of what they infer from their recalled observations, e.g. "Nobody smokes there, so it must be banned." This means that such reports may not

be able to help differentiate between policy existence and policy implementation. Indeed, generally there are difficulties in directly determining implementation, especially for complex policies independent of their effects. This is only a problem when the research questions include asking whether problems with a policy occur at the level of policy content, or are a problem of implementation.

2. *Identifying changes in the environment or factors that might moderate policy effects.* The challenges of doing this differ by the environment under consideration.

a) Mass media. Monitoring of national and regional media, with sampling of communities for audit of local media, is the most objective source of what is potentially available. This does not cover some important sources like the Internet. An aggregated respondent report is useful where there are sufficient observations per community unit. Individual reports are subject to sensitivity bias, such that when thinking about quitting, or trying to quit, the person is likely to be sensitized to mentions or images of tobacco or smoking. This means that respondent reports should not be used as indicators of exposure in most individual-level analyses.

b) Physical environment. These consist of rules about public tobacco use and cues to tobacco use from things like point-of-sale displays, billboards, and posters. They can be collected through observation in sampled settings. They may also be estimated from reports from relevant organisations (e.g. of workplaces as to the restrictions on smoking), but are assessed more often by reports from ordinary citizens as to what they experience, or for smokers, what they actually did (e.g. "when last at a restaurant, did you smoke?"). These reports can be averaged across communities to estimate overall levels of these features. Like other respondent reports, these are subject to sensitivity bias, limiting their use for individual-level analyses.

c) Production and sales data. Various forms of sales data, or proxies for sales data, may be available, usually related to reporting on taxes and excises. These may be national-level, but in some cases can be separated by type of outlet or locality. At a national level, there are some international repositories of this information (see Section 4.2). Self-report of price paid is a fairly accurate indicator of prices, but little is known of possible systematic biases.

d) Characteristics of tobacco products on the market. These include composition and engineering features of products and performance characteristics. These can either be gathered from the manufacturers or through independent testing.

3. *Effects on and characteristics of individuals*

a) Self-report data. Characteristics of individuals (knowledge, attitudes and behaviour) are generally only available from self-reports (some scope for proxy reports, but limited beyond smoking status). Self-report data can be of internal cognitive states that are not independently verifiable (e.g. of attitudes, knowledge or experiences), as well as of things that can, at least in theory, be validated, such as behaviours. Sometimes answers to questions can also be used to infer internal states of which the respondent is either not aware or not thought able to report accurately (e.g. personality traits). Many countries have routine behavioural risk factor surveillance studies and/or tobacco specific surveillance studies, and these can be useful in a range of contexts. Many countries use standardised methods and questions, and are working towards common repositories of data (see Section 4.3). Self-reports are affected by question wording and by other aspects of the ways in which the information is collected (see Section 2.2 for some examples).

b) Physical measures. This includes biological and chemical measures (e.g. of cotinine levels). These are often used to measure behaviour indirectly, but this should be done with caution. Limitations of these measures as well as their

strengths are well documented (Benowitz, 1996a; Matt *et al*., 1999; Al Delaimy, 2002).

c) Proxy reports. For observable aspects of behaviour, reports of others who know the target individual may be useful.

Survey methods for evaluation

Survey methods are crucial to many forms of policy evaluation. These can range from surveys of individuals to surveys of informants about the activities of organisations (e.g. of governments or workplaces). Two key issues are addressed here: the sampling frame and the way the questions are asked and answered.

Sampling: To be able to generalise to a population, the sample needs to be representative of the population. This is a function of both the sampling frame and participation. It is thus desirable to have broadly representative samples, recognizing that true representativeness is unattainable. Participation is also crucial. Any biases in participation threaten representativeness. Because often nothing is known about all or some of those who do not participate, quantitative estimation of biases is either impossible, or partial at best, meaning their likely effects need to be inferred. The higher the response rate, the less likely major biases are, but unless the rates are close to 100%, biases can occur.

Sample size is another important consideration. The two main factors to consider here are the size of effects that are expected (or the required power to detect) and the desire to explore potential moderator effects. In principle, making a study larger does not improve its representativeness. However, because size does increase power to detect moderator effects, larger samples can be used to increase confidence in the generalisability of the findings to all groups who have a sufficient sample size for such possible interactions to be tested.

Question asking: The main issue with surveys is inconsistency and bias in the ways in which people respond to questions. This is part of a general phenomenon of the frame of reference or context for the question affecting how it is understood, and thus how it is responded to. Variation in frame of reference includes mode of surveying (e.g. face to face vs. phone interview vs. self-completion). There is emerging evidence that some modes of surveying result in better response rates for sub-sections of the population. There is an urgent need for research to develop optimal methods for calibrating both questions and sample characteristics across modes (see Dillman & Christian, 2005, for a discussion of general issues concerning mixed-mode surveying). As it is beyond the scope of this volume to document the entire range of issues corresponding to questions (there are several excellent texts on this topic; e.g. Foddy, 1993; Fowler, 2001), we deal only with two issues in this chapter. These are the time frame over which answers apply, and cultural factors in interpreting question meaning.

The time interval over which the response is deemed to be valid is a crucial issue in testing causal models. Causes precede effects, so one must assume that predictor variables when measured at the same time as outcomes, predated the occurrence of the outcomes. Sometimes questions are given a time frame or timing of events is asked for to assist in determining sequences. Self-reports of periods or of dates are subject to biases in reporting with events sometimes displaced in time. Self-reports are typically better for recent events (due to memory effects). Salient events may be reported as experienced more recently than in reality, and less salient events are prone to be forgotten.

Aside from issues concerning the context of survey delivery, the way in which respondents interpret questions and response formats affects their answers. One key aspect is the extent to which the conceptual framework underpinning the questions reasonably applies across the cultural contexts under consideration. As research moves from studying issues like tobacco within Western European and North American cultures, to studying tobacco use across cultural settings where there may be different values and assumptions, there is a need to question the underlying assumptions that frame the research. Within all cultures, there will be variation that researchers should try to characterise and understand. The possibility that cultural differences may compromise the

utility of some questions needs to be reviewed on a case-by-case basis. Some of these issues and methods for overcoming them are covered in Section 2.2.

In principle, the response to a question can be directly compared when the respondents are answering the same question. People generally assume this means the same wording. However, under some conditions, the same wording can result in quite different questions being answered, and different wording may be required to achieve equivalence. The most obvious example is asking questions in different languages, but it can occur for the same language where respondents' assumptions about what is being asked can vary systematically, and achieving equivalence requires different contextualising words for different individuals. This can be caused by words having different nuances in different cultures, or effects due to the familiarity and or normativeness of the issues being asked about.

As surveys become standardised, there is a tendency for surveys to converge on common ways of asking questions, thus implicitly operationalising the constructs they are interested in. To the extent that either the operationalisation has an arbitrary element or the measure is flawed, there is a risk of institutionalizing error. To avoid this, it may be important to analyze whether different ways of asking questions may improve the ability to measure a construct. There is always a role for asking questions in different ways. Where the answers are relatively invariant to the form of wording, one can have considerable confidence in generalisability across the inevitable wording differences between languages. However, where responses are sensitive to wording, it is less likely that different forms are actually measuring the same construct, and extra care will be required in translation.

Study designs for evaluating population interventions

To best understand the implications of policy change (including community-wide dissemination of interventions), research designs should be as strong as possible. In Section 2.1 the relative strengths of various evaluation designs are can-vassed. In short, evaluation is strengthened with more observations (both before and after the intervention) within the population an intervention occurs in, the more populations that are studied in parallel, and the more alternative explanations for outcomes that are assessed within each study. In addition, the use of cohorts adds considerable power by allowing mediation and moderation effects to be tested more precisely. Finally, representativeness of the sample to the study population can increase the generalisability of findings. The ITC study (Fong *et al.*, 2006a) is a good example of what can be achieved by attempting to implement as many of these attributes as possible.

Achieving the strongest possible evaluation involves putting in place measures of key outcomes (at least) as long as possible before the policies are implemented. Obviously the best way to do this is if the measures can be part of the country's ongoing surveillance system. Where this is not possible, the studies should be implemented as early in the process of discussing policy change as possible.

For detection of trends, it is important that both sampling frame and participation rates remain constant. This is to maximise the likelihood that biases are likely to remain constant so that any changes are unlikely to be due to a sampling effects. Repeatability is more important than representativeness for determination of trends because it requires comparability between estimates over time.

Such a research agenda requires monitoring of all relevant variables in a diverse range of communities or jurisdictions over a period of time in which there are differences in policy implementation between those communities. This will include use of repeated cross-sectional surveying, and where possible, more in-depth longitudinal cohort studies of samples of relevant individuals (e.g. smokers, and young people at risk of uptake), to begin to explore how the changes come about and whether some groups are affected differently to others. This surveying will need to be complemented by longitudinal monitoring of ecological variables. The level (nation, state, local area) of the variable measurement will determine the

practicality of maintaining ongoing monitoring of all activity or whether some sampling is necessary.

Such a program of data collection is needed to provide the infrastructure necessary for understanding the mechanisms of population level change. Among other things, it would increase understanding of which factors are culture-sensitive, and which are not, and how the roles of various factors change as a person's position towards changing and adopting target behaviour changes. Similarly, it would allow for an understanding of how community readiness to change affects realized change and how readiness can be modified, as well as the conditions that facilitate the institutionalization of change. For policy makers, it can provide information on need for further action.

Drawing conclusions about causes

The approach the WG has taken to evaluation shares more with the methods used in epidemiology to determine causes of illness, than the reliance on RCTs to assess clinical interventions. As a result, when considering criteria to use in drawing conclusions about the effectiveness of policy interventions, we have adapted the criteria used in the epidemiology of disease (Hill, 1965). The adapted criteria are:

- Magnitude of the observed effect, particularly in relationship to known naturally occurring variations;
- Temporal relationship between intervention and change in target outcome;
- Exposure-response gradient;
- Biopsychosocial plausibility; that is, the effects can be explained as occurring through a plausible mix of biological, psychological and/or social process;
- Coherence across lines of evidence with different threats to validity, e.g. similar results using aggregate data and self-reported consumption could rule out response biases;
- Coherence of results from demonstrations of effects on different parts of the theorised causal pathway, or by demonstrating efficacy of components (e.g. the evidence of efficacy of many cessation aids makes it more likely that they have effects when delivered as part of programmes of help);
- Evidence that this type of intervention can have effects on other comparable outcomes (e.g. on other behaviour patterns);
- Consistency of observed effects across studies and populations, or clear patterns in the variability to demonstrate limits to generalisability;
- To which we would add: Elimination of theoretically possible alternative mechanisms for explaining the observed effects.

Policy evaluation has added challenges to other forms of outcome evaluation, because policies usually occur in a mix and policies are only one set of factors that are responsible for the outcomes of interest. Smoking prevalence or rates of quitting are determined by multiple factors, and establishing the contribution of each individual intervention is difficult. The task of differentiating the contribution of all possible contributors to the observed effects is difficult.

In providing a summative evaluation of the effects of an intervention, we need to not only consider the size and nature of effects, we also need to consider the possibility that there is no meaningful effect. In particular, it is important to make a clear distinction between evidence of the absence of effects, and the situation where there is a lack of evidence; that we really do not know whether an intervention works or not. We recognize that science cannot prove the null hypothesis, but it can and should make statements about interventions where there is a consistent failure to find evidence of any meaningful effect.

We need to qualify effects with a statement about generalisability. Some interventions have similar effects in most contexts, others can be quite context-specific. This consideration needs to cover cultural adjustments to the intervention itself, as well as factors in the environment that might affect its potency (effect moderators). It is also important to consider the direction of effects. Some interventions might prove counter-productive. Clearly less evidence should be required to stop an intervention where the evidence suggests that it is counter-produc-

tive, than if it suggested no effect or only a small positive effect.

The levels of evidence framework used to evaluate discrete interventions is not appropriate for use in evaluating policy interventions. We see more promise in adapting the criteria used by the International Agency for Research on Cancer (IARC) for its Cancer Prevention Handbooks. This is essentially a four-level system: *Sufficient evidence* of an effect, *Limited evidence, Insufficient evidence*, and *Evidence suggesting lack of effect.* The WG's concerns with adapting this framework to our purposes, is that it does not allow for gradations in confidence of concluding no effects, it does not clearly differentiate adverse effects, and it does not consider issues of generalisability, all of which are desirable qualifiers in the policy context. One possibility would be to adopt a matrix as shown on this page, with additional statements on effect size (for established effects) and on generalisability.

The effect size could be rated as: Small, Medium, or Large (or undetermined). Consideration needs to be given to whether the highest level of certainly could be applied to interventions where there had not been a direct demonstration of effects on the target outcome, or whether inferred effects could ever be rated as better than Probable. For example, it has been shown that larger health warnings lead to more thought about quitting, and that more thoughts predict future quitting. However, nobody has shown that there is more quitting in the context of stronger health warnings being introduced. How reliably can one conclude that stronger health warnings stimulate quitting?

Finally, once the effectiveness of an intervention is established, less powerful research designs will be needed to monitor continuation of effects and/or to assess whether similar magnitudes of effect are attained with new populations. It is only when there is reason to believe that there are real differences that stronger research methods might need to be reapplied.

How to use this Handbook

This Handbook is designed as a guide for program and policy evaluators. The WG hopes it will be used as a tool for training new evaluators and those who need to understand evaluation principles. It can act as a reference source for arguments about the role of evaluation and the way to think about evaluation, and by extension the development of effective interventions. In doing so, we hope it provides a framework for increasing the scientific credibility of the field, by helping to show that policy evaluation has rigorous methods and can make important contributions to knowledge.

We also hope it will act as a stimulus for further action to improve evaluation methods and measures. As such, this Handbook will need to be kept as up-to-date as possible. This might involve periodic revisions once the principles have been tested, or some other mechanism for moving our expected standards forward. There is a particular need to update the material on specific measures and on the status of data repositories, as these are in a constant state of change.

We hope this Handbook will provide a stimulus to work towards greater coordination of the ways in which policy evaluation operates and the development and/or expansion of international repositories to collect the relevant data and reports, and user-friendly ways to extract this information and synthesise it.

Some future actions the WG would like to see:

- Work to coordinate and arrive at a set of core terms that are most useful for our field.
- Work on what the criteria for validation should be for the

The evidence matrix

No evidence is available			
Possible effect:	**Negative**	**Not meaningful**	**Positive**
Probable effect:	**Negative**	**Not meaningful**	**Positive**
Established effect:	**Negative**	**Not meaningful**	**Positive**

various kinds of measures used, and how that relates to the different types of measures.

- Development and agreement on use of prototype formats for reporting on frequently repeated interventions, such as mass media campaigns. This will facilitate their combination into meta-analytic studies, especially important for understanding where and when things work.

In conclusion, this volume should be thought of as an important step in a process, rather than as a static recipe book for evaluating tobacco control interventions. The methods described and the measures provided are the best available today. The principles outlined in this volume will persist, but those principles require that methods and measures be adapted to the changing world. The WG has built into this Handbook some guidelines for seeking out the latest methods and some guidance in assessing the need to move beyond the measures and methods described here. We believe that this dynamic but systematic approach is the best way to approach the future because it provides a framework that allows evidence to guide action both before and after programmes or policies are implemented.