2.1 The importance of design in the evaluation of tobacco control policies

Introduction

The goal of this section is to describe elements of research design for evaluation studies and how they can form the basis for stronger conclusions about the impact of policies. The groundwork for evidence-based medicine has come from painstaking evaluation studies of treatment options. It follows then that the foundation of an emerging evidence-based public health policy must begin with building a database from rigorous evaluation of public health policies. It should be noted that the elements of research design that we offer in the domain of population-level tobacco control can easily be applied in efforts to evaluate any population-level policy or intervention in public health. Just as surely as the laws of gravity operate in Mumbai as they do in Lyon, the principles of causality, and the methods employed to make more confident judgments about causal relations, are not constrained by location nor area of research.

This section does not offer a comprehensive review of evaluation research design. (see Cook & Campbell, 1979; Shadish *et al.,* 2002; Rossi *et al.*, 2003 for discussions of evaluation research, and Rootman *et al.*, 2001 for the evaluation of health interventions). We focus on *impact evaluation*, that is, whether the implemented policy led to desired outcome(s), rather than other forms of evaluation, such as *process evaluation* (e.g. identifying and evaluating the processes that led to the creation and/or the implementation of a policy).

More specifically, our aim is to highlight how the inclusion of specific features in the design of a policy evaluation study can lead to more concrete conclusions about the possible causal impact of that policy. This section focuses mostly on the structural aspects of research design. Good evaluation design involves the selection of appropriate measures of high validity and reliability. Guidelines and recommendations for such measures, across tobacco policy domains, are provided in other sections of this Handbook.

This section does not provide a review of the statistical analyses that are employed in evaluation studies. However, we do wish to point out one common misconception about the role of statistical methods in attempts to ascertain causality from data*: causality is to be found in the design, not in the statistics.* No statistical method, not even those whose name may imply some special status in this regard (e.g. *causal models*) can confirm causal direction. A structural equation model (with or without latent variables) that yields a significant coefficient for A→B cannot be used by itself to conclude that A causes B rather than B causes A. To do so would be to fall prey to the logical error of *affirming the consequent:*

Statement: If A causes B, then the A→B path will be statistically significant
Observation: The A→B path is statistically significant
False Conclusion: Then A causes B

The advantage of more advanced statistical techniques is that they can take into account characteristics of the data to yield a "better" estimate of the A→B path coefficient. For example, structural equation modeling with latent variables (Bollen, 1989; Hoyle, 1995; Kline, 2005) explicitly models the measurement error from multiple measures of a construct (latent variable), so that the resulting estimate of the relation between that latent variable and another variable is free of the measurement error

that would otherwise have biased the estimate[1]. However, this statistical method does not advance in any way the argument that A causes B rather than B causes A. In fact, a system of variables with paths going in one direction will yield exactly the same model fit as if that same system of variables had all the paths going in the opposite direction.

The key to advancing the quest for causality is to be found instead in the design of a study. Here we offer a review of the elements of the design of evaluation studies that will increase the confidence with which causal statements can be made between and among variables (e.g. whether a tobacco control policy had a desirable causal impact on behaviour).

In our review of research design features for the evaluation of tobacco control policies, we describe the framework of the International Tobacco Control Policy Evaluation Project (ITC Project), which incorporates a number of the design features that are discussed here (Fong *et al.*, 2006a; Thompson *et al.*, 2006).

The importance of pre-evaluation knowledge in the design of evaluation of policies

The planning and design of evaluation efforts should be the first step in the process of formulating and implementing a policy (or any kind of intervention). This suggestion is part of the recommendations for "best practices" that the US Centers for Disease Control and Prevention created for tobacco control programmes in 1999. They strongly recommended that 10% of the total budget for a comprehensive tobacco control programme be allocated for evaluation and surveillance efforts associated with the programme (1999a).The WHO EURO Working Group on Health Promotion Evaluation made a similar call for resources for proper evaluation (Rootman *et al.*, 2001).

Planning should first identify the constructs that are theorized to be affected by the policy being evaluated (i.e. outcome variables and mediators), as well as those that could influence the strength of the impact of policies on those outcome variables and mediators (i.e. moderators). The choices of which constructs to include in an evaluation study come from this process. This Handbook provides descriptions of the constructs, and their measures, for many of the Framework Convention on Tobacco Control (FCTC) policy domains.

Identification of other possible events that might act as confounding factors (e.g. other tobacco control policies being implemented and programmes in operation, tobacco industry initiatives) should also be addressed in the planning stage. Knowledge of possible confounders may allow additional variables to be measured or design features to be incorporated, so that the evaluation of the policy can explicitly take them into account.

Causality

Ultimately, the goal of scientific inquiry is to attempt to identify causal relationships. The concept of cause has challenged and vexed philosophers and scientists alike through the centuries. The seminal work of epidemiologists, such as Doll and Hill (1950,1954), Wynder and Graham (1950), and Levin *et al.* (1950), on the association between smoking and lung cancer, stimulated the thinking about identifying criteria that would be used in the determination of causality in epidemiology. This influential work was the basis of the US Surgeon General's Report of 1964, and was summarized in several articles including one by A. Bradford Hill (1965). We have adapted the original nine considerations of Hill, in assessing the strength of evidence, into seven criteria concerning the possible causal impact of a tobacco control policy:

- Consistency of observed associations across studies and populations
- Magnitude of the reported association
- Temporal relationship between intervention and change in target outcome
- Exposure-response gradient
- Biopsychosocial plausibility

[1]This assumes that the common variance of the multiple measures of the construct perfectly capture the latent variable that the measures are intended to capture.

- Coherence of results across other lines of evidence
- Evidence that this type of intervention can have effects on other comparable outcomes (e.g. other behaviour patterns).

From criteria for causality to research design: the framework of Cook and Campbell

Cook and Campbell's (1979) seminal treatise on the relationship between research design of a study and the strength with which a causal relationship might be ascertained, is our starting point for a discussion of how design features can be employed to evaluate the impact of population-level tobacco control policies.

Central to the Cook and Campbell framework is the concept of *validity*. Cook and Campbell defined four kinds of validity that are critical in assessing the validity of a causal statement: *construct validity*, *external validity, statistical conclusion validity,* and *internal validity.*

Construct validity refers to the extent in which a measure captures the construct that it is intended to assess. An issue that arises in considering construct validity is the method of measurement and whether there exists a close or distant relationship between those measurements and the construct. In the area of tobacco control, examples include: Is cotinine a valid measure of exposure to tobacco smoke? Is the Fagerstrom Test for Nicotine Dependence (Heatherton *et al.*, 1991) a valid measure of nicotine dependence? What are the most valid measures of perceived risk among smokers? These basic measurement issues must be dealt with in order for the validity of a causal inference to be addressed with any substance or meaning. Sections 3.1 to 3.3 of this Handbook review the construct validity of measures to assess the effectiveness of tobacco control policies.

External validity, also known as *ecological validity,* refers to the extent in which the conclusions of a given study are maintained across different persons, settings, treatments, and outcomes (Shadish *et al.*, 2002). External validity considers issues such as whether a phenomenon studied in a laboratory setting, often involving university undergraduates, will be obtained in a "real-world" environment, which includes individuals from the general population. However, in the public health realm, two issues of external validity (whether or not the issue is expressed in these terms) arise. First, there is the importance of sampling. In evaluating a tobacco control policy being implemented in a large and diverse population (e.g. in an entire country), probability sampling methods will provide the best assurance that the study sample will be representative of the population from which the sample has been drawn and to which the intended intervention is directed. To the extent that a sample deviates from a representative sample, the external validity may be correspondingly reduced; however, it should be noted that this conclusion is not automatic. It may be that the way in which a sample deviates from the population is not (strongly) associated with the variables being analyzed; thus, the net impact may not be as great as might have been expected.

Another way in which external validity applies to the evaluation of policies and interventions is in the distinction between efficacy and effectiveness (the former referring to a treatment effect in a controlled context, and the latter referring to the effect of that same treatment in a more "real world" setting). In general, effectiveness is lower than efficacy. Interventions originally developed and tested in highly controlled experimental settings are often not as effective when implemented in the real world. This necessitates changes in an intervention when brought into real world settings in order to maintain its effectiveness, as in the more controlled settings.

The two types of validity described above set the stage for the next two forms, which deal with the relationship between two variables and whether the measured association is indicative of a causal relationship. For simplicity, our discussion revolves around whether there is a causal relationship between two variables, although the logic applies to relationships among more complex sets of variables.

Statistical conclusion validity refers to whether there exists a statistical association between the two variables. Issues surrounding

the consideration of statistical conclusion validity include: statistical power, assumptions of the statistical tests being employed, the inflation of Type I error rates due to the conduct of multiple statistical tests, unreliability of measures, as well as the selection of "appropriate" covariates/control variables in estimating the relationship between the two variables. Though correlation is important and necessary, it is not sufficient to imply a relationship for causation, as captured in the dictum "correlation does not suffice to establish causation".

Internal validity refers to the extent to which the study's design is rigorous enough to support the conclusion that the statistical relationship between two variables is due, at least in part, to a causal relationship. Here we focus on issues of internal validity, as adding design features to a study (e.g. a control group) is largely prompted by the objective of increasing the internal validity of the study. The most relevant threats to internal validity in the evaluation of tobacco control policies are presented in Table 2.1.

Basic study designs and features

We now proceed to a description of aspects of an evaluation study, and make a distinction between study design and a study feature.

The *study design* is the structural aspect of an evaluation study, defined by three dimensions:

1. *Who* the study is collecting measurements from relative to the policy that is being evaluated. Some evaluation studies only measure the impact of the policy by collecting measurements from those who were exposed to the policy; other evaluation studies, however, measure the impact by also collecting parallel measurements from those who were NOT exposed to the policy.
2. *When* the measurements were collected relative to the policy's implementation. Some evaluation studies only collect measurements after the policy was implemented; others collect measurements both before and after the policy was implemented.
3. *How many* measurements are collected. Evaluation studies vary in the number of measurement time points, ranging from a pre-post design involving one pre-policy and one post-policy time point, to a time series design involving many measurements over time.

A further design parameter arises in evaluation studies involving more than one measurement over time; that is, whether those multiple measurements are obtained on the same individuals (the longitudinal or cohort design) or on different individuals (the repeat cross-sectional design).

In contrast, a *study feature* is a non-structural aspect of a study whose inclusion will enhance the ability to address threats to internal validity. One such feature is the inclusion of multiple measures within the domain of the policy that is being evaluated, toward the goal of achieving convergent validity (multiple measures of the same construct should be related to each other). For example, in a study of the impact of graphic warning labels, we would have greater confidence that there was a causal impact of the labels if, after being exposed to them, smokers were significantly more likely to: (1) self-report that the warnings made them think about the health risks of smoking, (2) more likely to call a quit line, and (3) more likely to cite the warnings as a reason for seeking assistance for quitting, than if only one of these measures was included in the study.

Another study feature is the inclusion of measures that are relevant to some other policy that is NOT being evaluated, as it is not changing in the study population toward the goal of establishing discriminant validity (i.e. measures of different constructs should NOT be so related to each other). In the policy evaluation context, measures of the non-changing policy should NOT show change that is comparable to that in measures of the policy under evaluation. In addition, inclusion of measures that will allow the testing of mediational models are designed to elucidate the causal pathways between the policy and an important outcome variable, such as a quit attempt. For example, in an evaluation study of graphic

AMBIGUOUS TEMPORAL PRECEDENCE: Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.

- Cross-sectional survey data are particularly vulnerable to this threat.

SELECTION: Differences in respondent characteristics between groups that could also cause the observed effect.

- For example, observed differences between countries could be due to characteristics of the inhabitants rather than to differences in policies. Cross-sectional studies are particularly vulnerable to this threat.

CONCURRENT EVENT CONFOUNDING (HISTORY): Events occurring concurrently with treatment could cause the observed effect.

- For example, observed differences between countries could be due to other events or some other intervention (e.g. mass media campaign) rather than to differences in policies. This kind of confounding also includes activities of tobacco companies, which may be covert. These other events can cause the observed effect to seem stronger or weaker, positive or negative, compared to the policy/intervention's "true" effect. Concurrent event confounding could occur in longitudinal (cohort) studies, as well as in cross-sectional studies.

TEMPORAL TREND CONFOUNDING (MATURATION): Naturally occurring changes over time could be confused with a treatment effect.

- For example, trends over time occurring prior to the policy being evaluated, that are unrelated to the policy, could mimic the expected impact of policy or an adverse impact of policy (e.g. bar revenues dropping prior to the implementation of the policy could be the cause of a decrease in bar revenues observed after a smoke-free law compared to before the law).

ATTRITION: Loss of respondents to treatment or to measurement can produce artefactual effects if that loss is systematically correlated with conditions.

- Artefactual effects due to attrition can occur in cohort surveys of different groups (e.g. countries) where the attrition rate varies across the groups, and that attrition is linked to the outcome variable either directly or indirectly, via its linkage with an important predictor of that outcome variable. Related to attrition is non-respondent bias, in which non-respondents in an evaluation study could be differentially affected by the intervention (e.g. the very disadvantaged, who may be missed by both the intervention and its evaluation). Note that attrition effects in cohort surveys and selection effects in cross-sectional studies both involve biases in the sample that could lead to artefactual effects.

CONDITIONING (TESTING): Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect.

- An example of this threat is the presence of time-in-sample effects in cohort studies: participation in prior waves of a survey change the responses at the current wave (e.g. knowledge items, if repeated, can lead to observed higher levels of knowledge because of taking part in prior surveys).

Table 2.1 Selected Threats to Internal Validity and Examples

warnings, confidence that the introduction of graphic warning labels was responsible for an increase in quit line calls, rather than a mass media campaign, would be greater if there were measures included of the mass media campaign (e.g. recall measures of the campaign), and that these measures were not correlated with the likelihood of quit line calls.

In short, the internal validity of an evaluation study can be increased by including multiple measures of the policy, or other intervention, that is hypothesized to be responsible for the policy's impact, as well as measure(s) of other possible causes.

Designs for evaluation studies

In considering designs, we use the terminology of Cook and Campbell (Cook & Campbell,1979; Shadish *et al*., 2002) in which X stands for the treatment/policy that is being evaluated (e.g. introduction of graphic warning labels, increase in taxation, smoke-free legislation), and O stands for an observation (e.g. a survey data wave, quarterly report of cigarette consumption, or a set of data gathered by an air quality monitoring device).

Designs without control groups

The one-group posttest-only design:

In this design, the researcher has conducted one post-policy observation on some relevant unit of analysis. For instance, the unit could be human respondents to a survey, consumption figures from an economic database, or a venue at which the levels of respirable suspended particulates are being measured. The diagram of this design is as follows:

$$X \quad O_1$$

O_1 occurs after the policy X has been implemented.

In this post-only design, there is no sense of what the observations would have been in the absence of X ; therefore, this design alone is very poor. It does not defend against any of the threats to internal validity except ambiguity about temporal precedence. The history effects, and all threats associated with changes over time, are uncontrolled.

Given that none of the threats to internal validity are dealt with in this design, its value for evaluating policies, or interventions of any kind, is low. And yet it should be noted that the absence of a pre-test in this design often arises when the need for evaluation is recognized too late for a proper pre-test to be planned and implemented. This highlights the need for evaluation strategies to be established well before the intervention is applied, as discussed earlier.

In an effort to estimate the impact of X, researchers sometimes ask post-only respondents to recall their behaviour, opinions, or attitudes prior to X, or to make a judgment as to how X has affected them since. One should be cautious about the findings of studies relying solely on such strategies, as considerable experimental and survey evidence has demonstrated that such recall is subject to strong retrospective biases related to the respondent's theories on how the intervention might have affected them. These recall biases can occur when the respondent remembers the past as being more similar to the present than it actually was (consistency bias). When asked to estimate whether an intervention affected them, the recall bias could be in the direction of greater *contrast* (i.e. remembering the past as being more discrepant from the present than it actually was, with the magnitude of this contrast bias being correlated with the respondent's belief about the strength of the intervention (Conway & Ross, 1984; Ross, 1989; Pearson *et al*., 1992)).

Another more promising method of amplifying the value of the one-group posttest-only design is to incorporate data about pre-policy observations that are available from other sources. For example, if a new tobacco surveillance survey were created after a tobacco policy had been implemented, incorporating prevalence data from *other* surveillance surveys conducted prior to the policy would offer some comparison with a pre-policy measurement. The adequacy of this strategy would depend on the similarity between the two surveys (e.g. sampling,

method of measuring the outcome variable(s)).

The one-group pretest-posttest design:

This design adds a pre-policy observation to the previous design, and is denoted as follows:

$$O_1 \quad X \quad O_2$$

Here the addition of the pre-policy observation allows the computation of the difference score, $O_2 - O_1$, some portion of which might be causally attributable to the intervention X. The presence of an explicit measurement of the pre-post difference makes this far superior to the post-only design.

This design is considerably better than the one-group posttest only design. There is an explicit measurement prior to the policy that is not inferred or reliant on the validity of a respondent's memory or estimate of effect. The O_1 acts as a control against which the post-policy measurement O_2 can be assessed. In a repeat cross-sectional design, when O_1 and O_2 are taken from different samples in the same population, the control exists at the level of the group. In a cohort design, when O_1 and O_2 are measured from the same individuals, there is an additional level of power: each individual acts as their own control. Thus, response tendencies (e.g. the tendency to use the high end of a response scale, or to agree with survey questions (also known as acquiescence bias)) are controlled for at the individual level. This leads to greater statistical power, and the magnitude of this increased statistical power is a function of the extent to which individuals' responses at O_1 and O_2 are correlated.

Multiple pretest-multiple posttest design:

This design extends the single-group pretest-posttest design by the inclusion of additional pretest measurements and multiple posttest measurements within the group that received the policy/interventions, as in this example with 3 pretest and 3 posttest measurements:

$$O_1 \; O_2 \; O_3 \; X \; O_4 \; O_5 \; O_6$$

With many time point measurements, this design becomes a time series design. Variations within this multiple time point model include multiple pretest-single posttest and the single pretest-multiple posttest designs. These designs provide opportunities for assessing the impact of policies/interventions on the time related trends in the outcome variable that are unrelated to the policy, but which without knowledge or measurement of those trends, would bias the measurement of the policy's impact. When present, time related trends constitute an important confounding factor against which the effect of the policy must be evaluated. An example of the importance of taking into account these time related trends is presented later in this section.

In addition, designs with multiple measurements over time allow the evaluation of policies/interventions whose intensity varies over time, permitting the possibility of correlating intensity of intervention (e.g. measured by programme expenditures) with its corresponding impact. An example of this approach was used in studies evaluating the California Tobacco Control Programme, which distinguished between three time periods characterized by different levels of program intensity: pre-programme, early programme, and late program (Pierce *et al.*, 1998a).

Designs with a separate control group but with no pretest

Posttest-only design with non-equivalent groups:

In this design, a control group is added to the one-group posttest-only design. This design can be utilized if the evaluation process started too late to conduct a proper pretest measurement. If individuals were randomised to conditions, the groups would be "equivalent" on average, as randomisation equates groups with respect to all features of the individuals being measured. However, in the evaluation of national-level tobacco control policies, or in other cases where

the unit of intervention is a jurisdiction or organization, there is no possibility of randomisation, and hence, no possibility of equating groups[2]. The resulting design is the posttest-only design with nonequivalent groups:

$$X \quad O_1$$
$$\quad\; O_2$$

Case-control studies fall into this category, and often include various procedures to enhance the possibility of causal inferences, such as methods for matching the two nonequivalent groups. Issues surrounding these methods are well-identified in the epidemiological literature (Rothman & Greenland, 1998), but it should be noted that some of them, although possible with medical records among patient populations, may not be possible for implementation in evaluation studies of national-level policies.

Pretest-posttest designs with a control group:

This design is the basic "quasi-experiment" in which the pre-post measurement of the group that received the policy is compared to another group that did not receive the policy:

$$O_1 \quad X \quad O_2$$
$$O_3 \qquad\; O_4$$

The quasi-experimental design combines both elements that were used to enhance the internal validity of the one-group posttest design; added is a longitudinal component and a between-groups component. In this design, the critical starting point for an assessment of the causal impact of *X* is the construction of a multiple difference score; the change over time of the intervention group is compared to the change over time of the group that was not exposed to the intervention. The expectation, if the policy was effective, is that the pre-post difference in the policy group will be greater than the pre-post difference in the non-policy group.

The internal validity of the quasi-experimental design, although generally greater than the single group pre-post design, is dependent on the extent to which the non-policy group is similar to the policy group (e.g. similar levels of economic development, tobacco use prevalence). The greater the similarity, the more reasonable the comparison will be.

Randomisation to conditions is impossible in studies of policies. The strategy of strengthening an evaluation study via control groups depends on the selection of those control groups and their similarity. Various strategies can be used to enhance the selection of control groups that are objectively similar to the policy/intervention group on dimensions that matter (e.g. smoking prevalence, socio-economic status, similar levels of tobacco control intensity prior to the policy/intervention that is being evaluated in the study).

It would be more reasonable, for instance, to compare the impact of graphic warnings in Canada to a control group in the USA than to a control group in Bangladesh. It should be noted also that the "similarity" is not limited to the characteristics of the group. Relevant concurrent events should also be similar in the two countries. If, for example, the impact of graphic warnings in Canada were compared over time with a control group in the USA, but during that time between the pre- and post-policy measurements there was a large decrease in taxes in the USA, but not in Canada, the test of the graphic warnings would be confounded by the fact that the control group had changed in ways that would mimic the hypothesized impact of the warnings. Although the discrepancy of the difference scores would be consistent with the

[2] It should be noted that even in a fantasy world where people are actually randomly assigned to live in two different countries, one of which implemented a policy that the other did not, the randomisation would simply equate the personal characteristics of the respondents across the two groups. On average, the two countries would be populated by people who were equal on age, gender, age of initiation, number of past quit attempts, attitudes about the tobacco industry, etc. But left uncontrolled, would be the concurrent events that might occur along with the intervention that was being evaluated. The randomisation of people would offer no assistance for eliminating the possibility that observed differences between the two countries was due to differences in concurrent events. This demonstrates the limitations of randomised trials in the real world, even if such were possible.

conclusion that the graphic warnings had a desirable impact, the pattern of the data could also be explained by a significant unfavorable change in the difference score in the US control group due to the decrease in taxes.

This example points out that the structural features of the design endow an evaluation study with the *potential* for teasing apart possible alternative explanations, but that full realization of this potential is found in the selection of measures and analytic strategies that are designed to test for the *causal mechanisms* that underlie an observed difference between a policy group and a non-policy group. These strategies are described below in the section on mediation.

Threats to internal validity and methods for reduction

Having described some of the basic designs and strategies used in evaluation studies, we now proceed to a discussion of the threats to internal validity and methods for reducing them. As mentioned earlier, the rigor of an evaluation study is not only found in its design, but also in the features added to a study to enhance its power and internal validity. Examples are provided below.

Ambiguous temporal precedence:

A necessary, but not sufficient condition for causality is that a cause must precede the effect. The temporal priority condition provides challenges to cross-sectional studies by measuring possible causes and effects at the same point in time. It should be noted, however, that the temporal priority condition refers to the temporal ordering of the underlying constructs that are being measured, rather than the temporality of the data collection or observances per se.

In most cases, it is relatively simple to establish that the policy precedes a measurement. Even in a posttest-only design, temporal precedence is established: the measurement followed the implementation of the policy. However, because the key question is whether the evaluation measure changed as a result of the policy (i.e. whether the policy caused a change in the evaluation measure), the single measurement made in the posttest-only design is insufficient even as the temporal precedence condition is satisfied.

This discussion highlights the importance of multiple time point studies in assessing the causal impact of a policy/intervention, and is illustrated in greater detail below.

Selection: systematic differences over conditions in respondent characteristics that could also cause the observed effect:

Selection bias refers to the fact that individuals in different groups (e.g. different states, provinces, countries) are non-equivalent; that is, they could differ on dimensions that are correlated with the outcome measures used for the evaluation of the policies. Selection biases are difficult to identify and eliminate. Randomisation to conditions of an experiment is a powerful method for equalizing potential biases due to the non-equivalence of characteristics of individuals. However, randomisation is not possible in studies evaluating national-level tobacco control policies; therefore, selection bias in some form remains in all evaluation studies.

One approach to dealing with selection bias within a given evaluation study is to select control groups that are as similar as possible to the policy group. Thus, in evaluating the impact of policies in Canada, using the USA as a non-policy control group would be advantageous, as they are quite similar on many cultural and societal dimensions. If a policy in Canada were evaluated using, say, Kenya, as a control group, the inherent differences in the two countries would be much greater, leaving room for many more confounding factors.

A second approach is to measure differences between countries on constructs that might vary and act as possible confounding factors in the evaluation of policies. For example, in evaluating a policy in China compared to the USA, a possible confounder might be the fact that China is known to be a more collectivistic society, while the USA is a more individualistic

society. Knowing this difference, the evaluation study could add a measure of individualism-collectivism (Triandis & Gelfand, 1998), and correlate this variable with the policy-relevant variables in each country. If individualism-collectivism was uncorrelated with the policy-relevant variables, then this would suggest that, even though the two countries differed on this, it was not correlated with the policy and thus could not be a viable alternative explanation for observed policy impact.

The third approach considers multiple evaluation studies of the same policy in different settings and different times (i.e. of the overall consistency of the effects). This is adopted from one of Hill's criteria. If graphic warning labels are found to be effective in motivating individuals to quit smoking in Canada, Thailand, Venezuela, Brazil, and Belgium, then our confidence increases in making a general conclusion about the causal impact of graphic warning labels. Making general conclusions about policy impact will not and cannot occur on the basis of a single study, but rather after the consideration of multiple studies across multiple countries and time points. This principle is not limited to the evaluation of tobacco control policies.

It is worth noting that lack of consistency across studies provides an opportunity to examine what factors might be responsible for that variance. It may be that studies with weak designs yield different conclusions than those with stronger ones. In tobacco research, it has been shown that tobacco industry-funded studies of secondhand smoke are much more likely to conclude that it is not harmful, which is at odds with the very large number of non industry-funded studies concluding that secondhand smoke is harmful (Barnes & Bero, 1997,1998; for review, see Bero, 2005)

History: events occurring concurrently with treatment could cause the observed effect:

The internal validy of studies that evaluate the impact of policies over time, is threatened by events occurring concurrently with treatment/target policy which could cause the observed event. It is often the case that one treatment/policy intervention is implemented in conjunction with other policies/initiatives relevant to tobacco control. There are often other events, programmes, and interventions that are ongoing at the time of the policy that is being evaluated. Therefore, a major challenge is to estimate the impact of a specific policy in the field of other interventions that are ongoing simultaneously.

This is likely a common occurrence. If a government launches a comprehensive tobacco control programme, a frequent and recommended strategy would be to implement multiple policies and interventions. This comprehensive approach might include mass media campaigns, higher taxation, advertising/ promotion/-marketing restrictions, bans, increased resources for cessation programmes, and/or campaigns to raise awareness of existing cessation programmes.

For example, in 2003, countries of the European Union implemented new tobacco-use warnings, which were prominently displayed covering 30% of the package area. This corresponded with the minimal standard of warning labels under the Framework Convention on Tobacco Control (FCTC). The ITC Four Country Survey was launched in October 2002, in order to collect the pre-policy data for evaluating the impact of this enhancement of the warning labels. In May 2003, the second wave was conducted in the same manner as the first post-policy data collection.

By the time of the second survey, another important tobacco control policy had been put into action. In February 2003, the United Kingdom implemented a comprehensive ban on advertising and promotion of tobacco-related products, via billboards, magazines and newspapers, direct mail, domestic sponsorship (May 2003), website advertising and promotions, and exterior signs in store windows. This second policy complicated the quest for measuring the impact of the enhancement of the European Union's warning labels. Below, we outline an empirical strategy for distinguishing the effects of different interventions.

Factors that also influence the outcome measures of an evaluation study of a specific

tobacco control policy include activities of the tobacco industry, which are designed to reduce or neutralize the effect of tobacco control policies and programmes. Without consideration of these countermeasures (which could include explicit inclusion of industry activity variables), a policy evaluation study could lead to incorrect conclusions.

Although the importance of identifying and measuring the impact of tobacco industry activities cannot be over-emphasized, the impact of such activities will vary depending on the outcome measure. Broad, downstream outcome measures, such as prevalence rates, quit attempts, etc., are likely to be most strongly affected by tobacco industry activities. In contrast, more policy-specific outcome measures, such as label salience or the self-reported extent to which a smoker states that the warnings have made them think about the health risks of smoking, would be less likely to be influenced by industry activities. And here there is a trade-off: the measures of policy impact that are specific to that policy are less vulnerable to influence by tobacco industry counter-activity; as the measures become broader (e.g. going from label salience to perceptions of risk to intentions to quit to quit attempts), they are more vulnerable to impact from tobacco industry influences.

Maturation: naturally occurring changes over time could be confused with a treatment effect:

Typically, the term "maturation" refers to natural changes in individuals over time, such as changes that children undergo as they grow older. However, the concept might instead be called "time-dependent changes that are unrelated to the treatment." An example of how this concept must be identified and controlled for, comes from the claim made by opponents to the comprehensive smoke-free legislation in Ireland that sales volume in pubs had declined as measured before and after the March 29, 2004 ban (Figure 2.1).

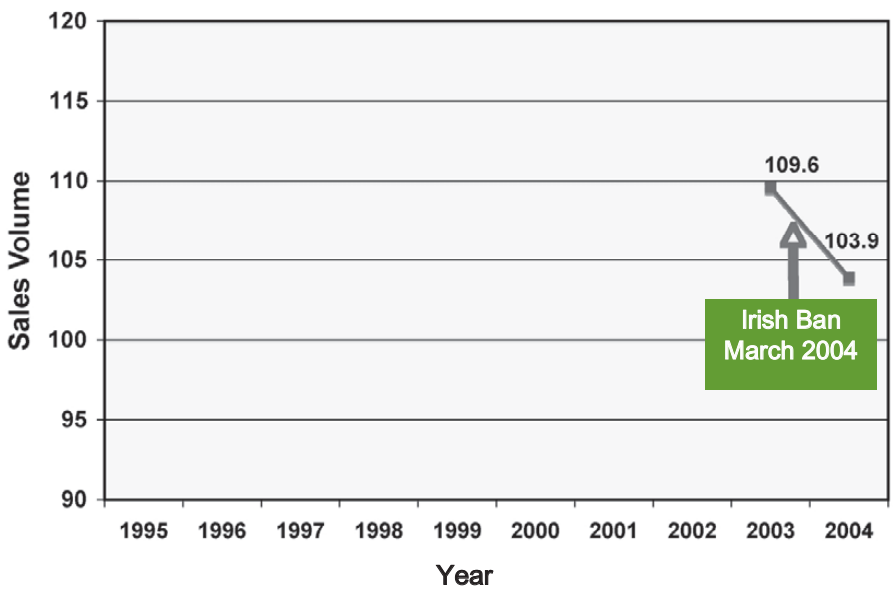


Figure 2.1 Pub sales volumes immediately before and after implementation of the Irish smoking ban in 2004

Source: Central Statistics Office of Ireland

Sales volumes are indexed so that sales volume in 1995 = 100

The data on the volume of pub sales before 2003 and after the 2004 ban, as shown in Figure 2.1, reveals that the volume of pub sales (indexed at 100 for volume of pub sales in 1995) in 2004 was lower (103.9) than it was for 2003 (109.6). With just those two data points, it might be concluded that the Irish ban caused a decline in sales in pubs.

However, Figure 2.2 presents the volume of pub sales for nine years (1995–2003) prior to the Irish ban. Taking into consideration the data from years prior to 2003 leads to a very different conclusion.

Sales volumes had been rising steadily since 1995, hit their peak in 2001, and then began to fall fairly steeply. When the full nine year profile is considered, the decrease between 2003 and 2004 does not appear to be any different than what would be expected by the secular trends. The decline between 2003 and 2004 was not significantly more dramatic than the declines experienced between 2001 and 2002, and between 2002 and 2003. When the more long-term "maturation" trends are considered, there was no greater decline after the smoke-free law had been implemented. Thus, the hypothesis that the Irish ban had a detrimental impact on the volume of pub sales is not supported.

Time trends can also work in the opposite direction. Suppose that the ban in Ireland was implemented between 1997 and 1998. If the evaluation study had been conducted with data from only those years, it would have shown an increase in sales, which might lead to the false conclusion that the ban was the cause of this increase. Again, consideration of the pre-policy time trends would reveal that the secular trend was indicative of increasing sales, and taking that trend into account would likely lead to a more proper conclusion that the ban had no impact on sales.

The implications for research design are clear: evaluating the

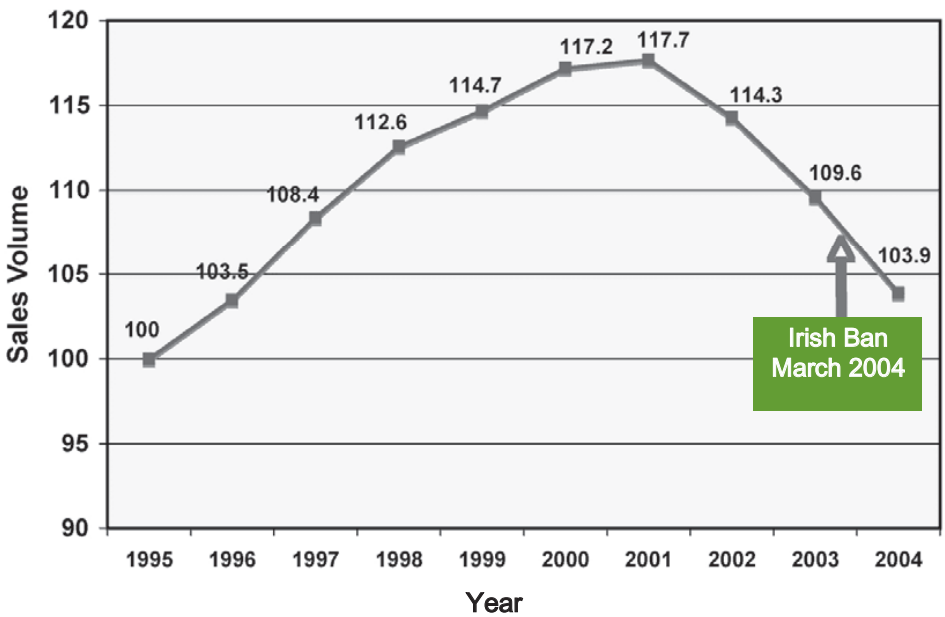


Figure 2.2 Pub sales in volumes in Ireland for the period 1995-2004
Source: Central Statistics Office of Ireland
Sales volumes are indexed so that sales volume in 1995 = 100

impact of policies is best conducted with the inclusion of data that allow the evaluation to take place within the context of time trends. This example highlights the value of having a surveillance system in place for collecting data over time on outcome variables of interest.

Although the Irish pub data illuminate the importance of time trend data, it also provides an example of how even good time trend data alone can sometimes be incapable of yielding a clear estimate of policy impact. To illustrate this, suppose the ban occurred in 2001 instead of 2003, and the evaluation was conducted with pub volume data from just 2001 and 2002. Here, consideration of the time trend might be taken to mean that the ban definitely reduced sales; however, it was still positive up to that point.

If only the time trend were taken into account, one might be even more confident of the conclusion that the ban decreased sales. However, in 2001, Ireland passed a law that limited the use of alcohol, which had an adverse impact on sales volume. Because of the presence of this known negative causal factor, the impact of the Irish smoking ban would remain ambiguous. Although time trend data are important in resolving some threats to internal validity, they fail to eliminate the threat to validity represented by concurrent events in the absence of information on the impact of such events.

A research design that is also concerned with understanding the impact of an intervention over time is the *interrupted time series* design (a specific version of this general design is the *regression discontinuity* design). In these designs, which require a fairly lengthy series of observations over time, the impact of an intervention can be measured by its impact on the mean function of the time series. In the regression discontinuity analytic framework, a distinction is made between the regression line that fits the data points (capturing the relation between the outcome variable and time) before the intervention, and the regression line that fits the data points after the intervention. The analysis compares the two lines; the effect of the intervention is measured as the difference in the slope, the intercept, or both parameters of the line. This kind of design can provide powerful evidence for the impact of a policy in its temporal context. There are a number of sources that describe these models (Trochim, 1984; Trochim *et al.*, 1991; Box *et al.*, 1994).

Time series approaches have been used in evaluating the impact of tobacco control programmes. For example, Pierce *et al.* (1998a) used piecewise regression analysis on time series data on cigarette consumption from 1983-1997 in California, versus the rest of the USA, to demonstrate that the California Tobacco Control Programme, initiated in 1989, led to declines in consumption. They also found that the impact of the programme was greater for the first five years than for the subsequent three. Biener *et al.* (2000) used similar methods to analyze prevalence data in Massachusetts versus the remaining US states (except California because of their similar comprehensive programme), and concluded that the Massachusetts programme led to a continued downward trend in prevalence, compared to the flattening of the downward trend in the other US states during that same time period.

Keeler and colleagues (1993) examined monthly time series data from 1980 to 1990 in California in their analysis of the association of cigarette prices, taxes, income, and anti-smoking regulations with cigarette consumption. Reduced consumption was found to be associated with tobacco control policies. They highlighted the impact of the tax increase in 1989, which led to a greater decline in consumption, followed by additional tax increases at other points along the time series.

In general, multiple time point data, particularly if such data are also available with control groups, provide strong potential for teasing out possible confounding due to time related alternative factors, and for providing confirmatory evidence for the impact of policies and programes. The strength of this potential (and therefore confidence in attributing changes in behaviour or some other important outcome measure) grows with the number of post-intervention data points, which means that more definitive conclusions might be reached

only after a greater delay than would be desired. The ability to come to more definitive conclusions increases with the number of other evaluation studies of a particular policy, or type of policy; within a specific (well-designed) study, the ability grows with the passage of time. Both require greater effort/time than is possible within a single pre-post evaluation study.

Attrition: loss of respondents to treatment or to measurement can produce artefactual effects if that loss is systematically correlated with conditions:

Attrition is a major concern in cohort surveys. In surveys about smoking, for example, those who quit are less likely to stay in the survey, even when specific provisions have been made for those who quit to move to a non-smoker/quitter survey, as in the ITC Surveys (Thompson *et al.*, 2006). Thus, it may be that if a policy or intervention is successful in increasing the proportion of individuals who quit, the greater attrition rate in the policy group, skewed as it is for those that quit, will attenuate the observed treatment effect (i.e. it will make the statistical test of group differences more conservative). Another potential bias due to attrition is seen in respondents with low socioeconomic status (SES), who are more likely to drop out. If the policy/intervention is more likely to have an impact on high SES individuals, the differential drop out will lead to an artificial enhancement of the treatment effect. The cumulative result of attrition will be the net effect of conservative and liberal biases, which will lead to uncertainty regarding the overall impact of differential attrition in any given survey situation.

Although attrition is unique to cohort surveys, non-response bias is a problem in cross-sectional studies, as well as cohort surveys. Non-response bias occurs when the surveyed sample differs from the population, because some types of respondents are less likely to agree to participate in the survey, or are less apt to be contacted in the first place. This poses the same problems as attrition; many factors contributing to non-response bias are present in biases from attrition.

As with all threats to validity, an approach to dealing with attrition is to measure its impact. The goal is to develop a model of the correlates of attrition that identifies variables that are associated with the likelihood of attrition and the strength of the relationship. Toward this end, it is valuable in cohort designs to replenish cohort members lost to attrition at each stage with newly recruited respondents from the same sampling frame. Differences between the responses of the cohort and the newly recruited replenishment sample can then be attributed to biases in attrition, and to time-in-sample effects, to which we turn next.

Time-in-sample: exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect:

A *time-in-sample effect* (also known as *rotation group bias*) is a phenomenon whereby an individual's responses to the same question over time varies as a function of how many times the individual has responded to the same question in the past (i.e. the number of prior survey waves the individual has participated in (Duncan & Kalton, 1987)). In a cohort survey of nutrition, respondents were systematically rotated out of the survey, so that at each survey wave there were respondents who had participated 1, 2, 3, and up to 9 times before. It was found that respondents reported eating smaller quantities of food purely as a function of the number of prior survey waves they had been administered (Nusser *et al.,* 1996). It is valuable to take into account the time-in-sample effect in the analysis of cohort data.

Additive and interactive effects of threats to internal validity: the impact of a threat can be added to that of another threat or may depend on the level of another threat:

This statement reminds us that, as with any study, there exists more than one threat to internal validity and more than one source of bias in the estimate of an intervention effect. Some of these biases may

be in the direction of overestimating the effect; others may be in the direction of underestimating the effect. The impact of one source of bias can depend on the level of a second source of bias. For example, the overall impact of participation bias over time will depend on the level of attrition.

Cost effectiveness in the design of evaluation studies

On some dimensions, study design can be guided by a calculation of costs in relation to its benefits. The allocation of total sample size to number of clusters, and number of individuals within clusters, is one example where prior information (e.g. the incremental cost of conducting the study in an additional cluster; the intraclass correlation, a measure of the correlation of individuals within a cluster compared to the correlation of individuals belonging to different clusters) can be entered into formulas to create the "optimal" sampling design given specific resources available for the study.

In principle, the same is true for designing an evaluation study to reduce threats to internal validity, that is, a study that stands to yield a more confident judgment about the causal impact of the policy/intervention. But here, however, the process cannot be guided by formula or algorithm in the same way as can be accomplished in creating an optimal sampling plan. The increment in internal validity due to the addition of a second or third post-policy time point, for example, cannot be measured quantitatively. The reason is that the actual value is dependent on knowledge of the impact of spurious causal factors. The value of the second or third time point depends on whether the other causal factors would have exerted a policy-consistent or policy-inconsistent impact, which is unknown. In fact, if we actually felt confident enough about the impact of the other causal factors to put them in such a formula, there would be little need to actually conduct the evaluation study in the first place! Even though we cannot be specific about the value of a certain design feature in an evaluation study, we can make some general statements about the likely relative value of one feature or design element over another.

As described earlier, the single-group post-only design is not sufficient for evaluation of a policy (or any other intervention). So what could be added to this single measurement? There are two basic possibilities: (1) create a one-group pretest-posttest design by adding a pre-policy measurement from the same sampling frame as the post-policy measurement: either the same individuals who will be measured at post-policy (cohort design) or other individuals (repeat cross-sectional design); and (2) create a posttest-only design with nonequivalent groups by adding a post-policy measurement from another group who is not receiving the policy/intervention.

For example, suppose a researcher is planning an evaluation of the graphic warning labels introduced in Thailand in 2005 knowing that a post-policy measurement is required. But when adding another group to the design, should this second group be a pre-policy measurement in Thailand, or a post-policy measurement in another country, such as the neighboring country of Malaysia? It is strongly recommended that a pre-intervention measurement be added. This is because the starting point for all considerations of measuring the causal impact of an intervention is in the difference between pre- and post-policy (i.e. how respondents changed from pre- to post-policy on a label-relevant variable). Having an explicit measurement of this pre-post difference is much preferred to adding a control group (Malaysia), as the researcher would still have to infer what the outcome variable would look like in the absence of the policy at a time prior to the policy's implementation. As long as there is sufficient time to collect pre-policy data, this recommendation is also the easiest to implement. In the evaluation of national-level policies, it is simpler to obtain multiple measurements within one's own country than it is to obtain the same measurements in a different country.

Thus, the single expansion would favor the addition of pre-policy measures. In addition, the logistics of setting up the parallel study (e.g. a survey) in another country, with the establishment of a second research team, and the challenges of making the two parallel research efforts com-

parable in method and measures, would be great.

Summary of study design considerations

To summarize, in the absence of a randomised trial, there are two study design strategies that can be employed for the rigorous evaluation of the effects of policies. First is the use of measurements both before and after the policy's implementation. These measurements can be taken from either units (usually, but not limited to, individuals; the same logic would apply if the measures were of households, schools, or other venues) that are either the same (as in a cohort design) or different, but drawn from the same sampling process (as in a repeat cross-sectional design). The second design strategy is the use of a quasi-experimental design, in which one group that is exposed to a policy is compared to a similar unexposed group, as discussed above. Combining these two strategies in a single study yields a two-group, pre-post design, which offers a higher degree of internal validity than either feature alone. The utility of longitudinal designs is strengthened if there are multiple data collections before and/or after policy implementation, allowing more precise specification of effects (e.g. taking into account temporal trends that were occurring before the implementation of the policy).

Considerations of study features in the evaluation of policies

We have made a distinction between study designs and study features. In addition to the two design considerations, there are two study feature strategies that contribute to increasing an evaluation study's internal validity. The first is the measurement of policy-specific variables that are theorised to be affected initially after the policy is implemented. For example, in evaluating the impact of a new warning label policy on behaviour, one might reasonably predict that for the policy to exert its effect on behaviour, the target population must first report noticing the new warning labels (Hammond *et al.*, 2006). A second strategy is the measurement of policy-specific variables for policies that have not changed; such variables act as another form of control. In a country where labels have been enhanced and where taxation has not, for example, we would expect that label salience would be improved over time, but taxation-relevant variables (e.g. perceived cost of cigarettes) would not. Recommendations for measures in each FCTC policy domain are provided in other sections of this Handbook.

Combining the two design and two study feature strategies, along with the inclusion of other explanatory variables (covariates) that might help explain differences between two jurisdictions, creates a powerful research design allowing more confident inferences to be made about the causal effects of policies and/or combinations of policies. We now turn to an illustration of the use of these strategies in the International Tobacco Control Policy Evaluation Project.

The International Tobacco Control Policy Evaluation Project (ITC Project)

The ITC Project was established with the goal of measuring the psychosocial and behavioural impact of key policies of the FCTC on tobacco use among adult smokers (Fong *et al.*, 2006a; Thompson *et al.*, 2006). As smokers are directly affected by tobacco control policies, this understanding is crucial to assessing the extent to which the FCTC objectives are met, and of desirable and undesirable collateral effects. The ITC Surveys were explicitly shaped by the four strategies described above. To date (as of December 2007), the ITC Surveys are a set of parallel prospective cohort surveys of representative samples of adult smokers in 15 countries—Canada, USA, UK, Australia, Ireland, Thailand, Malaysia, South Korea, Mexico, Uruguay, France, Germany, The Netherlands, New Zealand, and China, with additional ITC Surveys under development in other countries (Bangladesh, India and Bhutan).

With these additions, the ITC project will be conducting

evaluation of FCT policies in countries inhabitated by over 50% of the world populations, 60% of the world smokers, and 70% of the world's tobacco users.

The ITC evaluation framework utilises multiple country controls, a longitudinal design, and a pre-specified, theory-driven conceptual model to test hypotheses about the anticipated effects of specific policies.

Conceptual model of the ITC Project:

The first step in creating the ITC Surveys was to determine how policies may achieve their desirable effects. How do policies work?

In order to address this important issue, a couple of assumptions need to be described. The first is that the most appropriate level of analysis, to understand the mechanisms by which policies may ultimately change public health outcomes, is that of the individual person. It is the individual who smokes or does not smoke, the individual who is influenced by anti-smoking media campaigns or by marketing campaigns of the tobacco industry, the individual who is or is not influenced by societal norms or by influences from close friends and family, and the individual who does or does not form intentions to quit and then either does or does not engage in an attempt to quit.

Having said this does not preclude the possibility, indeed the reality, that the individual can be influenced by forces at broader levels of analysis (e.g. social structure and organization), and by factors at even finer levels of analysis (e.g. individual differences of genetic susceptibility, such as high versus low metabolism for nicotine). Ultimately, however, it is individuals whose behaviour will or will not be influenced by policies, and in order for us to understand these behaviours, we must focus on the individual.

The second assumption is that there exists a causal chain of changes within the individual *through which the impact of policy flows*. This assumption directly relates to the idea of mediation: that policy causes changes in one or more constructs, and/or a chain of constructs within the individual, which then eventuates in behavioural change. The ITC Project team created a conceptual model of how tobacco control policies might work based on a combination of existing models from the psychosocial literature and from health communication theories. The resulting conceptual model, which is presented in Figure 2.3, guided the selection of questions included in all ITC Surveys.

The ITC conceptual model assumes that each policy ultimately has an influence on behaviour through a specific causal chain of psychological events. It is a general framework for thinking about policies and their effects on a broad array of important psychosocial and behavioural variables, and for testing how policy distinctions relate to their effectiveness. Several key characteristics of this conceptual model require further explanation. First, the model focuses on how policies affect the behaviour of individual smokers, and thus circumvents the potential hazards of making inferences about individuals from aggregates (i.e. policy studies in which countries are the unit of analysis, or individual-level studies that are repeat cross-sectional analyses conducted over time).The presence of macro-level causal forces that exert pressure on an individual, are acknowledged in the ITC conceptual model. For example, societal norms toward smoking, economic conditions, messages from the media that are either pro- or anti-tobacco use, and the influence of family and friends are taken into consideration. The model specifies, however, that the impact of those macro-level causes must be measured at the level of the individual through their perceptions of the presence of such factors (e.g. beliefs about the norms and expectations of society, close friends, and family on smoking). In the end, it is the individual who takes up smoking, who increases or decreases tobacco consumption, who does or does not attempt to quit, who is successful or unsuccessful at attempting to quit, and who may contract a smoking-related disease and die. Of critical importance, and a focus in the ITC conceptual model, is to capture and measure the influences of the many macro-level causes as

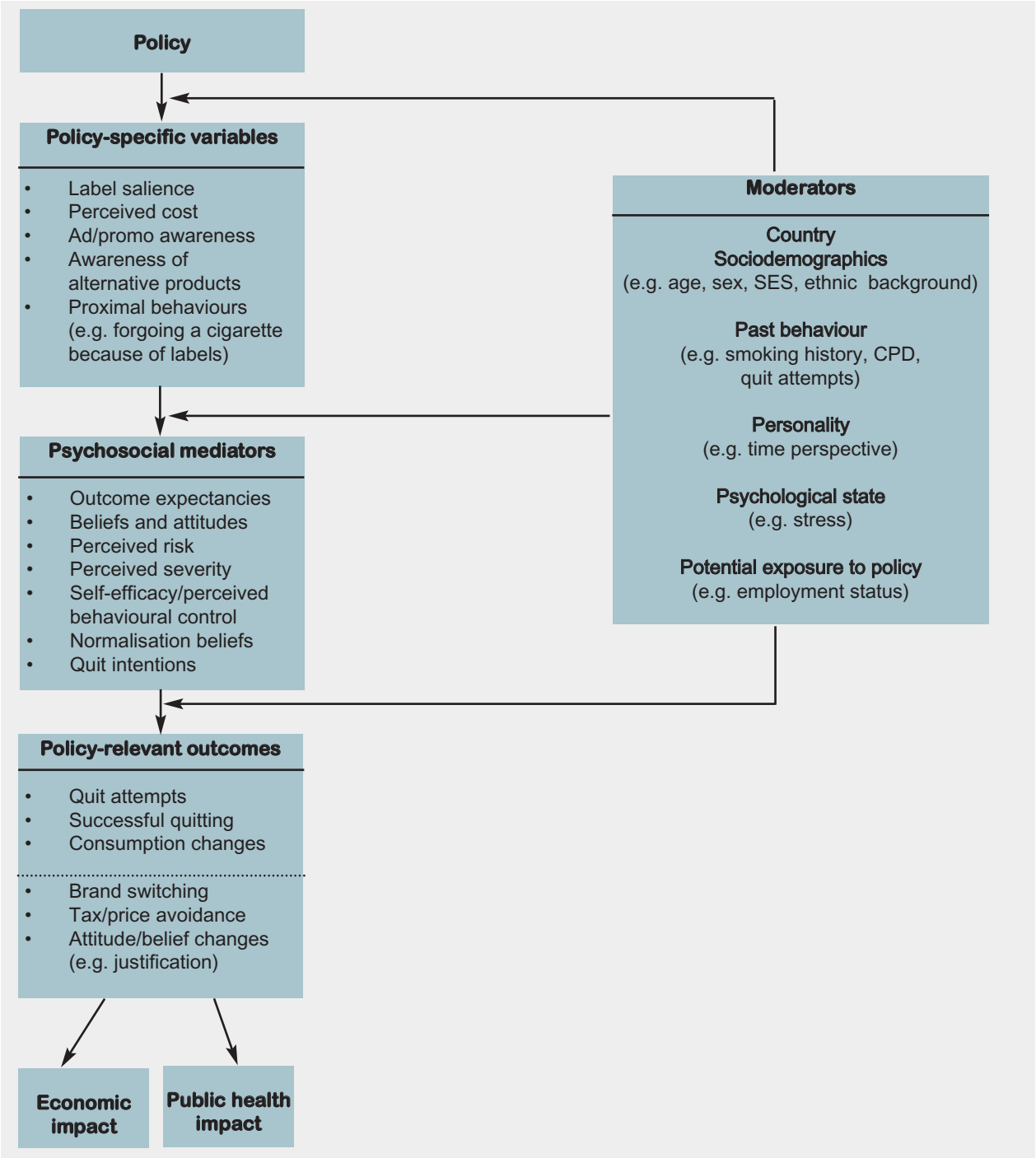


Figure 2.3 Conceptual model guiding the formulation of questions in the ITC Surveys
Adapted from Fong *et al.*, (2006a)

experienced by the individual. Ultimately, in order for us to understand the impact of policies and other macro-level influences on populations, it is essential to measure them at the individual level. It is a fallacy that the presence of macro-level causal forces requires that macro-level modelling be conducted.

Second, policies are seen as potentially affecting individuals along a variety of psychosocial and behavioural variables, of which there are two classes. The most immediate effects are those on the *policy-specific variables* (those variables that are proximal (conceptually closest), or most specifically related to the policy itself). Thus, new graphic warning labels should increase salience and the ability to notice warnings; price should affect perceived costs of cigarettes (for example, belief that cigarettes have become too expensive); and lifting of restrictions on alternative nicotine products should lead to increased awareness of the availability of those products. These effects may also increase the likelihood of discrete behaviours specifically linked to the manifestations of the policy such as smokers hesitating, or even forgoing or stubbing out cigarettes because of the warning labels. Examples of survey questions designed to measure policy-specific variables are presented in Table 2.2. Other sections of this Handbook describe these and other measures of policy-specific variables in each of the FCTC policy domains.

The more downstream effects are on the non-specific *psychosocial mediators*, which are conceptually distant from the policy and theorised to be affected by multiple influences, not just policies. Among these are variables such as self-efficacy and intentions, which come from well-known psychosocial models of health behaviour, including the theory of planned behaviour (Ajzen, 1991), social cognitive theory (Bandura, 1986), the Health Belief Model (Becker, 1974), and Protection Motivation Theory (Rogers & Prentice-Dunn, 1997). The ITC conceptual model holds that policies will affect these general mediating variables indirectly, through their prior effects on the policy-specific variables. As each policy has its own policy-specific variables, there exists potential to estimate the relative contributions of various policies to the outcomes of interest.

Third, the ITC conceptual model explicitly identifies the mediators of policy and articulates the goal of understanding the psychosocial processes that explain how and why a given policy may lead to changes in smoking behaviour. The longitudinal design allows the explicit testing of the causal chain of effects that is depicted in the model. With a repeat cross-sectional design, the capabilities of modeling the dependence of change in an outcome on the changes in an explanatory variable are diminished as data on the same individuals are not collected prospectively.

The policy-relevant outcomes that are measured in the ITC surveys include those that confer public health benefits (for example, quitting), but also include important compensatory behaviours that the smoker may engage in that, although responsive to the policy, may not lead to the economic and public health benefits that are ultimately the goal of such policies. For example, smokers may switch to discount brands in response to price increases, which would confer no public health benefit. The ITC Project thus attempts to provide a more complete account of the effects that may result from the implementation of a tobacco control policy, and includes both the detection of desirable effects and of unintended, undesirable side effects.

In summary, the ITC conceptual model is a causal chain model, and, as such, suggests that the policy-specific variables play a critical mediating role because they reside between the policy and the outcome variables that are important in public health (e.g. quitting behaviour). These causal paths, from policy-specific variables to behaviour, could be direct, but more typically will be through the more general mediators. In some cases, there may be pathways through several kinds of mediators, both the policy-specific, proximal variables, and the more general, distal variables. Policies are theorized to vary in the psychosocial ‘‘routes’’ that they take to affect behaviour, that is, each policy has a different

Policy Domain	Examples of Questions Measuring Policy-Specific Variables
Warning Labels	In the last month, how often, if at all, have you noticed warning labels on cigarette packages? Warning labels make me think about the health risks of smoking (level of agreement or disagreement with this statement)
Smoke-Free Legislation	Which of the following best describes the rules about smoking in drinking establishments, bars, and pubs where you live? – Smoking is not allowed in any indoor area – Smoking is allowed only in some indoor areas – There are no rules or restrictions For each of the following public places, please tell me if you think smoking should be allowed in all indoor areas, in some indoor areas, or not allowed indoors at all? – Hospitals – Workplaces – Drinking establishments (e.g. pubs/bars) – Restaurants and cafés
Price/Taxation	Where did you last buy cigarettes for yourself? How much did you pay for your cigarettes? The last time you bought cigarettes for yourself, did you buy them by the carton, the pack, or as single cigarettes? The last time you bought cigarettes or tobacco for yourself, did you use any coupons or discounts to get a special price?
Pro-Tobacco Advertising	In the last 6 months...how often have you noticed things that promote smoking? In the last 6 months, have you noticed cigarettes or other tobacco products being advertised in any of the following places: television, radio, at the cinema/movie theatre before or after the film/movie, on posters or billboards, in newspapers or magazines, on shop/store windows or inside shops/stores where you buy tobacco? Now I would like you to think about advertising or information that talks about the dangers of smoking, or encourages quitting. In the last 6 months, how often, if at all, have you noticed such advertising or information?
Product Regulation	Do you agree or disagree with this statement about “light” cigarettes: “Light cigarettes are less harmful than regular cigarettes”?

Table 2.2 Examples of Questions Designed to Measure Policy-Specific Variables in the ITC Surveys
Adapted from Fong *et al.* (2006a)

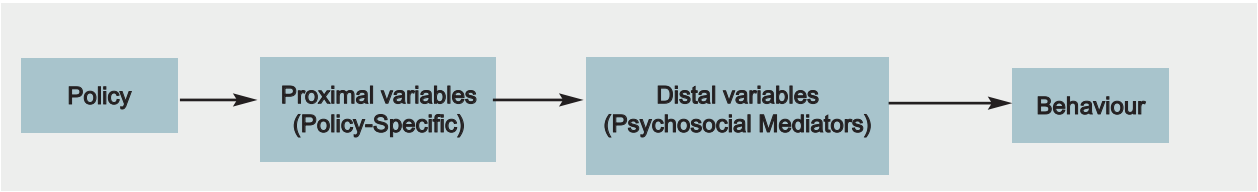


Figure 2.4 Schematic model of how a policy intervention might work (general pathway)

Figure 2.5 Schematic model of how an intervention such as warning labels on cigarettes might work

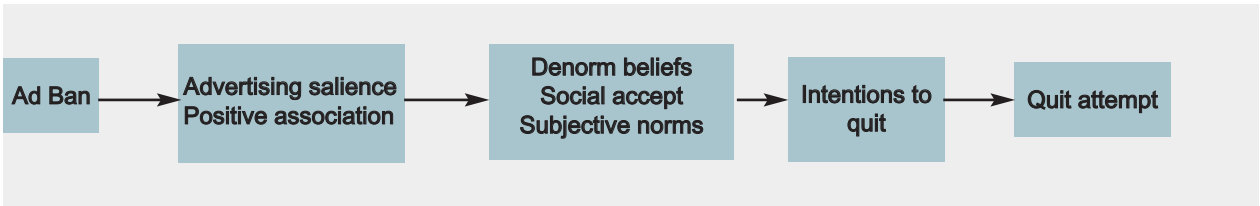


Figure 2.6 Schematic model of how an invervention such as banning of pro-tobacco advertissement might work

mediational model for how it is theorized to operate (Figure 2.4).

For example, an enhancement in warnings may first increase salience/noticing, depth of processing, and other constructs that have been identified by communication theory as being an important initial step for a communication attempt to be effective. The resulting heightened perception of the risk or hazards of smoking should affect overall attitudes and outcome expectancies, which affect intentions, which in turn affect behaviour (Figure 2.5).

In contrast, advertising bans may first decrease awareness of tobacco-favorable messages, which may lead to reductions in the perceptions that smoking is a socially acceptable behaviour, then to the idea that subjective and societal norms are more negative toward smoking, which is theorized to lead to quit attentions and quitting behaviour (Figure 2.6).

The specific articulation of these mediational models leads to specific, theory-driven empirical tests. The strategy of testing the impact of policies through mediational models of this kind differs from the approach taken in dealing with threats to internal validity. That approach, which is a process of *falsification*, uses research design and analytic tools to determine that a possible confounding factor was NOT responsible for the observed pattern of data, whereas explicit

tests of mediational models provide the possibility for confirmatory analyses, which test whether a policy had its impact on an important outcome variable because it first caused changes in a policy-relevant mediator.

In general, the design of the ITC Surveys is guided by the possibility of disentangling the web of alternative explanations and competing forces through the careful selection of specific, theory-driven mediators.

The ITC conceptual model offers an opportunity to test how policies impact or fail to impact anticipated behaviour. For example, the mere existence of a policy, even if implemented properly, does not guarantee that smokers will be exposed to its consequences in the ways anticipated. Using the example of warning labels, some smokers barely look at a pack when they are smoking and may rarely or never notice the warnings. This, however, could be due to motivated avoidance, and it is important to measure whether this has an impact on behaviour. In a cohort survey of Ontario smokers, Hammond and collaborators (2003) found that avoidance of the graphic Canadian warning labels, by means such as covering them up or by putting them in a cigarette case, was not associated at follow-up with a decreased likelihood of a quit attempt.

Additional research questions can be addressed, such as whether is it sufficient for someone merely to notice warnings or whether it is necessary to read them closely, or process them at a deeper cognitive level. And what role do microbehavioural reactions, such as foregoing a cigarette as a result of noticing/reading warning labels, play in determining longer-term outcomes, such as quitting?

In order to address these and other conceptual questions about the impact of warning labels, the ITC Surveys include multiple measures to empirically identify from the service results which measures may be important in understanding the impact of warning labels. In this regard, it should be noted that the ''best'' measure for understanding the impact of warnings may depend on whether the warning is text-based or whether it includes graphic images.

Mediational models have the potential to identify causal mechanisms, and the importance of this is that knowledge of the causal mechanisms can inform the creation of interventions of potentially greater power. Thus, the general mediation model is realized differently in diverse policy domains; different policies are mediated by different constructs. Because the ITC Surveys measure all of these constructs, it is possible to begin to distinguish whether a change of behaviour (e.g. quit attempt) was due to a given policy, in the context of other policies, or to other alternative events that occurred at the same time.

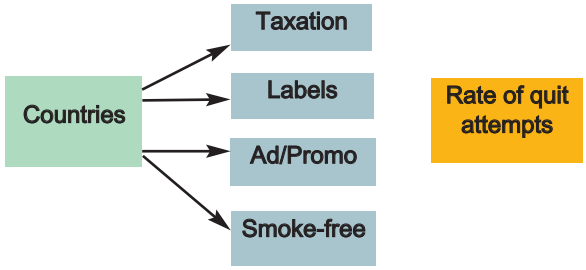
The use of mediational models as a mechanism for establishing the effect of policies:

As described earlier, an important and vexing hazard to internal validity is the concurrent events threat (also known as a *history* threat): the presence of events that occurred concurrently, such as multiple policies, or a mass media campaign that was implemented at the same time as the policy that is being evaluated. How can these threats be measured and dealt with?

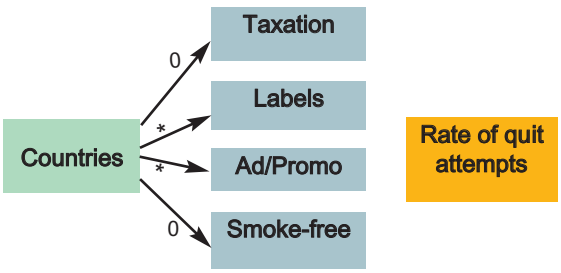
The only method of keeping possible alternative causes from becoming confounders is to measure their potential impact, and explicitly including them in a model that competitively tests their impact. For example, if a mass media campaign is being implemented at the same time as a policy to be evaluated, measures of noticing, and the impact of, that mass media campaign (see Section 5.6) could be included in a post-policy survey, and those measures used as covariates in an analysis of the impact of the policy. Although the study might originally have been conceptualized as evaluating the policy, including measures of the mass media campaign would augment the study as a simultaneous evaluation of the impact of both policy and the campaign. The general point here is that unconfounding of alternative events in the evaluation of a policy can only be attempted through the measurement of the possible impact of those alternative events.

It should also be noted that even randomisation to conditions does not eliminate the threat to internal validity posed by concurrent events. If randomisation were possible in policy evaluation

(a) Basic layout of mediational model designed to test whether any of the policies might have been causally responsible for the difference between countries in the rate of quit attempts.



(b) Between the two ITC survey waves, for each of the four policy domains, did any of the countries make a change?



(c) Between the two ITC survey waves, suppose there were two policy domains in which one country changed: Labels and Ad/Promo (starred paths from countries to those two policy domains). There were no changes over time in the other two domains. Thus, those paths are equal to zero, indicating that differences across countries in the rate of quit attemps could not have been mediated by changes in Taxation and Smoke-free policies.

(d) The reduced mediational model, having eliminated Taxation and Smoke-free policies as possible mediators

(e) We then examine the paths from each of the two policy domains (that is, the policy-specific measures for each of the domains) to rate quit attempts to test whether the change in those policy-specific measures is associated with differences in the Rate of quit attempts. We find that the Label measures are associated with the Rate of quit attemps (indicated by a star), but the Ad/Promo measures are not (indicated by a 0).

(f) Thus, Ad/Promo was not supported as a mediator between countries and rate of quit attempts. That is, changes in Ad/Promo do not help explain why countries varied in quit attempts. In contrast, the significant paths from Countries to Labels and from Labels to Rate of quit attempts supports the contentions that the change in warning labels mediated the pathway from Countries to Rate of quit attemps and that the change in warning labels was responsible for the increase in the rate of quit attemps.

Figure 2.7 The use of mediational models for isolating the effects of specific policies

studies, there would still be the need to measure the impact of other possible influences on behaviour that had occurred between the policy intervention and the post-policy assessment point.

A more complete articulation of the strategy of teasing apart the impact of multiple policies, and/or the presence of other possible influences/confounding factors can be found in the approach to mediational analyses (e.g. Baron & Kenny 1986; MacKinnon *et al.*, 2002; Mathieu & Taylor, 2006; and Spencer *et al.*, 2005). An extended example of the logic of the approach is provided in Figures 2.7 a-f. The scenario is that ITC countries varied in the rate of quit attempts. For simpliciity, four policies are listed: taxation, labels, ad/promo, and smoke free, and the analysis involved the policy-specific variables associated with each of the four policies.

Moderator variables in the ITC Project:

One of the most intriguing lines of inquiry in the ITC Project is to determine whether the impact of the same or similar FCTC policy differs across different countries. In the domain of health warnings (Article 11), the ITC Project is addressing whether the impact of graphic warnings differs across different countries. Among the ITC countries to date, Thailand and Australia have introduced graphic warnings since the beginning of the ITC surveys, and several other countries are anticipated to do so in the future.

The ITC Project is also examining the impact of smoke-free laws in several ITC countries. To date, the impact has been remarkably similar in Ireland (Fong *et al.*, 2006b) and Scotland (Hyland *et al.*, 2007). Ongoing ITC surveys will allow a rigorous comparative evaluation of the impact of smoke-free laws in other ITC countries including France, Germany, The Netherlands and China. Given that the ITC Surveys are using identical or very similar measures and parallel data collection methods across the set of ITC countries, the potential for making conclusions about the commonality or differences of the impact of smoke-free laws, graphic warnings, and the other FCTC policy domains will be strong.

Thus “country” and the environmental and cultural factors that “country” embodies, constitutes an important moderator variable in the ITC conceptual model.

Further, within a country, it is possible to test for differential policy impact on subgroups of a population, by including variables to determine which subgroups are more favourably (and less favourably) influenced by FCTC policies. These moderators fall into five broad classes: *socio-demographics* (age, sex, SES, ethnic background); past behaviour (smoking history, current consumption (cigarettes per day), quit attempts); *personality characteristics* (time perspective, depression, sensation seeking); other environmental effects (stress levels); and *potential exposure to policy* (unemployed people should be less affected by workplace smoking policies).

Dealing with hypothesised moderators is relatively straightforward when they are postulated merely to add predictive power to linear models. The issues become more complex when different mediational pathways are postulated for subpopulations. For example, individuals who avoid warnings might change behaviour through more emotion-related pathways, while those who take in the information on warning labels might be influenced through more cognitive pathways. The ITC Surveys have the design and the measures that will allow the creation of separate models for these different subpopulations, which will make it possible to test whether different subpopulations within a country, as well as between different country populations, respond in the same way or differently to tobacco control policies.

Conclusions

This section has provided some basic principles of how evaluation studies can be designed to offer more confident judgments about the causal impact of tobacco control policies. It has also illustrated the use of study designs (the structural aspects of an

evaluation study) and study features (the selection of measures to be used in an evaluation study, including theoretically guided mediators and moderators).

The eventual outcome of rigorous evaluation studies does not end with a causal statement, however. If mediational analyses demonstrate that a given policy works through changes in one putative mediator but not another, non-policy interventions (e.g. mass media campaigns) can be tailored to influence those mediators that had been identified in the evaluation study to be the operating causal forces leading to favorable changes in behaviour.

Thus, rigorous evaluation of FCTC policies has the potential not only to demonstrate the impact of these policies on tobacco use, but also to provide valuable insights into the development of more effective non-policy efforts to reduce the burden of tobacco use throughout the world.