

Chapter 4 Tobacco Labeling Toolkit

# EVALUATING HEALTH WARNINGS & MESSAGES



Prepared by:

David Hammond  
Department of Health Studies  
University of Waterloo

February 2009

This chapter is taken from the **Tobacco labelling and packaging toolkit**.

A complete copy of this toolkit and additional resources are available at:  
[www.tobaccolabels.org](http://www.tobaccolabels.org) , or by contacting the author directly:

David Hammond  
Department of Health Studies  
University of Waterloo  
200 University Ave West  
Waterloo, ON  
Canada N2L3G1  
Email: [dhammond@uwaterloo.ca](mailto:dhammond@uwaterloo.ca)

Financial support for this work was provided by Tobacco Control at The Union  
(International Union Against Tuberculosis and Lung Disease) [www.tobaccofreeunion.org](http://www.tobaccofreeunion.org)

## BACKGROUND

---

The focus of this section is to describe how to pre-test and evaluate the impact of warnings. Although some jurisdictions have conducted extensive evaluation work before implementing health warnings, others have selected and implemented warnings with no pre-implementation evaluation. Although a lack of resources should never act as a barrier to implementation, even modest evaluation work is likely to increase the effectiveness of warnings.

The goal of this section is to describe a range of evaluation activities that can be adapted to local needs and the availability of resources. As with the previous section, special consideration has been made for jurisdictions with minimal resources for evaluation.

### **A. Pre-implementation: Pre-testing the layout and design of warnings**

#### ***Primary Objectives***

Jurisdictions that wish to explore new design features, or jurisdictions that require evidence of the impact of larger, pictorial warnings may wish to evaluate individual components of layout and design.

#### ***Priorities***

The following layout and design features may be considered a priority for pre-testing:

- Text-only vs. picture warning
- Position of text vs. picture
- Inclusion of a government attribution
- Inclusion of a marker word
- Overall size of warning and relative size on the “front” and “back”
- Colour schemes, including contrast between background and text

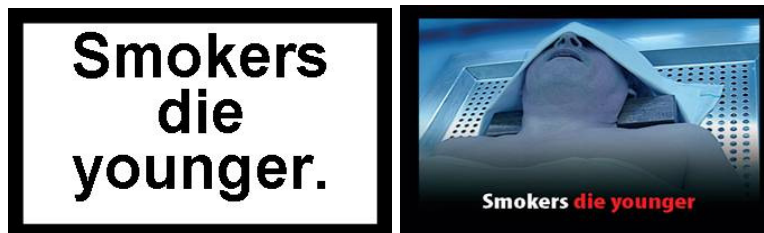
## Methods

The basic principle is to systematically evaluate each design feature so that strong conclusions can be drawn about their effectiveness. This involves creating different versions of the same warning that are identical, except for the feature that is being examined. In marketing research, this type of approach is often called “conjoint analysis.” For example, if the value of pictures vs. text-only warnings were being evaluated, two warnings should be created that are identical, except for the addition of the picture. Therefore, the picture vs. text-only should be the same size, have the same text message, same border width, etc. An example is provided below.

Set A: Picture vs. text-only



Set B: Picture vs. text-only



This approach ensures that if the warnings are rated differently by participants, the differences in scores can be attributed to the use of pictures—the only point of difference between the two warnings. Note that the picture that is selected to go along with the text will have some influence

on whether the picture warnings are rated as more effective. Therefore, it is important to repeat the process more than once, to ensure that the success or failure of the picture warnings is not simply due to a particular image. The best option is to use pairs of text-only and picture warnings across different themes. If the same pattern of results is found for both Sets A and B above, for example, the findings will be more robust.

#### *Presentation of warnings and participant ratings*

There are two approaches to presenting the warnings to participants. The first approach is to show both warnings in each “set” at the same time, and ask participants to directly compare the warnings. In the example above, participants would be shown both warnings from Set A and B, and asked which of the two warnings had greater immediate impact. The second approach is to show participants each warning one at a time and have participants rate the warnings separately. In other words, each warning would receive a score for immediate impact using a standard rating scale (see below), and these scores could then be compared to examine which warning was rated more highly. The advantage of the second approach—having participants rate each warning individually—is that warnings can then be compared across “sets” or themes fairly easily, without using statistical techniques. In other words, the impact ratings for each of the warnings in Set A could be compared with each of the warnings from Set B.

#### *Developing the questions and rating scales*

The design and layout of health warnings can be evaluated on a range of different “outcomes.” Potential outcomes include the overall effectiveness of a warning, immediate “impact”, noticeability, and the credibility of the information. The choice of outcomes should be guided by what is being evaluated. For example, if you are testing whether a government attribution should be included, you may be most interested in outcomes regarding the

credibility of the warning.

Participants are often required to use a rating scale when responding to questions. Participants may be asked to rate each warning by selecting a number or symbol that corresponds to a particular category. The category is often written directly below the number or symbol, and typically ranges from “Very bad” to “Very good”, or some version of these words. The use of a number or symbol along with the category helps to ensure that the rating scale is easily understood by low literacy smokers.

Examples of rating scales:

1	2	3	4	5
★	★★	★★★	★★★★	★★★★★
☹☹	☹	☹	☺	☺☺
<b>Very bad</b>	<b>Bad</b>	<b>In the middle</b>	<b>Good</b>	<b>Very good</b>

Different questions should use the same rating scale for consistency. In other words, questions about immediate impact, noticeability, and credibility can all use the same 5-point rating scale. At the end of the process, each warning will have a set of ratings that can be compared across questions.

Note that in addition to questions specifically related to the health warnings, basic demographic variables should also be collected from participants, including smoking status, age, gender, and education level. Demographic variables can help to indicate whether different types of participants are providing different patterns of scores or ratings.

© **Should I use a focus group or survey when pre-testing warnings?**

*Focus groups*— A focus group is a form of qualitative research in which a group of people are asked about their attitude towards a product or concept. Questions are asked in an interactive group setting where participants are free to talk with other group members. Focus groups are an effective method for generating new ideas and concepts, particularly during the early stages of development. One limitation of focus groups is that the findings can be somewhat difficult to summarize given the unstructured nature of the group setting. In addition, the responses of each individual can be influenced by the group setting.

*Brief Survey*— In contrast to focus groups, surveys collect responses from each respondent individually, using more structured word and response options. The main advantage of conducting a survey is that responses can be collected more systematically for each individual, without social influences from other members in a group setting. One of the disadvantages to using surveys is that they are less effective than focus groups at exploring new ideas and concepts, although open-ended questions are capable of this to some extent. Surveys that are used to evaluate warnings will need to be conducted in-person or “face-to-face”, rather than by telephone so that respondents can view images. “Self-completed” mail surveys and internet surveys are possible, although are less favourable in most cases.

*A combined approach*— The most effective and efficient approach may be a combination of surveys and focus groups. For example, participants can be recruited to a group setting, which may begin with a brief background survey on smoking status and demographics. The group can then be presented with the series of warning labels to be evaluated and instructed to complete written survey questions after each presentation. This

should be done individually using structured questions, without group discussion or sharing of information. After all the warnings have been presented and the surveys have been completed, the warnings can then be presented a second time with group discussion following each presentation. It is important to wait until all of the warnings have been presented and all survey questions have been completed before beginning any group discussion; otherwise opinions from different group members may affect how each individual responds to subsequent survey items. This “combined” approach yields structured responses at the individual level, as well as additional context from the group discussions that follow.

There are many other types of studies and techniques available to evaluate the layout and design of health warnings. Methods used to date include eye-tracking, fMRI, and other physiological responses which are all used to examine general levels of attention and the strength of first impressions. Each of these methods can be informative, but they are largely used for basic research purposes and are not necessary as part of a standard approach.

#### *Target audience*

A primary goal is to ensure that health warnings are easily understood among all smokers. To this end, it is absolutely critical that evaluation work includes participants with low levels of literacy and diverse socio-economic backgrounds. This is especially important given that, in most countries, smokers have lower levels of education than the general public. In order to ensure a suitable mix of participants, participants should be recruited from public areas with a cross-section of people, such as shopping areas and other public meeting places. In some cases, it may be necessary to specifically target and recruit participants from lower SES areas or occupations. Although many individuals are willing to participate in surveys, providing a small compensation in the form of a small gift or small amount of

money can help to increase participation rates.

📌 **RESOURCE: How to conduct focus groups**

The International Development Research Centre has assembled an overview of how to conduct focus groups, as well as general guides on developing surveys, recruiting participants, and basics of data analysis. The book is available free of charge on the internet: [http://www.idrc.ca/en/ev-56615-201-1-DO\\_TOPIC.html](http://www.idrc.ca/en/ev-56615-201-1-DO_TOPIC.html)

There are a number of government reports that describe findings from previous focus groups conducted to test health warnings: [www.tobaccolabels.org](http://www.tobaccolabels.org)

## **B. Pre-implementation: Concept and content testing**

### ***Primary Objectives***

The main objective of concept and content testing are to evaluate the most effective health warning concepts for each theme and subject. Jurisdictions with both the time and resources often conduct this type of evaluation in several stages; initially to generate feedback on early concepts, as well as to test “final” versions before implementation.

### ***Priorities***

The main priorities are to ensure that each warning under consideration meets the following criteria:

- Strong initial impact.
- Consistency between text and picture.
- All text is clear and easily understood.
- Engaging and interesting text.
- Personal relevance and emotional impact.

- Credibility of message.
- Overall perceptions of effectiveness.

### **Methods**

The basic principles are the same as evaluating layout and design: the process should be as systematic as possible, while also allowing for the possibility of broad feedback. Early testing of concepts and content is usually somewhat less structured. Often, very different concepts will be presented to participants to collect general feedback on which direction to pursue. However, as the content in the warnings becomes more defined, testing should become more systematic: the best way to test a specific concept is to develop similar versions of the same warning that differ only on one aspect of the content.

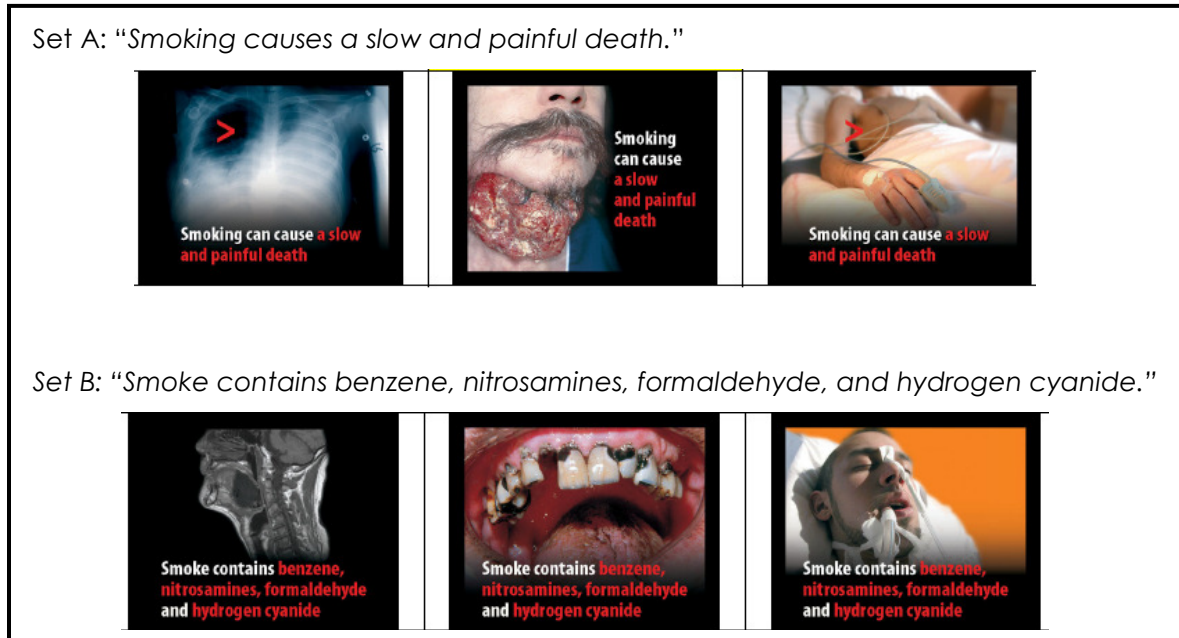
#### *Developing the questions and rating scales*

The main difference between layout/design and content/concept testing is the types of questions that will be asked. The questions should focus to a greater extent on how the information and content is received. Questions should be developed to measure immediate noticeability and impact, consistency between text and picture, clarity and meaning of text, interest in text, personal relevance, emotional impact, credibility of the information, and overall perception of effectiveness.

It is often helpful to ask about specific components of the warnings, such as the picture, text message, the cohesiveness of the pictures vs. text, etc. These types of questions provide important feedback about how to improve specific aspects of the warning. A typical approach would be to begin with a general question on the overall effectiveness of a warning before asking separate questions about each main component.

## Presentation and Ratings

Several concepts should be developed for each theme or “message”. See examples below, where three concepts were tested for each message.



As with testing the design and layout, there are two approaches to presenting the warnings to participants. The first approach is to show each set of three warnings at the same time, and ask participants to indicate their preference. In the example above, participants would be shown all three warnings from Set A simultaneously and asked which of the three warnings had greater immediate impact, for example. The second approach is to show participants each warning one at a time and have participants rate the warnings separately. The advantage of the second approach is that it allows warnings from different sets to be compared without using statistical analyses. This is especially important to identify whether certain themes or subjects are performing poorly compared to others. In other words, it tells you not only which concept is the best execution of a particular theme, but which themes are having the greatest success.

Note that concept testing and evaluation of layout and design do not necessarily have to be completed in separate surveys or focus groups. In many cases, testing of layout and design is conducted prior to specific concept and content testing only because basic decisions about text and pictures need to be determined before developing concepts. However, jurisdictions that wish to examine only a few novel layout or design features can incorporate elements of layout/design evaluation and concept/content evaluation in the same survey or focus group. Regardless, the process should be as systematic as possible with respect to the types of information that are varied and the types of questions that are asked.

### **Summary**

Pre-testing of health warnings should be as rigorous as possible given available resources, but should not create any significant delays in implementation. It is possible to complete the entire process of development and pre-testing in several months if necessary, although you should allow at least 6-months for the process if possible. Longer periods will be helpful if time allows.

#### **📍 CASE STUDY: Using the internet to engage and evaluate**

In 2006, the Department of Health in the United Kingdom chose to develop a website as a way to engage the public on the issue of pictorial warnings and to solicit their feedback on different alternatives. Visitors to the website were asked to complete a number of demographic questions and to select the warnings they felt would be effective. Over 20,000 people completed the survey during the 3-months the website was in operation. The results were used to inform the final selection of images and received considerable media attention in the process.

## **C. Implementation evaluation: Monitoring & Compliance**

### ***Primary Objectives***

The primary objective of monitoring compliance is to examine whether health warnings have been implemented on packages as planned. This type of evaluation, often called “process” evaluation, is critical to measuring compliance to the regulations.

### ***Priorities***

The main priorities of this type of evaluation are to ensure that the warnings are appearing on packages as they were intended, as well as to ensure that the warnings begin appearing by the implementation deadline.

### ***Methods***

The most straightforward approach is to visit retail outlets to visually inspect packages. This type of approach is commonly referred to as an “environmental scan.” Although there are formal protocols for conducting an environmental scan, even information approaches may be sufficient. The number of retail outlets visited will depend greatly on the availability of resources. While the number of retail outlets need not be exhaustive, a range of retail outlets in different parts of the country should be visited. In many cases, this requires relatively little expertise, with the potential to involve advocates and other public health officials if necessary. In addition, some regulators have visited factories of domestic tobacco manufacturers to ensure that packages are being printed in accordance with the regulations. Another approach is to encourage members of the public to report non-compliance, although this requires resources to publicize the phone number or reporting mechanism.

Overall, implementation evaluation for health warnings is considerably less

resource-intensive than for other policies, such as smoke-free legislation. Efforts should focus on the immediate post-implementation period, after which relatively little monitoring is typically required.

## **D. Post Implementation: Impact Evaluation**

### ***Primary Objectives***

The primary objective of impact evaluation is to examine the potential effectiveness of health warnings after implementation. In general, impact evaluations are not used to evaluate the effectiveness of individual warnings, but rather the impact of the health warnings as a whole.

### ***Priorities***

One of the main priorities is to measure potential “wear-out” of the warnings and the point at which new warnings may be required. This requires measuring whether the health warnings have met and continue to meet their objectives. Although the objectives of health warning systems may differ to some extent across jurisdictions, common objectives include the following:

- Increases in health knowledge and perception of risk.
- Greater awareness of cessation services.
- Increases in motivation to quit and cessation.

### ***Methods***

Population-based surveys provide the most comprehensive method for evaluating the impact of health warnings. Ideally, surveys should be conducted before and after the implementation of new warnings. These surveys should also use similar questions and methodology so that changes in key outcomes can be examined. Whereas some jurisdictions have conducted entire surveys devoted to evaluating the impact of health warnings, it is also possible to insert a smaller number of questions into on-

going surveys that include other topics. Basic principles for survey design and analysis are provided in the IDRC “Focus Group” resource, presented earlier in this section, as well as the resource described below.

**📍 RESOURCE: Designing impact evaluation surveys**

A detailed discussion of questions used to evaluate the impact of health warnings is included in a Monograph from the International Agency for Research on Cancer. The Monograph Chapter can be requested from: [dhammond@uwaterloo.ca](mailto:dhammond@uwaterloo.ca)

*Questions*

The first step in developing questions to evaluate warnings is to identify potential outcomes of interest. Common outcomes include the following:

- Are the health warnings being noticed and how do they compare with other forms of health information?
- To what extent do smokers “process” the warnings in terms of thinking about and discussing warnings?
- Do smokers believe the information in the warnings is credible?
- Have the warnings increased levels of health knowledge and perceived risk?
- Are smokers more likely to quit due to the health warnings?
- Do health warnings reduce the appeal of the package?
- What is the level of public support for health warnings?

The resource listed above includes examples and a discussion of these and other survey questions.

© **“Can I use prevalence figures to evaluate the impact of warnings?”**

Prevalence rates from large national surveys provide the estimate of population-wide changes in smoking behaviour. However, there are several limitations to using prevalence data as a measure of whether health warnings have been effective in promoting cessation. The Canadian experience provides a good illustration of these limitations. In the six years since 2001, when large pictorial warnings were implemented in Canada, the prevalence of smoking has decreased by approximately 4%. This represents a substantial decrease of approximately one million smokers in six years—a considerable public health achievement. However, it would be inaccurate to suggest that the health warnings were responsible for all or even most of the 4% decrease in smoking. Indeed, over this six year period the price of cigarettes have increased, several mass media campaigns have been conducted, and smoke-free legislation has become considerably stronger in Canada. In other words, prevalence data are not specific to health warnings or any other single intervention. Therefore, while health warnings may have played an important role in reducing smoking in Canada, there is no way to precisely estimate the contribution.

**Other considerations**

*Timing of surveys*

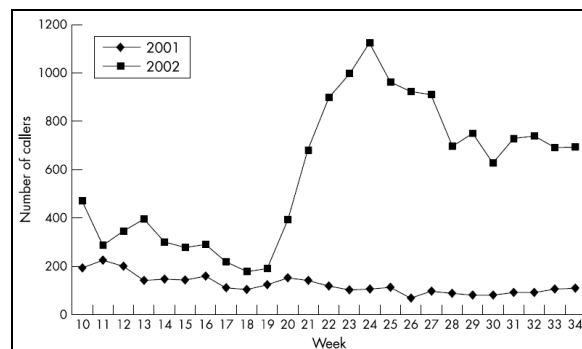
Several months often pass between the implementation date of new health warnings and the time at which they begin appearing on most packages. In addition, the cumulative impact of health warnings may build over time, with repeated exposures to the different messages. As a result, surveys that seek to measure the impact of warnings should wait at least 6 months after the implementation date. Ideally, regular surveys would be conducted to examine potential wear-out of the warnings in the long term, perhaps at 12 or 24-month intervals, if necessary.

### Target groups

Unlike some other aspects of evaluation, impact evaluations should include both smokers and non-smokers. The extent to which non-smokers notice and recall health warnings is a very good indication of their overall effectiveness in the general population.

#### 📍 CASE STUDY: Using different sources of data to evaluate health warnings

Concerns about health risks of smoking are among the most common and important reasons for quitting smoking; however, there are a number of other factors that also contribute to the decision to quit and whether or not a quit attempt is successful. Although it may be impossible to measure the precise number of smokers who quit as a direct result of health warnings, some jurisdictions have used alternative data sources to estimate the potential impact, such as tracking the use of cessation services. For example, the UK, the Netherlands, Brazil, and Australia have tracked calls to the free telephone “quitline” number that is displayed on packages in each country. In each case, calls to the national quitline have increased significantly immediately after the telephone number appears on packages. For example, the graph below shows the increase in calls to the Netherlands quitline service after the number was printed on the back of one of 14 package warnings, beginning in Week 19 of 2002. This type of data source indicates that, at the very least, the health warnings are helping to increase the use of effective cessation services.



Source: Willemsen M., Simons C, Zeeman g. Tobacco Control 2002;11: 381-2.