3

METHODS AND PROCEDURES IN WARNING RESEARCH

Tonya L. Smith-Jackson
Virginia Polytechnic Institute and State University

Michael S. Wogalter
North Carolina State University

ABSTRACT

This chapter reviews some of the basic research methodologies used to investigate warning effectiveness. Selected studies are used as examples to describe how specific methods are implemented. Although some results are discussed, the goal is to focus on the methods used to achieve the results. This chapter is organized based on the information processing stages of the Communication-Human Information Processing (C-HIP) framework (Wogalter, chap. 5, this volume). The stages and typical assessment methods include (a) attention switch and maintenance stages, using eye tracking, response time, and looking behavior; (b) comprehension/memory, using recall and recognition tests; (c) attitudes, beliefs, and motivations, using subjective and self-report measures. Issues including operational definitions, subjective measures, and validity/reliability in warning effectiveness studies are discussed.

INTRODUCTION

The last 2 to 3 decades have witnessed substantial growth in warning research. During this time research methodologies have been developed, borrowed from other fields, and refined across studies. Measurement of compliance behavior is one of the earliest methods to be developed (e.g., see Laner & Sell, 1960; Wogalter et al., 1987), but despite its usefulness, the method is relatively difficult to carry out (see Wogalter & Dingus, 1999; Kalsher & Williams, chap. 23, this volume) because of costs, effort, time, and ethical considerations. Researchers over the years have developed other methods to assess effectiveness.

This chapter describes some of the major research methodologies or techniques used to measure warning effectiveness at the information processing stages before behavioral compliance. It is intended for new researchers or consumers of research.

The stages involved in warning processing can offer insight on the conditions leading up to compliance. Potentially, the stages-of-processing perspective can provide reasons why a warning failed to produce appropriate safety behavior. The communication-human information processing (C-HIP) model (Wogalter, DeJoy, & Laughery, 1999; Wogalter, chap. 5, this volume) provides a framework to systematically examine precompliance stages. This model is used to structure this chapter's presentation of research methods. The primary reason for this approach is that the methods can target the cognitive processes involved. More information on the research methods that assess the precompliance stages of processing can be found in Young and Lovvoll (1999). For more information on research methods involving the measurement behavioral intentions and compliance behavior, see Wogalter and Dingus (1999) and Kalsher and Williams (chap. 23, this volume).

Handbook of Warnings, Edited by Michael S. Wogalter, © 2006, Lawrence Erlbaum Associates (Mahwah, NJ).

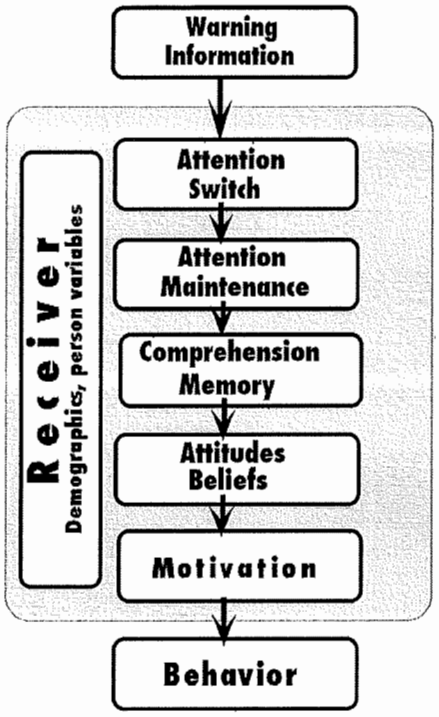

FIGURE 3.1. The information processing stages of the Receiver portion the Communication-Human Information Processing (C-HIP) Model (see also Wogalter, chap. 5, this volume).

C-HIP MODEL

Figure 3.1 is a representation of the C-HIP model (Wogalter, DeJoy, & Laughery, 1999; Wogalter, chap. 5, this volume) that shows a set of stages involved in warning processing from the source to compliance behavior. In this chapter, the focus is on the stages of the model within the receiver of the warning. This includes the processing stages of: attention switch and maintenance, memory and comprehension, attitudes and beliefs, and motivation. As mentioned earlier, these stages are precursor stages occurring before behavioral compliance, the last of the receiver's substages in C-HIP.

According to the C-HIP model, a warning may fail to affect behavior because of bottlenecks in the precursor sequence of stages. If successfully processed at a given stage, the information flows through to the next stage. If processing at a stage is unsuccessful, the bottleneck blocks or prevents the flow of information from getting to the next stage. However, assuming the warning is noticed and attended to, the blockage could be at the memory and comprehension stage. The individual may not understand the warning, and, as a consequence, no additional warning processing occurs beyond that point. Even if the message is understood, it still might not be believed; and even if believed, it may not motivate behavior. If all of the stages are successful, the warning process ends in behavioral compliance attributable to the warning information.

Although warning processing may not make it through all of the stages to the behavioral compliance stage, it can still affect and be effective at earlier stages. For example, a warning could be effective in terms of indicators showing increased understanding and appropriately affected beliefs, and these are valid indicators of effectiveness. But the warning may not affect behavior, perhaps because of blockage at the motivation stage. Although changed behavior may not result, the warning could be effective in influencing other stages, and at some later time affect behavior appropriately.

Besides its use in describing the information processing stages, another benefit of the C-HIP model is in investigations determining where the blockage occurs (for the purpose of removing the bottleneck and allowing processing to continue to subsequent stages). This use of the C-HIP model with respect to warning methodology is described in a later section.

The sections that follow provide an overview of some of the research methods used to assess processes at each of the receiver's substages. The research procedures described generally typify a class of research methodologies or paradigms that may be used to investigate processing at each of the stages. An additional purpose of the presentation is to highlight some of the critical aspects of the research so that future researchers and consumers of research can gain a better understanding of the methods that can be used to assess various aspects of warning effectiveness. The focus will be on research methods involving warnings presented in the visual modality, although many of the same methods are applicable to auditory warnings.

Unfortunately, internal mental processing cannot be viewed or measured directly. All mental processes are assessed indirectly by measuring something that is believed to reflect the mental process (Young & Lovvoll, 1999). Although measurements are indirect, there are techniques that can make overt what cannot be seen directly. The measurement techniques are operational definitions of constructs. In other words, mental processes, which cannot be seen (the construct), are measured by an indicator that can be seen (the operational definition). For example, thirst and hunger are hypothetical constructs that cannot actually be seen as internal mental processes are involved; but indicators of them can be viewed, in this case how much drink or food is consumed (which operationally define the concepts). The assumption is that greater consumption is at least somewhat related to the internal process involved with thirst and hunger. Operational definitions and hypothetical constructs are discussed further in a later section.

ATTENTION SWITCH AND MAINTENANCE

The first two stages of the receiver portion of the C-HIP model both concern attention. The initial stage is when attention is switched from other stimuli, thoughts, or tasks to the warning. To make that switch reliably, the warning must have characteristics that make it noticeable, conspicuous, prominent, or salient relative to its context or background. Attention maintenance refers to the process of holding attention onto the warning so that users can encode its message. A demonstration of poor attention maintenance is a person who notices a warning but switches his or her attention away to something else.

There are different methodologies to assess the two stages of attention. They are described in the sections that follow.

Attention Switch

Warnings that switch attention to them are more likely to be processed further. Warnings compete for attention with other stimuli in the environment, as well as ongoing thoughts and tasks.

Research related to attention switch has examined the conditions under which warnings are made more or less conspicuous. Methodologies to investigate attention switch include response time, eye movement, looking behavior, and subjective/self-report measures. The measures provided by these paradigms are indicators of attention, or in other words, the constructs of attention switch and maintenance are operationally defined as performance within these paradigms.

Response Time. Response times measure the speed of some process(es). Here it is assumed that faster times reflect ease of processing. With respect to noticeability, the assumption is that if participants are able to find a stimulus in one condition faster than in another condition, it indicates that the first condition is relatively more noticeable than the second. Search speed is an indirect measure, or in other words, the operational definition of stimulus salience/noticeability construct. Thus, an underlying assumption is that more salient warnings will capture attention faster compared to less salient warnings and that this will be reflected in the amount of time it takes to respond to the warning stimuli.

One example of a study involving response times was published by Young (1991). He tested a total of 48 alcohol-warning designs that included the systematic variation of several warning features: presence versus absence of an alert symbol (triangle exclamation point), color (red vs. black print), and a thin-lined border. These designs were presented to participants in the context of labels displayed on a computer screen for alcoholic beverages. Some labels had a warning with one or more of the above features and some labels had no warning. Participants were seated in front of a computer monitor. As participants sat in this position, labels were presented in a random order on the computer display. Before each label appeared, the participant was directed to look at a small fixation point, usually a "+" in the center of the display to stabilize the eye position before each trial. On being shown each label, participants were to determine as quickly and as accurately as possible whether a warning was present and, if so, to press one or two buttons. If no warning was present then they were press the other button. The warning-absent trials are "catch" trials used to keep the participants' response strategies "honest" so they would not press the "yes" button immediately after the onset of each trial. Usually the "catch" trials are not analyzed.

Young (1991) found significantly faster response times when the alert symbol and color were present in the warnings compared to their absence. The presence of a border produced a trend of faster response times than when no border was present, but the difference was not statistically significant.

In another study that used response time measures, Bzostek and Wogalter (1999) had participants press one of two computer keys when they found certain specific warnings in an over-the-counter cold medication label that was displayed on a computer screen. The warnings' locations in the label across trials were manipulated (e.g., vertical placement and left–right column), as were other components (e.g., presence or absence of a symbol and color). The results showed that the presence of warnings placed toward the top and left sides of the label and those having color and possessing a symbol produced the fastest response times. The slowest times occurred when these aspects were largely absent.

Eye Tracking. Measures of eye movement are used to assess the direction and dwell time of eye gazes when exposed to a warning. Eye movement recording provides an operational definition of attentional processing. The assumption is that attention at the cognitive level is manifested by the direction and dwell time of eye movements.

Laughery and Young (1991) used a camera with infrared lens, a pupil tracking system, and an auto calibration system to record eye movements, fixations, and pupil positions. Like Young (1991), the purpose of the study was to determine differences in the noticeability of certain warning features displayed in alcoholic beverage labels. In this study, the warnings were manipulated according to the presence or absence of an alert symbol (triangle exclamation point), color (red vs. black print), and a thin-lined border. Participants were seated in front of a computer monitor and their pupil positions were calibrated while the participant sat still and moved their eyes to designated corners of the screen. It is important for participants to sit still, but because people naturally move small amounts, it is common, at least with older equipment, to keep participants' head still by having them clamp their teeth on a bite bar (covered with a clean plastic bag). Still participants move a little anyway, and it is not uncommon for some data in eye movement studies to be lost. Pupil positions often need to be recalibrated. At the start of each trial presentation, the participant looked at a fixation point followed by the presentation of one of the labels. The participant's task was to make the determination whether a warning was present. In the trials with the warning present, fixation sequences were recorded on a computer (how the eyes moved from a center position fixation point to the warning) and time duration of eye fixations (dwell time) at designated zonal areas within the label. The data were later organized and statistically analyzed according to whether there were consistent differences in the eye-movement recordings as a function of manipulated warning features in the labels (presence or absence of symbol, color, and border).

Unfortunately, there are only a few other studies using eye movement measurement in the warnings literature. Horberry, Purdy, and Gale (1997) evaluated the noticeability of bridge warning signs by tracking eye movements and eye fixations while participants were in a simulated driving situation. In another study, Krugman, Fox, Fletcher, Fischer, and Rojas (1994) used eye tracking equipment to determine whether teenagers attended to cigarette warnings that were placed within cigarette advertisements.

Eye movements are an excellent way to determine relative conspicuity. Unfortunately this method has been plagued by problems, which have inhibited its use. These include the use of costly, difficult-to-use equipment and considerable patience on

the part of both the researcher and participants. Data is sometimes lost because of participant movement (which can produce statistical analysis problems), and the necessity of frequent pupil/line-of-sight recalibrations. There are sometimes issues with equipment malfunction and operator expertise. New technologies that reduce these problems will promote its greater use in warning research.

Looking Behavior. Several studies have used a more global form of head movements to evaluate warning effectiveness in terms of the noticeability of various features (i.e., color, borders, or font sizes). *Looking behavior* can be assessed through empirical observation of the positioning of the head as assessed by an observer. For example, Wogalter and Rashid (1998) tested the effectiveness of different border designs (ANSI-based and other designs) placed around the periphery of warning signs. Specifically they examined whether adding a rectangular border around the warning text to increase the sign's salience would promote greater attention as measured by two indicators of looking behavior. Six conditions were tested. In four, warning text was surrounded by one of four different borders (thick red, thick yellow/black alternating stripes, thin red, or thin black). The other two conditions were controls: warning text with no border and no warning (blank sign). The signs were individually posted in a university campus building and more than 1,200 people were unobtrusively observed on whether or not they looked at the sign and the amount of time they spent examining it. The results showed that signs with thick red and thick yellow/black diagonal stripes were looked at more frequently and examined for longer periods of time than the other signs.

Smith-Jackson (2004) also measured looking behavior in which participants were placed in a simulated manufacturing task requiring the use of a punch press and spindle machine. An observer watched participants as they performed the task using the instructions provided to them. Using a checklist, the observer recorded critical safety behaviors associated with the task, including whether the participant looked at a warning that was prominently displayed on the punch press. Another example of looking behavior research is Louch, Price, Esson, and Feistner (1999). They used unobtrusive observation to measure the amount of time visitors spent reading signs that were manipulated on the basis of pictures, color, and "grabber" headlines. The observers recorded whether a sign was looked at and the length of time spent looking at the sign.

An important methodological point in looking behavior research is reliability of the measure. It is advisable to have at least two researchers make independent observations of the same phenomenon so that the extent to which they agree can be determined. Multiple observers are advised whenever the behaviors measured may be quick, multidimensional, and subject to interpretation.

Self-Reports. Another way to assess attention switch is to ask participants after having been exposed to a warning if they noticed it. Self-report measures are commonly collected from participants in postexperiment questionnaires in compliance-type studies (see Kalsher & Williams, chap. 23, this volume).

For example, in a compliance study by Frantz, Rhoades, Young, and Schiller (2000), participants completed a task requiring them to unpack and assemble a file cabinet. There were four different warning conditions, and researchers observed whether the participants complied with the warning. After the task was completed, participants were asked if they noticed the warning label, and a *Yes* or *No* response was recorded for each participant. Similar procedures for gathering data on warning noticeability or attention switch have been conducted by other researchers (e.g., Wogalter, Allison, & McKenna, 1989).

Rousseau and Wogalter (chap. 11, this volume) point out that these measures are to some extent plagued by other intervening variables. One is memory; the self-reports are usually assessed some time after participants may have viewed the warning. Another is the potential for bias, sometimes called the good participant effect or social desirability, in which some participants may give an answer based, in part, on what they think is expected of them by the researchers. As a consequence, they may indicate having seen the warning when they actually did not (or vice versa). Although self-reports may be influenced to some extent by other variables, the method can be useful in investigating why a warning failed to produce compliance (a topic covered later).

Other measures can bolster self-reports. For example, sometimes video surveillance can be used in conjunction with self-reports. With a video, researchers would have an opportunity to compare judges' observations of participants' looking behavior with participants' self-reports.

Subjective measures such as self-reports and ratings are commonly contrasted with objective (overt) performance measures such as response time, eye movements, and looking behavior. With performance measures, behavior of some form is recorded and quantified in some way. Subjective measures are sometimes considered less direct and desirable than performance measures in assessing mental phenomena. Usually subjective measurements are the easiest and least costly of all of the warning research methodologies to use. But objective measures are usually preferred over subjective measures when available and feasible. With respect to attention switch, the previously mentioned objective performance measures of response times, eye movements, and looking behavior would be preferred as indicators over subjective measures such as self-reports. However, as we will see, subjective self-reporting is the only method available to researchers for evaluating some of the later stages of the C-HIP model (e.g., beliefs and attitudes).

Attention Maintenance

Attention maintenance follows attention switch during which the receiver's gaze is held to the warning stimulus. Attention should be held for a long enough time for the person to acquire the information from the warning.

Eye Tracking and Looking Behavior. Some of the same techniques used to assess attention switch can also be used to measure attention maintenance. Objective measures include eye tracking and looking behavior research to measure how long people fixate or dwell on the warning material. It is important

to note that dwell time in eye movement studies is not by itself an indication of adequate attention maintenance. A person may be attracted to a well-formatted, aesthetic warning and look at it for some amount of time. During this time, the person may learn relevant aspects of the hazard and know how to avoid it by taking the time to read it. However, with a poorly formatted, less-legible warning, the person could also take a lot of time to examine the warning but encode very little. Consider also the reverse, that both a good and a poor warning might yield very short dwell times. With the poor warning, the person might not look very long, moving their fixations to something else, and as a consequence, encoding little or nothing from the warning. But a short dwell time might also be yielded with a good warning. The good warning may enabling quick extraction of information or simply be serving as a reminder by quickly activating information the individual already knows. The point is that dwell time cannot be directly related to attention maintenance without other measures being collected such as memory and comprehension (discussed later). Dwell time is dependent on multiple factors, including existing knowledge and memory of the receiver.

Legibility. One important factor that can influence attention maintenance is legibility. Legibility concerns whether the letters of text and important markings of symbols are discriminable/distinguishable. It is relevant at the attention maintenance stage because a warning that has components that are difficult to discern or identify can negatively affect information acquisition.

Numerous legibility-type studies have been reported in the warning literature (e.g., Collins, Dahir, & Madrzykowski, 1992; Dewar, 1976; Wogalter, Murray, Glover, & Shaver, 2002). In the typical study, stimuli are shown degraded in some way, for example, very small, very short exposure, covered by smoke, visual noise, or obscured in another way, for example, a prohibition symbol (circle slash). The participant's task is to identify the stimulus presented.

In one study, Wogalter et al. (2002) had participants identify 16 pictorial safety symbols each presented for a very short time of .05 seconds (50 ms). The symbols were manipulated with respect to four different types of prohibition symbol (circle-slash) variants in which the slash was over, under, partial, and translucent with respect to the enclosing symbol. Participants first completed a questionnaire requesting general demographic information (e.g., gender, age). Initially, participants were presented with two practice trials of images that were not used in the main experiment to familiarize them with the quick presentation duration. Images were each presented for a very short duration of .05 sec (50 ms) followed by a checkerboard pattern to mask or reduce postexposure afterimages. After being exposed to each symbol, participants recorded what he or she believed to be its meaning on a numbered response sheet. Two independent judges scored participants' responses. The results showed that responses were more accurate for symbols having concrete, less complex, and familiar concepts, and in the conditions having under and translucent prohibition slashes.

In general, the utility of determining legibility investigations is to help select warning stimuli that remain discernable under degraded conditions. For more information on the effects of legibility in real-world conditions, see Glasscock and Dorris (chap. 39, this volume). For more information on legibility's role in attention maintenance, see Wogalter and Vigilante (chap. 18, this volume).

Self-Report and Subjective Measures. Interviews and questionnaires can assess the degree to which participants believed they looked at and acquired information from the warning. For example, Frantz et al. (2000) measured attention maintenance to a warning by asking participants to rate the extent to which they read each statement in the warning. They used a 7-point rating scale anchored with *None* (1) and *All of it* (7). Like with attention capture, subjective measures are not the preferred approach when seeking data about attention maintenance because of the issues mentioned earlier. Instead when possible, objective performance measures are preferred.

MEMORY AND COMPREHENSION

Even if a warning is attended to it may fail if the individual is not able to extract meaning from the warning and activate relevant information in memory. If previous knowledge, experiences, or other information stored in long-term memory cannot decipher the message, the warning will not only be unsuccessfully processed in the memory and comprehension stage, but also will cause failures in the stages further downstream in the C-HIP model. Thus, the process of matching the incoming message information from the warning with existing information in long-term memory is an important cognitive event with respect to effective processing of warnings.

Memory and comprehension can be measured in many ways. There is a large body of literature in cognitive psychology on various techniques. In this section, we describe in some detail a method of determining comprehension and memory that was employed in two experiments by Young and Wogalter (1990). The interest was whether warnings inside product (owner's) manuals could be enhanced so that people will more likely understand and remember them. The specific focus was whether the addition of safety symbols and conspicuous warning text would benefit comprehension and memory of the warning material inside a product manual. The study was advertised as concerning the use of consumer products. All participants were first asked to perform a set of tasks using a product (a computer) with a product manual present. After completing the computer tasks, participants were told that they would be doing another task with a product but this time the product manual would not be available during the time they used the product. However, they were also told that they could take a few minutes to look at the manual before they went to another room to use the product. In one experiment, a manual for a gas-powered electric generator was used and in the other, a manual for a natural gas oven was used. Participants were not told that they were being given one of four experimentally manipulated manuals and that their exposure time to the manual was being rigorously controlled for exactly 4 minutes. All four manuals had eight warnings spread across several pages. For example, one of these warnings

stated: "Warning: Operate generator only in well ventilated areas. The exhaust from the generator contains poisonous carbon monoxide gas. Prolonged exposure to this gas can cause severe health problems and even death." For one manual, all of the print in the manual, including the warnings, used regular text font and there were no safety symbols. The warnings in the other three manuals were: (a) conspicuous print, symbols absent; (b) conspicuous print, symbols present; and (c) regular print, symbols present. The conspicuous print was bigger, bolder with a highlighted orange background, and the safety symbols when present corresponded with the warning text. In other words, the warnings were manipulated in this experiment as 2 (conspicuous print: present vs. absent) X 2 (symbols: present vs. absent) factorial design. After the 4-minute study time was over, all participants were told that they would not actually be using the product (yet were led to believe this), but rather they would be given several questionnaires. The responses to the questionnaires served as the operational definition of the memory and comprehension constructs.

The set of questionnaires were ordered in a particular way. The first was an open-ended comprehension test. The second was a symbol identification test, and the third was a symbol comprehension test. The reason for ordering the tests this way was to ensure that the first and most important test was not contaminated by exposure and answering questions on the other two tests. Each test gave progressively more cues to assess what the participants knew. Their responses were self-reports.

The open-ended comprehension test asked brief questions about the hazards associated with the machines (i.e., information covered in the warnings). Care was taken in designing the questions so that the test itself would not contaminate responses. Sometimes these kinds of tests are called free or cued recall tests depending on how much information is given.

The second questionnaire, the symbol recognition test, presented the symbols that were shown in two of the manual conditions mixed together with other similar symbols (as distracters or foils). Participants marked next to each of them whether it was shown before. The third questionnaire, the symbol comprehension test, was similar to the test described in the annex to the American National Standard Institute's (ANSI, 2002) Z535.3's Criteria for Safety Symbols. The ANSI Z535 test is described in more detail in the chapters by Deppa (chap. 37, this volume), Peckham (chap. 33, this volume) and Wogalter, Silver, Leonard, and Zaikina (chap. 12, this volume). In Young and Wogalter (1990), all participants were given a sheet with the eight safety symbols that appeared in two of the manual conditions and were asked to describe what each meant with respect to the product (generator or oven depending on the experiment).

After the experiment was completed, the data from the questionnaires were scored. The first and the third questionnaire, because they were both open-ended type questions, required somewhat elaborate scoring procedures. One of the procedures was to first break down the content of the warnings into component parts and then have judges look for those component parts in the participants' answers. The data were scored by two judges on the accuracy and completeness of the responses with respect to each of the eight warnings given in the manuals. Two judges were used to determine the reliability of the scoring. The judges did the scoring *blind* or, in other words, they were provided no information on participants' assignment to conditions (by coding the response sheets in a way that the judges would not be able to decode). Also the scoring was done in two ways. One was a *strict* method in which judges scored the answers on whether the answers exactly matched the component warning text in the manual. These data were ascribed to memory performance. The other method of scoring was more *lenient*. Lenient scoring was based on whether the participant's answer was synonymous (not needing to be exactly identical) to the warnings in the manual. These data were ascribed as comprehension performance. These two scoring criteria were used for both the first and third questionnaires. The second questionnaire was easiest to score because participants were simply asked whether the symbols were presented earlier and so the correctness or incorrectness could be based on a prepared set of answers.

After completing the data collection phase, the scores were put into a computer database with each line (row) assigned to a different participant along with their component test scores and their assigned experimental condition. The component scores were totaled in different ways and each analyzed in a 2 (presence versus absense of conspicuous print) × 2 (presence versus absense of symbols) between subjects analysis of variance.

The third test was symbol comprehension. Given the procedure of Young and Wogalter (1990), it is likely that there was some effect on the symbol comprehension test from the other two tests preceding it. If their experiment had been mainly focused on symbol comprehension, then it would have made sense to give that test initially, but that was not the main objective of the study. Other chapters in this Handbook describe symbol comprehension testing (Deppa, chap. 37, this volume; Peckham, chap. 33, this volume; Wogalter, Silver, Leonard & Zaikina, chap. 12, this volume), so it will not be covered in detail here (see also Wolff & Wogalter, 1998). However, a few notable methodological issues are worth mentioning, as they may be important for new researchers and consumers of research. Symbol comprehension tests should be presented to participants together with a description (and pictures) of where the symbol might be placed. Open-ended comprehension tests are preferred over other kinds of tests, for example, multiple choice or matching. However in open-ended tests, participants can give incomplete answers even though they may know more. Probing or giving very general prompts to continue describing their answer may be beneficial in finding out what they actually know. There should be two or more judges to score the answers. Scoring open-ended responses can be difficult because judges commonly need to make inferences according to what they believe the participant meant. The scores assigned by judges are based on their internal subjective criteria. The judges' criteria should be externalized in research reports.

Subjective Measures

Ratings are another technique to measure comprehension Participants can be asked to rate the degree to which they or others

would understand the warning or its components. This method of measuring understanding is easier to conduct than formal symbol comprehension tests, and it is useful when culling a large set of prototype symbols down to a smaller set to be formally tested in the gold-standard symbol comprehension test (Dewar, 1999). The procedure involves (Brugger, 1999; Zwaga, 1989) asking participants to estimate the percentage of people who would comprehend each of the symbols (0% to 100%). Other types of rating measures (e.g., visualizability, construct correspondence) have been found to relate well with symbol comprehension scores (Hicks, Bell, & Wogalter, 2003; Young & Wogalter, 2001).

Generally, for the assessment of memory and comprehension, subjective rating studies, although easier to conduct, are less preferred compared to objective performance assessments such as open-ended testing.

ATTITUDES AND BELIEFS

Beliefs and attitudes are mental frames-of-reference based on cognitive and affective experiences of the individual. Attitudes are similar to beliefs except they have a more emotional component. They are important because a warning may be adequately noticeable and understandable, but is otherwise ineffective because it did not adequately influence people's hazard-related attitudes and beliefs.

In the first two stages of the C-HIP model, there were both objective and subjective measures a researcher may select to use in a study. In general, tests that provided objective performance data were preferred. However, beliefs and attitudes cannot be determined very well by methods involving objective (overt) performance measures. There are few objective measures that specifically target attitudes and beliefs, except in some behavioral compliance studies (see Kalsher & Williams, chap. 23, this volume). The main, and in some respects the only, widely used measure of beliefs and attitudes (and motivation—to be discussed later) is subjective, self-report tests. Subjective measures are useful because they provide data from the participants' perspectives, which may not be apparent in analyses of less sensitive measures of warning effectiveness (e.g., behavioral change derived from performance data). Also, direct reports of beliefs and attitudes from participants would seem to be a fair, direct, and simple way to measure them.

Perceived Hazard

Although "perceived" is in the name "perceived hazard," it is not tied directly to the senses; it is really a belief/attitude. Perceived hazard describes people's beliefs about the dangers associated with something, such as a product, task, or environment. The concept goes by various names in the literature such as hazardousness, risk perception, danger, and urgency (see Leonard, Otani, & Wogalter, 1999; Wogalter, Young, Brelsford, & Barlow, 1999). In one study, Wogalter, Desaulniers, and Brelsford (1987) had participants rate a generic set of 72 products on various scales. Hazard perception was measured by a 9-point Likert-type scale ranging from 0 to 8 with the following textual anchors: 0 (*not-at-all hazardous*), 2 (*somewhat hazardous*), 4 (*hazardous*), 6 (*very hazardous*) and 8 (*extremely hazardous*). Two additional rating scales determined users' beliefs about the severity of injury and likelihood of injury if they were to use the product. They had similar rating scales, but the verbal anchors were modified for each construct. The severity ratings were on the scale of 0 (*not severe*), 2 (*slightly severe*), 4 (*severe*), 6 (*very severe*) and 8 (*extremely severe*). The likelihood ratings were on a scale of: 0 (*never*), 2 (*unlikely*), 4 (*likely*), 6 (*very likely*) and 8 (*extremely likely*). The questions were randomized for each participant, and the product names were presented in one of four random orders. The ratings were collapsed across participants forming 72 means for each question and then these data were analyzed using correlation and regression analyses. Wogalter et al. (1987) found that perceived product hazard was better predicted by severity of injury ratings than likelihood of injury ratings.

Researchers also evaluated other belief dimensions using ratings such as familiarity (e.g., Wogalter, Brems, & Martin, 1993) and source credibility (e.g., Wogalter, Kalsher, & Rashid, 1999). See also Leonard, Hill, and Karnes (1989) and Braun, Holt, and Silver (1995).

MOTIVATION

Motivation is a difficult term to define, and there is disagreement on the definition. Motivation can be thought of as a *push* or a *pull* toward a goal, something that energizes behavior. It is also useful to think of motivation as the likelihood of doing or not doing something.

Like beliefs and attitudes, subjective ratings are the measurement of choice regarding motivation. Although some studies have used objective performance to measure the effects of various motivational variables, the number of studies is few and they are covered more fully elsewhere in this Handbook (see e.g., Kalsher & Williams, chap. 23, this volume) and will not be discussed here. Instead, the focus will be on subjective measures.

Generally motivation is measured by behavioral intent responses, usually ratings. Other names for this concept in the warnings literature include willingness to comply, precautionary or behavioral intent, estimated compliance likelihood, and intended carefulness.

Wogalter, Brems, and Martin (1993) collected ratings of precautionary intent for 18 consumer products categories (e.g., bleach, ladders, skateboards). These ratings were made on a 9-point scale anchored from (1) *no precaution at all* to (9) *extreme precaution*. The products were also rated on several other scales such as injury severity, likelihood of reading a warning, likelihood of a major injury, and likelihood of a minor injury. Among the results found was that precautionary intent was best predicted by the severity of injury expected.

Another study investigating intended carefulness was conducted by Barzegar and Wogalter (1998). They examined speech-based warning signal words. Forty-three signal words such as RISKY, PREVENT, URGENT, ALARM, SEVERE,

ATTENTION, DEADLY, DANGER, WARNING, CAUTION, and NOTICE were presented by male and female speakers and pronounced in 3 ways: monotone, emotional, and whisper. They rated each presented word on: "How careful would you be after hearing each word?" The verbal anchors were similar to the form presented earlier: (0) *not at all careful*, (2) *slightly careful*, (4) *careful*, (6) *very careful*, and (8) *extremely careful*. They found similar results as found in the visual modality (see Hellier & Edworthy, chap. 30, this volume) and that females speaking in an emotional voice produced the highest ratings of intended carefulness by both male and female participants.

Empirical research has shown an association between intent to display certain behaviors and actual observed behavior, but the relationship is also driven by beliefs such as perceived control and perceived vulnerability. Researchers such as Ajzen and Fishbein (1977) and Brannon and Feist (1992) found these relationships in the context of health behaviors and other social settings (see Kalsher & Williams, chap. 23, this volume). Warning researchers have applied this relationship in warnings research, for example, see Smith-Jackson and Durak (2000).

Motivation (see Riley, chap. 21, this volume; Vredenburgh & Helmich-Rich, chap. 28, this volume) can be assessed through direct observation or by asking participants after a manipulation why they did or did not comply with a warning. The answers can be used to identify any restrictions, barriers, or other factors that block or reduce individuals' motivation to comply with a warning. Behavioral methods relevant to motivation are discussed in Wogalter and Dingus (1999) and Kalsher and Williams (chap. 23, this volume).

C-HIP AS AN INVESTIGATIVE TOOL

Suppose, for example, a warning fails to change behavior in a substantial number of participants. The reason(s) for this outcome will not be clear if the only index is behavioral compliance. A warning that fails to show an effect on compliance could be due to a number of reasons. Is it due to an attention stage failure because people did not notice the warning? Or, did people notice it, but the problem is they did not comprehend it? And so on through the stages of C-HIP. One of the benefits of the C-HIP model is not only to aid in understanding the processes involved, but also to serve as an investigative or diagnostic tool to assess the reasons for warning failure.

Consider, as an example, that a manufacturer finds that a critical warning on their product label is not performing adequately to prevent accidents. The first reaction to solving the compliance problem might be to increase the size of the sign or label so more people are likely to see it. But noticing the warning (the attention switch stage) might not be the problem. Potentially, user testing could show that all users report having seen the warning (attention switch stage), read the warning (attention maintenance stage), understood the warning (comprehension and memory stage), and believe the message (the beliefs and attitudes stage). If so, the problem with the manufacturer's warning in this case may be at the motivation stage—users may not be complying because they are not considering the severity of injury that might result. Thus, a solution according the C-HIP model would be to enhance the motivational characteristics of the warning, for example, by describing the severity-of-injury consequences more explicitly. By using the C-HIP model as an investigative tool, one could potentially track down the failure (similar to detective work) and make the appropriate corrections. One reason for noting this benefit of C-HIP is that it can systematize the investigation, rather than making attempts at fixes more or less blindly.

OTHER CONSIDERATIONS

There are several important methodological considerations when conducting research on warning effectiveness. This section discusses two.

Data Type

Subjective measures are relatively easy to administer. Some constructs such as beliefs/attitudes can only be acquired directly by asking participants (Burt, Bartolome, Burdette, & Comstock, 1995). Subjective measures are sometimes confused with qualitative measures, because subjective measures can produce quantitative or qualitative output. The use of the term *subjective* in social science research methods simply refers to the origins of the data output. When participants report their personal experiences, this output is referred to as subjective data. However, subjective data may take the form of quantitative or numerical outputs or qualitative, non-numerical outputs. A Likert rating scale (Likert, 1932) or semantic differential rating scale (Osgood, Suci, & Tannebaum, 1957; Snider & Osgood, 1969) both produce quantitative outputs derived from participants' subjective ratings. Open-ended questions, however, would produce qualitative, non-numerical subjective data. In some instances, researchers will assign frequencies to recurring concepts or ideas in qualitative data. When frequencies are assigned, these data are then classified as quantitative. Objective or performance-based measures are measures that are reported by someone or something (e.g., an observer using a stop-watch) that is external to the participant. Subjective measures can be quantitative (e.g., ratings) and objective measures can be qualitative (e.g., verbal descriptors assigned by observers).

Validity and Reliability

If not designed well, subjective measurement instruments may not provide valid and reliable data on relevant user constructs. Validity and reliability of objective measures such as performance (observed compliance, task completion time, etc.) require their own careful design. But, it can be even more challenging to design subjective measures that meet these standards. For example, a common problem with subjective measures that may undermine reliability is the context effect (Colle & Reid,

1998). Context effects can occur if participants are shown several warning designs and are asked to rate each design on some dimension. In a situation where all warning designs are poor, some designs that are only slightly better than the others may receive higher ratings (and possibly very high ratings) because of participants' tendencies to provide relative judgments. Context effects can only be eliminated by using between-subjects designs, where users see only one of the test designs. However, between-subjects designs introduce other kinds of problems, such as needing many more participants. Using within-subjects design, or allowing all participants to rate all designs, will lead to context effects. However, context effects can be minimized by giving careful instructions to participants to emphasize the importance of rating each warning design independently. Also, the order in which warnings are presented to participants can be randomized or counterbalanced.

Validity and reliability are important to the development of meaningful subjective measures. There are many types of validity (predictive, construct, face, and statistical conclusion, for example) and there are many types of reliability (test-retest, internal consistency, and inter-rater, for example). Discussions of these types of validity and reliability can be found in most basic social science research methods and survey methodology texts. However, one often neglected attribute of subjective measures is social validity. Social validity is the acceptability and meaningfulness of a measurement or treatment process or method and its acceptability and relevance in a specific social context (Finney, 1991; Kennedy, 1992). Certain types of subjective measures may not be meaningful to specific groups of users or in specific contexts. For example, when warnings are tested with samples drawn from the general consumer population, can we be certain that the subjective measures used to elicit self-reports are valid for these groups? Some consumer groups may not be as "survey savvy" as others. Preliminary testing should be considered to determine whether there might be problems in participants' understanding the questionnaire as intended. For example, the rated dimension or the scale itself may not be meaningful to some consumers. Open-ended questioning will likely give more and better feedback from respondents with greater validity with respect to important warning constructs.

CONCLUSIONS

This chapter provided an overview of methodologies used to identify bottlenecks that undermine the effectiveness of warnings, from the perspective of the receivers. The C-HIP model provides a useful framework to select methods and to diagnose problem areas of a warning based on the results. Attention switch and maintenance can be measured by using response time, eye movements, and looking behavior. These methods provide data to draw inferences about the noticeability of a warning.

Memory and comprehension are generally measured using recall and recognition tests. In addition, open-ended tests or symbol identification tests provide data on the extent to which warning information is retained (remembered after exposure) or understood. Subjective measures are sometimes used in memory and comprehension, but they are not preferred compared to objective measures. However, subjective measures are the main method for assessing attitudes and beliefs and are used most frequently in assessing motivation. Thus, when available, objective performance measures are preferred over subjective ones. There is this choice in measuring attention and memory/comprehension, but objective performance measures are generally unavailable for measuring attitudes/beliefs and motivation.

Careful planning and design are strongly recommended when conducting studies to assess warning effectiveness. The outputs of studies on warning effectiveness are only as good as the research designs that are employed. Additional considerations such as validity and reliability are important to consider, and steps should be taken to ensure that the methods used will yield results that are useful to the design or evaluation of warnings.

ACKNOWLEDGMENT

The authors would like to thank Dr. S. David Leonard for his support in the development of this chapter.

References

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84,* 888-918.

ANSI (2002). Z535.1-5, *Accredited Standards Committee on Safety Signs and Colors*. Washington, DC: National Electrical Manufacturers Association.

Barzegar, R. S., & Wogalter, M. S. (1998). Intended carefulness of voiced warning signal words. In *Proceedings of the 42nd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1068-1072). Santa Monica, CA: Human Factors and Ergonomics Society.

Brugger, C. (1999). Public information symbols: A comparison of ISO testing procedures. In H. J. Zwaga, T. Boersema, & H. C. M. Hoonhout (Eds.), *Visual information for everyday use: Design and research perspectives* (pp. 305-313). London: Taylor & Francis.

Brannon, L., & Feist, J. (1992). *Health Psychology, (2nd ed.)*. Belmont, CA: Wadsworth.

Braun, C. C., Holt, R. S., & Silver, N. C. (1995). Adding consequence information to product instructions: Changes in hazard perceptions. In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 346-350). Santa Monica, CA: Human Factors and Ergonomics Society.

Burt, J. L., Bartolome, D. S., Burdette, D. W., & Comstock, J. R. (1995). A psychophysiological evaluation of the perceived urgency of auditory warning signals. *Ergonomics, 38,* 2327-2340.

Bzostek, J. A., & Wogalter, M. S. (1999). Measuring visual search time for a product warning label as a function of icon, color, column, and vertical placement. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 888–892). Santa Monica, CA: Human Factors & Ergonomics Society.

Colle, H. A., & Reid, G. B. (1998). Context effects in subjective mental workload ratings. *Human Factors, 40,* 591–600.

Collins, B. L., Dahir, M. S., & Madrzykowski, D. (1992). Visibility of exit signs in clear and smoky conditions. *Journal of the Illuminating Engineering Society, 21,* 69–84.

Dewar, R. E. (1976). The slash obscures the symbol on prohibitive traffic signs. *Human Factors, 18,* 253–258.

Dewar, R. E. (1999). Design and evaluation of public information symbols. In H. J. G. Zwaga, T. Boersma, & H. C. M. Hoonhout (Eds.) *Visual information for everyday use: Design and research perspectives* (pp. 285–303). London: Taylor & Francis.

Finney, J. W. (1991). On further development of the concept of social validity. *Journal of Applied Behavior Analysis, 24,* 245–249.

Frantz, J. P., Rhoades, T. P., Young, S. L., & Schiller, J. A. (2000). Assessing the effects of adding messages to warning labels. In *Proceedings of the International Ergonomics Association/Human Factors and Ergonomics Society 2000 Congress* (pp. 4.818–4.821). Santa Monica, CA: Human Factors and Ergonomics Society.

Hicks, K. E., Bell, J. L., & Wogalter, M. S. (2003). On the prediction of pictorial comprehension. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 1735–1739). Santa Monica, CA: Human Factors and Ergonomics Society.

Horberry, T. J., Purdy, K. J., & Gale, A. G. (1997). Mind the bridge! Drivers' visual behaviour when approaching an overhead obstruction. In S. A. Robertson (Ed.), *Contemporary ergonomics* (pp. 110–115). London: Taylor & Francis.

Kennedy, C. H. (1992). Trends in the measurement of social validity. *Behavior Analyst, 15,* 147–156.

Krugman, D. M., Fox, R. J., Fletcher, J. E., Fischer, P. M., & Rojas, T. H. (1994, November/December). Do adolescents attend to warnings in cigarette advertising? An eye-tracking approach. *Journal of Advertising Research,* 39–52.

Laner, S., & Sell, R. G. (1960). An experiment on the effects of specially designed safety poster. *Occupational Psychology, 34,* 153–169.

Laughery, K. R., & Young, S. L. (1991). An eye scan analysis of accessing product warning information. *In Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 585–589). Santa Monica, CA: Human Factors and Ergonomics Society.

Laughery, K. R., Young, S. L., Vaubel, K. P., & Brelsford, J. W., Jr. (1993). The noticeability of warnings on alcoholic beverage containers. *Journal of Public Policy & Marketing, 12,* 38–56.

Leonard, S. D., Hill, G. W., & Karnes, E. W. (1989). Risk perception and use of warnings. *In Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 550–554). Santa Monica, CA: Human Factors and Ergonomics Society.

Leonard, S. D., Otani, H., & Wogalter, M. S. (1999). Comprehension and memory. In M. S. Wogalter, D. M. DeJoy, & K. R. Laughery (Eds.), *Warnings and risk communication* (pp. 149–188). London: Taylor & Francis.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 1–55.

Louch, J., Price, E. C., Esson, M., & Feistner, A. T. (1999). The effects of sign styles on visitor behaviour at the orangutan enclosure at Jersey zoo. *Journal of the Wildlife Preservation Trusts, 35,* 134–150.

Osgood, C. E., Suci, G. J., & Tannebaum, P. H. (1957). *The measurement of meaning.* Urbana-Champaign: University of Illinois Press.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuasion: Classic and contemporary approaches.* Dubuque, IA: Brown.

Sanders, M. S., & McCormick, E. J. (1993). *Human factors in engineering and design* (7th ed.). New York: McGraw-Hill.

Smith-Jackson, T. L. (2004). Cultural ergonomics: Exploring risk disparities. In L. J. H. Schulze (Ed.), *Proceedings of the International Society for Occupational Ergonomics and Safety, 18th Annual Meeting: Building Bridges to Health Workplaces* (pp. 106–109), Houston, TX.

Smith-Jackson, T. L., & Durak, T. (2000). Posted warnings, compliance, and behavioral intent. In *Proceedings of the 14th Triennial Conference of the International Ergonomics Association and the 44th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 115–118). Santa Monica, CA: Human Factors and Ergonomics Society.

Snider, J. G., & Osgood, C. E. (1969). *Semantic differential techniques: A sourcebook.* Chicago: Aldine.

Wogalter, M. S., Allison, S. T., & McKenna, N. A. (1989). Effects of cost and social influence on warning compliance. *Human Factors, 31,* 133–140.

Wogalter, M. S., Brems, D. J., & Martin, E. G. (1993). Risk perception of common consumer products: Judgments of accident frequency and precautionary intent. *Journal of Safety Research, 24,* 97–103.

Wogalter, M. S., DeJoy, D. M., & Laughery, K. R. (1999). Organizing theoretical framework: A consolidated communication-human information processing (C-HIP) model. In M. S. Wogalter, D. M. DeJoy, & K. R. Laughery (Eds.), *Warnings and risk communication,* (pp. 15–23). London: Taylor & Francis.

Wogalter, M. S., Desaulniers, D. R., & Brelsford, J. W. (1987). Consumer products: How are the hazards perceived? In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 615–619). Santa Monica, CA: Human Factors Society.

Wogalter, M. S., & Dingus, T. A. (1999). Methodological techniques for evaluating behavioral intentions and compliance. In M. S. Wogalter, D. M. DeJoy, & K. R. Laughery (Eds.), *Warnings and risk communication* (pp. 53–81) London: Taylor & Francis.

Wogalter, M. S., Godfrey, S. S., Fontenelle, G. A., Desaulniers, D. R., Rothstein, P. R., & Laughery, K. R. (1987). Effectiveness of warnings. *Human Factors, 25,* 599–612.

Wogalter, M. S., Kalsher, M. J., & Rashid. R. (1999). Effect of signal word and source attribution on judgments of warning credibility and compliance likelihood. *International Journal of Industrial Ergonomics, 24,* 185–192.

Wogalter, M. S., Murray, L. A., Glover, B. L., & Shaver, E. F. (2002). Comprehension of different types of prohibitive safety symbols with glance exposure. In *Proceedings of the Human Factors and Ergonomics Society? 46th Annual Meeting* (pp. 1753–1757). Santa Monica, CA: Human Factors and Ergonomics Society.

Wogalter, M. S., & Rashid, R. (1998). A border surround a warning sign affects looking behavior: A field observational study. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (p. 1628).* Santa Monica, CA: Human Factors and Ergonomics Society.

Wogalter, M. S., Young, S. L., Brelsford, J. W., & Barlow, T. (1999). The relative contribution of injury severity and likelihood information on hazard-risk judgments and warning compliance. *Journal of Safety Research, 30,* 151–162.

Wolff, J. S., & Wogalter, M. S. (1998). Comprehension of pictorial symbols: Effects of context and test method. *Human Factors, 40,* 173–186.

Young, S. L. (1991). Increasing the noticeability of warnings: Effects of pictorial, color, signal icon, and border, In *Proceedings of the 35th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 580–594). Santa Monica, CA: Human Factors and Ergonomics Society.

Young, S. L., & Lovvoll, D. R. (1999). Intermediate processing stages: Methodological considerations for research on warnings. In M. S. Wogalter, D. M. DeJoy, & K. R. Laughery (Eds.), *Warnings and risk communication* (pp. 27–52). London: Taylor & Francis.

Young, S. L., & Wogalter, M. S. (1990). Comprehension and memory of instruction manual warnings: Conspicuous print and pictorial icons. *Human Factors, 32*, 637–649.

Young, S. L., & Wogalter, M. S. (2001). Predictors of pictorial symbol comprehension. *Information Design Journal, 10*, 124–132.

Zwaga, H. J. (1989). Comprehensibility estimates of public information symbols: Their validity and use. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 979–983). Santa Monica, CA: Human Factors Society.