CHAPTER FOUR

Methodological Techniques for Evaluating Behavioral Intentions and Compliance

MICHAEL S. WOGALTER
North Carolina State University

THOMAS A. DINGUS
Virginia Polytechnic Institute and State University

The ultimate criterion of warning effectiveness is actual behavioral compliance. Given its importance, there are surprisingly few behavioral studies in the warnings literature, probably because their implementation is difficult. Instead, many studies use questionnaires to measure behavioral intentions to warning-related variables. While a link between behavior and behavioral intentions has been established in the social psychology literature, the association has not been confirmed in the warning literature. Nevertheless, sometimes questionnaire type studies that include measures of behavioral intention are the best one can do given limited resources. The main purpose of this chapter is to describe the techniques for examining behavioral intentions and actual behavioral compliance. Self-report, observational, physical trace and epidemiological methods are described. It is hoped that researchers will incorporate (or adapt) some of the techniques in future studies.

4.1 INTRODUCTION

The ultimate goal of warnings is to reduce personal injury and property damage. For this goal to be met warnings need to influence people so that they do not behave in ways that lead to personal injury and property damage. Chapter 3 by Young and Lovvoll describes methods that permit measurement of warning effects at the intermediate or pre-behavior stages of the communication-human information processing (C-HIP) model. While there is no doubt that the processes of attention, memory, and the other stages are important for effective warnings, it is the last stage of the model, the behavior stage, that is the most important. The occurrence of safe behavior is the ultimate measure of whether the

Warnings and Risk Communication, edited by Michael S. Wogalter, David M. DeJoy, Kenneth R. Laughery. Published in 1999 by London: Taylor & Francis. ISBN 0-7484-0266-7.

warning works. If a warning is effective at the behavioral stage, the warning is probably adequate at the earlier stages. Indeed, behavioral data are so important that if only one measure of warning effectiveness can be obtained, a compliance test is the best one to do. It is superior to all other tests. While other measures (e.g., of attention or memory) are capable of evaluating important aspects relevant to warning effectiveness, the effects may not always translate into behavior.

This chapter focuses on the methods of measuring behavioral intentions and actual behavioral compliance. Behavioral intentions usually are assessed by questionnaire, while behavioral compliance usually is assessed by observing whether warning-directed behavior occurs. The main goal of this chapter is to increase understanding of the techniques that have been used or could be used. By familiarizing researchers with the range of potential methods, we hope that this chapter will facilitate the conduct of future compliance research. Chapter 11 by Silver and Braun describes specific outcomes and conclusions from studies using these techniques.

The term 'behavior' can be interpreted very broadly. Frequently it is defined as being some observable, measurable overt response with respect to some internally or externally generated stimulus. This definition is quite general and could include just about everything. For example, this definition would permit questionnaire responses to be classified as behavior. In this chapter we take a more restricted view of behavior. We define behavior as whether people do what the warning asks them to do. Although behavioral intentions (e.g., judgments of how careful they would be, etc.) probably are related to warning pertinent behavior, we do not consider them the same as actual behavioral compliance.

Given that compliance is such an important outcome for warnings, one might expect that most research would measure it. In fact, there are relatively few behavioral compliance studies in the warning literature. Most studies on warnings use the kinds of measure and technique described in Chapter 3 by Young and Lovvoll, mainly rating scales and questionnaires. More surprisingly, very few of these questionnaire studies measure behavioral intentions, which is perhaps the closest indication of behavior that paper and pencil techniques are capable of measuring. Why do so few studies measure behavior when we know it is the ultimate criterion of warning effectiveness?

The main answer to this question is that behavioral compliance research and testing are difficult to do for various reasons. The most compelling reason is that it is unethical to expose research participants to hazards. For example, it would be improper to test warnings in true life threatening circumstances such as diving into shallow water. For example, one could not ethically remove a NO DIVING sign to make a comparison to warning-present conditions. More generally, it would be unethical to use any but the best possible warning in this and other hazardous situations. Most behavioral studies involve specially created 'hazardous situations' which appear realistic but where the safety of the research participants is always protected.

A second reason for the difficulty of conducting behavioral compliance tests is control. One may not have the opportunity to directly manipulate the conditions that prompt a hazard warning. Warnings for severe weather where the events are random and infrequent is an example. Also one can not usually measure compliance to products like drain cleaners and condoms in natural environments such as people's private homes.

Finally, beyond the above-mentioned problems, compliance tests frequently are prohibited because of limited resources and capabilities. Behavioral compliance research is time and labor intensive and expensive. Additionally, it is difficult to manipulate warnings printed on labels of products sold in the stores. Permissions and appropriate label making capabilities would be needed to conduct such studies. Limited resources is a

particularly common plight for university researchers studying warning issues as there are virtually no federal grants available to fund the research. However, funding for warnings should not be a problem for large companies who may be selling hazardous products. US tort law says that product manufacturers are responsible for providing adequate warnings for associated hazards that can not be (practically) eliminated. It would seem then that companies—with their significant financial resources, superior knowledge, and legal responsibility—would take steps to assess the adequacy of the warnings for their products through behavioral compliance testing.

We recognize, however, that situations do occur where it is not feasible to conduct behavioral testing. In some instances, the warning may be needed before behavioral testing can be conducted. However, generally it is still possible (and advisable) to perform follow-up warning effectiveness tests after the initial warning has been placed in the stream of use. Some of these 'post' tests could be behavioral in nature, and could assist in determining whether the warning is adequate or should be replaced with a better warning. Thus, while behavioral testing can not always be performed, we want to encourage its use whenever possible.

In the remainder of this chapter, we describe and comment on methodological issues associated with assessing behavioral intentions, followed by a similar but more extensive review of the techniques involved in behavioral compliance research.

4.2 BEHAVIORAL INTENTIONS

In this section we describe some of the methods that could be and have been used to measure behavioral intentions. Given that sometimes behavioral testing cannot be done, an important issue is whether behavioral intentions predict behavior. Considerable research in social psychology (Ajzen and Fishbein, 1977; Eagly and Chaiken, 1993, 1996; Kim and Hunter, 1993; Eckes and Six, 1994; Kraus, 1995) as well as in other fields such as medicine and health (Taylor, 1991; Brannon and Feist; 1992) indicate that behavioral intentions do predict behavior. In other words, what people say they intend to do reflects to varying degrees what they actually do. Not all studies show a relationship between intentions and actual measured behavior, but in general the bulk of the research indicates that prediction of actual behavior from intention judgments depends on whether the context/scenario/state-of-mind during which the intentions are taken are similar to the specific situation in which the behavior is to occur. The closer the match, the better the prediction. Nevertheless, the fact that some research shows no match between intentions and actual behavior should make interested parties somewhat uncomfortable when questionnaire/interview results are not subsequently confirmed by behavioral data of some type. Prudence would recommend withholding judgment until the results are backed up by other studies—preferably each using a different research methodology. Multiple methods showing the same effect would allow stronger statements regarding the generalizability of the warnings-related phenomenon.

Additionally, we would be more confident about the utility of behavioral intention measures with respect to warnings if there were a lot of research showing a direct tie to behavior. However, specific data on the relationship between behavior and behavioral intentions in the warnings domain is virtually nonexistent, and much of the supporting evidence is indirect. The necessary research would determine people's perceptions, beliefs and intentions *before* behavioral compliance is assessed. That is, behavioral intentions measurement needs to be made at a point in the task sequence before compliance

might take place. Behavioral intention data concordant with actual behavior would indicate useful prediction capability of the behavioral intentions measure. As we have said, no studies in the warnings to date have done this. Nevertheless, we believe that knowledge gained in other domains on this topic has at least some generalizability to the warning domain.

There is one side note to this issue. A number of warning compliance studies have had participants complete a questionnaire *after* the main behavioral compliance measurement phase is over. DeJoy (1989; see also Chapter 9) reviews several behavioral compliance studies that also include a questionnaire asking whether they noticed and read the warning. The data show an across-the-board drop in the percentages of participants who report noticing the warning, who report reading the warning, and who comply with the warning. Frequently these questionnaire measures show significant correlations with behavioral compliance (e.g., Friedmann, 1988; Jaynes and Boles, 1990; Otsubo, 1988; Wogalter, Godfrey, Fontenelle, Desaulniers, Rothstein, and Laughery, 1987; Wogalter, Kalsher, and Racicot, 1993b). One example of such data is by Otsubo (1988). For a task involving a circular saw, 74% of the participants noticed the warning, 52% said they read it, and 38% were observed to comply. On the other hand, for a task involving a jig saw, 54% of the participants noticed the warning, 25% read it, and 13% complied. Thus, we can see an indication that noticing and reading roughly concurs with compliance levels. However, besides the fact that it is problematic that these predictors are assessed after the event they are supposed to predict, there also is the possibility that compliance or noncompliance might influence how people answer the questions on the follow-up questionnaire. For example, participants who did not comply in the situation may subsequently respond that they did not see the warning when they actually did. Additionally, answers to the questionnaire items may be affected by social desirability and demand characteristics (e.g., they answer in the way that they think the experimenter wants them to). Thus, we can not say with certainty that the responses on the post-task questionnaire are accurate and useful because the noticing and reading measures are retrospective reports that are subject to various biases. More useful for predicting of behavioral compliance would be the recording of precursor events, such as looking behavior (whether people are observed looking in the direction of the warning and appearing to be reading it).

4.2.1 Methodology

Behavioral intentions go by different names in the literature: precautionary intent, cautionary intent, intended carefulness, likelihood of complying, and willingness to comply. In behavior intentions research, participants typically are asked one or several questions about whether they would comply with a warning for a particular product or environmental hazard. The questions sometimes request dichotomous 'yes' or 'no' answers, but most research employs Likert-type rating scales. Participants are asked for a judgment on the extent to which they would comply with a warning. The points on the scale range might range from 1 to 7 (or some other set of numbers) with the values labeled along some or all of the points of the scale with verbal anchors. For example, the end point anchors might be 'definitely would not comply' and 'definitely would comply' with the intermediate values labeled 'somewhat likely to comply,' 'likely to comply,' and so forth. The number of scale points can vary from study to study (and within studies). The lowest ratings on most scales is zero or one. In some studies, participants are asked to estimate the percentage of people who would likely comply with the warning (the number of

people out of 100 who would comply) which is similar to asking for a judgment along a 101-point (0–100) scale.

In attempting to predict behavior from behavioral intentions, some variability in the scores is necessary. Suppose that the scores on a behavioral compliance measure are all very low (near 0%) or all very high (near 100%). With very low or very high scores, it is not possible to show a statistical relationship with another variable—in this case, compliance with something else. In other words, if no one complies then you can not predict when compliance will occur from another variable. Some variability in the scores is needed to allow prediction.

It is important also that researchers do not misrepresent behavioral intentions research as behavior-based. Researchers should tell readers that the measurement involves intended or self-reported behavior. For example, Staelin (1978) uses the term 'actual behavior' to describe what is really a behavioral intention. To be fair, the point of Staelin's paper was to compare people's normative behavior (what they are supposed to do) to what they would personally do (the self-reported 'actual' behavior). Nevertheless, it is important for warning researchers to be specific and to use unambiguous terms. Readers of research articles should pay close attention to the study's method to be sure that the 'behavior' discussed is actual compliance behavior.

4.2.2 Other Behavioral Intention Measures

In this section we describe several other kinds of behavioral intention measures used in published research. Farid and Lirtzman (1991) used a very interesting behavior intention measure while assessing Egyptian workers' perceptions of different warning labels for hazardous chemicals. They assessed the workers' intention to quit, and found that workers exposed to high-hazard labels were significantly more likely to say that they would quit their jobs than workers exposed to low-hazard labels.

Intention to purchase is another behavioral intentions measure (Ursic, 1984; Silver, Leonard, Ponsi, and Wogalter, 1991; Laughery, Vaubel, Young, Brelsford, and Rowe, 1993). Sometimes this can be a highly appropriate way to assess how effectively a warning conveys the message that a particular product (e.g., an over-the-counter medication) is or is not appropriate for certain people with certain health conditions or that it should not be used in certain tasks and environments. If people say that they will purchase it for the right condition and not purchase it for a contraindicated condition, then this provides evidence that the labeling is doing its job. Research has also asked how much people would pay for a product with different warnings labels (Barlow and Wogalter, 1991; Wogalter, Forbes, and Barlow, 1993a).

4.3 BEHAVIORAL COMPLIANCE

Since the mid 1980s there has been solid growth in the number of published articles using behavioral compliance as a measure of warning effectiveness. These studies have been conducted in various creative ways and a description of the methods used in these studies is the main focus of the remainder of this chapter.

Careful thought and planning are required to create a situation, either experimental or observational, that provides interpretable behavioral compliance data. Some of the main approaches considered in compliance testing are described in the following sections.

4.3.1 Value Added

An important concept related to the influence of warnings is the extent to which a warning or a component of a warning adds value. One of the main reasons for including a control group (in which no warning is present) in experimental research is to determine the extent to which people would perform the safety behavior anyway, without the warning being present; the no-warning control condition provides a 'base' rate for the target behavior. A warning condition that shows higher levels of behavior than a control's base rate (with everything else held constant) in essence shows the value added by the warning. For example, consider the comparison between a condition without a warning (control) with a condition that has a warning added. If the control condition had 20% compliance and the warning condition had 35% compliance, the added value of including a warning is the difference between the two, or 15% compliance. Lehto and Miller (1988) call this effect 'efficiency,' which considers prior incident rate of the desired response and the rate after the warning is presented. Thus, the value-added/efficiency measure expresses whether and how much the warning makes a difference.

The added value of particular warning features can also be determined by comparing two experimental conditions that systematically differ on some dimension, e.g., comparing two colors like yellow versus blue. If a yellow warning has higher compliance than a blue warning, it indicates that the yellow color adds value to the warning's effectiveness relative to the blue color. Additionally, experiments can be designed so that warnings differ in multiple systematic ways so that one can determine quantitatively the relative value of the different features and the interactions among them. For example, Adams and Edworthy (1995) showed that the effect of changing character size produces a greater impact on effectiveness ratings than a change in border thickness. Assuming that this finding is confirmed in a behavioral compliance study, this result would have the implication that warning designers should take greater effort in trying to increase the size of the print than the thickness of the warning border. By having more than one factor in a single experiment, it is possible to see whether and how they interact. For example, one might find that combining larger sized characters with a thick border adds value beyond that expected by the simple linear addition of each of the component effects. Unfortunately, except for a few studies (e.g., Braun and Silver, 1995; Wogalter *et al.*, 1993a), researchers have not yet extensively employed experimental designs that can give the relative effect sizes between component factors (Cox, Wogalter, Stokes, and Murff, 1997). Our knowledge about warning design would be benefited greatly by research manipulating more than one factor in the same experiment. Such research would aid our understanding of the relative importance of certain warning features. See Edworthy and Adams (1996) for an extensive discussion of this point.

4.3.2 Incidental Exposure Paradigm

Many of the behavioral compliance studies in the literature use an incidental exposure experimental paradigm. In this paradigm, participants are not informed that the study deals with warnings. Participants are led to believe the purpose of the research is something other than warnings, i.e., a 'cover' story is given. The warning is presented to participants in the context of a set of tasks that they are trying to accomplish—it occurs incidentally, simulating how people are most often exposed to warnings in real life. For example, some behavioral compliance research studies have used a chemistry demonstration

task. In this protocol, participants are led to believe that the study concerns how people perform a series of steps in a chemistry laboratory demonstration procedure. Across the many studies that have used this paradigm, the warning is exposed in various systematic ways (e.g., in task instructions, on a sign, from a digitized voice recording) without any explicit mention (or implicit suggestion). The warning occurs as part of the situation in which the participants' goal is to measure and mix various chemical substances and solutions. Before and during these procedures, nothing is mentioned about warnings until the study is over when participants are debriefed about the true purpose of the research. In short, the incidental paradigm makes the experimental situation realistic in that people are trying to perform a set of tasks where a warning could be present.

The incidental paradigm is similar to work conducted in the human memory and social psychology literature. In the human memory literature, the incidental paradigm often is contrasted with an intentional memory paradigm. With intentional conditions, participants are told that they need to learn the material (e.g., explicitly told to memorize a list of words) and the point of the study is readily apparent to the participant. With incidental conditions (e.g., Craik and Lockhart, 1972), participants are led to believe that they are being exposed to the material for some other reason (e.g., to get their subjective, qualitative judgments of the material), and later they are given a surprise memory test of the material. Most questionnaire research on warnings could be categorized as being intentional, given that participants are asked explicitly to evaluate a set of warnings. Most behavioral compliance studies are incidental in that participants are led to believe initially that the research has some other (non-warning) purpose.

The incidental exposure paradigm does involve some level of deception. Ethics committees (Institutional Review Boards or IRBs) at universities and other organizations are sometimes concerned that deception is contradictory to the notion of 'informed consent,' a hallmark procedure enabling participants to play a role in choosing their own exposure to risk. Nevertheless, IRB committees will grant permission to conduct incidental exposure studies if (a) the situation is carefully planned so that there is virtually no risk of harm (e.g., the participant stopped prior to any hazardous action); (b) the study's rationale shows that the potential benefits of the research outweigh the costs of not immediately informing participants of the ruse (the benefit, of course, is that the results can help to produce better warnings, and ultimately, reduce injuries); and (c) a full/complete debriefing takes place immediately following the study's completion that describes the true nature of the study and what specific factors were being examined. Compared to certain areas of social psychology, the deception in warning studies is quite mild. IRB committees also like to see that participants' names and their performance are held in a confidential manner and that the consent form makes it clear that participants can discontinue their participation in the study at any time without penalty. With these safeguards in place, most oversight committees will approve the procedures.

Behavioral compliance studies can also make valid use of the intentional exposure paradigm. Consider the following hypothetical evaluation of warning effectiveness that is similar to that used in the work of Geller and associates (Geller, Casali, and Johnson, 1980; Johnson and Geller, 1984; Roberts and Geller, 1994). Suppose people are explicitly warned that compliance to a seat belt regulation will be recorded by hidden cameras and that failure to comply may lead to substantial fines (let us say $50 or so). This blatant intentional warning will no doubt be highly effective in getting people to wear their seat belts. One way to measure the warning's effectiveness in this scenario is to collect base line seat belt use prior to a warning announcement and then compare that wearing rate to a similar period after the warning is presented. More complex variations of this procedure

also could be employed. The point is that it may be appropriate to use an intentional exposure protocol in a behavioral compliance study in certain circumstances.

4.3.3 Participant Populations

Participants in warning research should be representative of the intended target population. However, in practice most studies fall short of this goal. One reason for this limitation is that it is difficult to bring nonstudents to university laboratory locations for a variety of cost and logistical reasons. Most laboratory-based behavioral compliance studies employ participants from a pool of undergraduate students taking introductory psychology courses for which there is usually a research participation requirement. It is important to take steps to ensure that the details of an incidental-type study do not get communicated to participants before they arrive at the laboratory. One can reduce contamination by asking participants during debriefing not to tell anyone about the study until after the study is completed. Also it helps to tell them why they need to withhold telling others—that it is important that future participants know very little about the study beforehand because otherwise it would affect the results adversely. With this information, most participants are willing to adhere to the request. Also, if one is using good experimental procedures like randomly assigning participants to conditions, then any 'compromised' participants will be equally distributed across conditions and should not bias the final outcome seriously.

4.3.4 Demand Characteristics

The concept of demand characteristics is important to behavioral compliance research because an otherwise well executed experiment sometimes can provide incorrect or misleading conclusions about the effects. Specifically, an experiment has demand characteristics when participants are forced to behave in a certain way because of the circumstances of the particular experimental situation. For example, suppose that research participants are told to use a hammer to accomplish some task. Suppose further that there is a warning on the hammer that says 'Do not use if handle is cracked,' and the only hammer provided has a cracked handle, then essentially the research participants are being encouraged to ignore the warning. Participants may believe that failing to do the task may jeopardize the receipt of course credit or promised monetary compensation associated with their participation in the study. Now let us consider a slightly different situation. Here the situation is identical to the one described above except that another hammer is available that does not have a cracked handle. Assuming the crack in one hammer is apparent and the warning on both hammers is prominent and conspicuous, the results would surely be different than the earlier-described single hammer study. Generally people will comply with warnings when it does not require much effort to do so. If there is no alternative hammer, then the only correct solution would be to discontinue participation in the experiment, which as we have said, might be perceived as 'costly'. In other words, certain characteristics of the situation can sometimes 'demand' participants behave in a certain way.

4.3.5 Data Collection Techniques in Behavioral Compliance Studies

Behavioral compliance research can be classified in a number of ways. We have already discussed one distinction: incidental versus intentional. Another categorization involves

the extent to which the research is laboratory-based, field-based or something in between. Laboratory studies tend to be (a) more highly controlled, and thereby have greater internal validity, (b) more sensitive to manipulations between conditions in the study, and (c) frequently involve the use of undergraduate students as participants in the research. Field studies tend to be (a) less well controlled, (b) less sensitive in detecting effects between manipulated variables except when many participants are involved, (c) more externally valid (i.e., concurs with real-life situations), and (d) tend to involve nonstudents (i.e., a wider range of participant demographics). Field studies tend to be more similar to real life situations than laboratory studies.

Some behavioral compliance studies are not so easy to categorize as 'laboratory' or 'field.' Thus, a study conducted at a shopping center mall involving dish washing cleaning solutions is an example in this gray area. Because the types of behavior are being studied outside of the normal place that dishes are washed, i.e., the home kitchen sink, we would not classify this research as a true field study or a true laboratory study. This type of study might be termed a quasi-field study (see also Cox *et al.*, 1997).

There are other ways to classify behavioral compliance research. In large part, most of the studies that have been performed to date have been experimental; that is, they have involved some explicit manipulation of variables by the researchers. This kind of study has the potential for giving the most solid cause–effect conclusions. Another kind of study is an *ex post facto* (after the fact) study. Several examples were mentioned briefly earlier in this chapter. Naturally occurring severe weather conditions are extreme circumstances where one has very little control over the conditions that prompt a warning. To study the effectiveness of such warnings, one might have to examine the effects after they were used. Different locales receiving different warnings would need to be matched with regard to relevant criteria so that a valid comparison can be made. Consider another example of *ex post facto* research, in this case the implementation of a law that requires a warning on a hazardous product. If no warning effectiveness measures were collected before the mandated warning is placed on the product, then the best a researcher can do to investigate the effectiveness of the warning is to find another population (matched on multiple demographic characteristics) that does not have the law and then make a comparison between the two groups. The basic weakness of an *ex post facto*-type study is the lack of researcher-controlled manipulation of conditions and absence of baseline data, and therefore, one can not draw strong conclusions about a factor's influence (despite extraordinary efforts to match conditions for everything except the variable under consideration) as would be the case with a tightly controlled experiment. Nevertheless, *ex post facto* studies sometimes make extremely valuable contributions to our knowledge that otherwise might not be obtained (see e.g., Greenfield and Kaskutas, 1993; MacKinnon, 1995; Mayer, Smith, and Scammon, 1991).

4.3.6 Method-oriented Taxonomy

Besides the above-mentioned categorizations of behavioral research, a method-oriented taxonomy is perhaps the most informative in the context of this chapter, because our primary objective is to delineate useful techniques for data collection and analysis. 'Method' in this context refers to both data collection and data analysis techniques. While most studies to date have used classical experimental methods, the actual implementation of these methods varies greatly with respect to design complexity, venue selected, measures collected and data quality. Though most studies have used classical experimental

techniques, some of the most valuable contributions have come from innovative, non-experimental, observations. The methodological categories include the following.

Self-reports of behavioral compliance. The data generated by this method are subjective, gathered via questionnaire or interview. Unlike behavioral intention data, however, they are collected *ex post facto* and therefore might be contaminated, as we pointed out earlier. Such reports are different and perhaps more reliable than intention data since the participant is not predicting future behavior, but is instead reporting previous behavior. Self-report data are especially useful in field research where often the act of compliance cannot be practically observed.

Observation of behavioral compliance. Many studies simply observe participants either complying with, or not complying with, a warning. Such observation can be made directly by human observers or indirectly by other means such as a video or still camera.

Physical traces of behavioral compliance. A number of innovative techniques have been employed which allow the objective measurement of compliance without direct observation of the behavior itself. This methodology has the advantage of unobtrusively measuring compliance, thus increasing the internal validity of the data generated.

Epidemiological analysis of compliance. This technique involves the use of prospective or retrospective analysis of objective data generated either from archival or observational sources. An advantage of the use of this method is that data from a very large sample, or an entire population, can be utilized.

We describe each of these categories in more detail in the sections below.

4.4 SELF-REPORTS OF BEHAVIORAL COMPLIANCE

In many cases, it is either impossible or detrimental to observe behavioral compliance directly. In some circumstances, the behavior simply cannot be observed directly. Such is the case when a study must be performed in a location to which the experimenter cannot gain access to observe the compliance behavior. For example, Dershewitz (1979) studied two groups to determine if mothers would use safety devices to 'safety-proof' their homes. The experimental group consisted of 101 families receiving health information on home safety-proofing. The control group consisted of 104 families. Each of the 205 families was given Kindergards (plastic locking devices for cupboards and cabinets) and electric outlet covers. Once the mothers were given the free devices, two methods were available to determine whether or not they were installed: direct observation and self-report. To effectively employ direct observation, participants must be willing to let the experimenter enter their home, so that he or she can observe the outlets and cabinets in question. Direct observation in this instance is also expensive and time consuming, because it involves scheduling visits to each of a large number of locations.

Another method used by Dershewitz (1979) is self-reports. Participants were asked via telephone whether they had installed the devices. The problems associated with self-reports are similar to those discussed previously for behavioral intentions. The data are not as accurate as directly observing compliance where, barring any observation error, scoring reliability is generally 100%. While pertinent data are limited, self-reports have been shown to match well with actual behavior. For example, a study by Hunn and Dingus (1992) compared self-report data to actual physical evidence of compliance in a consumer product warning scenario. The authors found evidence that self-reported compliance and physical evidence of compliance differed by less than 5%.

Self-report compliance measures have been used by researchers in a number of domains. Planek, Schupack, and Fowler (1972) studied the impact of the National Safety Council's defensive driving course (DDC) on over 8000 drivers from 26 states. These drivers self-reported their accidents and violations for the previous year by completing a questionnaire prior to participating in the course. One year after the program, graduates of the program responded to a similar questionnaire. State records were analyzed to assess self-report accuracy. The results showed substantial agreement in what people said and what was known to have happened from driving violation records.

Self-reports of compliance are used quite often in the health domain (Taylor, 1991; Brannon and Feist, 1992). Compliance, or adherence to medication or other treatments is often critical to health maintenance or recovery. Self-reported compliance has been shown to be useful when supplemented by other measures.

4.5 OBSERVATION OF BEHAVIORAL COMPLIANCE

The primary way to measure behavioral compliance is to see whether people follow the warning-directed behavior or types of behavior while engaging in some task. A critical feature is that the observation be unobtrusive; that is, the experimental circumstances, including the experimenter's act of observing or the presence of a camera, do not influence compliance.

4.5.1 Measuring Observed Compliance

In most behavioral studies, direct observation provides information on whether individuals performed the appropriate safe behavior. Usually whether the person complied or not is completely clear and easily observed and recorded. However, sometimes, the question of whether compliance has occurred is less clear, and additional methods must be employed to handle the ambiguity. These include (a) enhancing or tightening up the classification of what constitutes acceptable compliance versus noncompliance behavior, (b) training of the experimenters or judges so that they know specifically which behavior types are recordable, and/or (c) using two experimenters or judges to record the observations concurrently as a reliability check.

Sometimes only a single measure of compliance is recorded. For example, Wogalter, Racicot, Kalsher, and Simpson (1994) used an electronic LED sign that directed the participants to behave in a single safety-related manner: to put on gloves to protect against chemical irritants involved in the task that they were asked to perform. When the participants complied, they were given a score of '1,' and when they did not, they were assigned a score of '0.' This kind of scoring can be transformed readily to usable descriptive summary statistics: a mean of these numbers gives the proportion that complied (and multiplied by 100 gives percentage complied).

Some warnings have multiple directives. For example, the warning in Jaynes and Boles (1990) requested participants to wear (a) a mask, (b) gloves, and (c) goggles; and the warnings in Wogalter, Barlow, and Murphy (1995) described the proper connection of an external disk drive to a computer and directed participants (a) to turn off the computer, (b) to eject a protective transport disk from the external drive, and (c) to physically touch the metal connection on the back of the computer to discharge any static electricity present.

When there are multiple directives, the data may be scored for analysis in several ways. One very basic method is to determine whether participants did *everything* that the warning requested. Thus, for the two example studies cited above, the answer to the question depends on whether they behaved in all three ways. If they did, participants were given a score of '1,' but if only two, one or none of the three warning-directed types of behavior occurred, they are assigned a score of '0.'

Some studies employing warnings with multiple directives analyze each of the compliance behavior types separately. So for the examples given above, three separate sets of scores for each participant would be collected (one for each type of warning-directed behavior) and analyzed. In the Wogalter *et al.* (1995) study separate analyses were performed for the behavior pattern of turning off the computer, another analysis for the behavior pattern of ejecting the transport disk, etc. A potential benefit of analyzing separate compliance behavior patterns is that a richer and more complete picture of compliance-related behavior may be obtained.

Statistical analysis of '1' and '0' scores generally requires chi-square or nonparametric statistical tests to determine whether there are significant differences between conditions. This is the conventional method of evaluating categorical or nominal data. However, Cochran (1950) argues that analysis of variance (ANOVA) is a valid, reasonably robust test of binomial (dichotomous) data. There are two important advantages of using ANOVA techniques: (a) we can use conventional follow-up tests (e.g., simple effects, Tukey HSD) to compare the mean (proportions) for significant ANOVA effects, and (b) we can more easily detect interactions among simultaneously manipulated independent variables. Nevertheless, it should be noted that some conservative statisticians might object to using ANOVA for data of this type because of the very small added chance of error.

Dichotomous scores (e.g., yes/no, 0/1, etc.) are by their very nature limited. Large numbers of participants are sometimes needed to have sufficient statistical power to detect small differences between conditions. This can increase the study's costs dramatically. Another way of scoring compliance to multiple directives is to sum each participant's scores across the types of warning-directed behavior. Thus, in the example of the 3-behavior-type studies given above, if a participant complies with two out of three types of behavior then that person's score is a 2. if he/she complies with all or none of the directives, then these individuals are given scores of 3 or 0, respectively. So rather than having three sets of dichotomous data, we now have scores on a 4-point scale which can be analyzed more readily using more statistically powerful (more sensitive) analyses such as ANOVA.

When simply summing the scores for the individual warning-related types of behavior, the component scores are given equal weight. However, some types of behavior may be more serious than others. For example, breathing a chemical may be worse than touching it, and thus the failure to use respiratory protective equipment is worse than failure to wear gloves. In such instances, one could differentially weight each of the multiple directives according to their importance (which in this case is related to injury severity). The weightings are multiplied against the 1 and 0 scores before summing the values. Although differentially weighting the component scores is both logical and reasonable, we do not know of any warning compliance studies that have done this.

Sometimes compliance levels produce floor or ceiling effects. A floor effect occurs when scores across all conditions are very low (at or near the lower limit, e.g., 0% or 10% compliance rates). A ceiling effect is the opposite situation, when scores are all very high (e.g., from 80% to 100%). In either case it is possible to miss a true effect because the scores cannot move or differ by much between conditions. For example, if the base

rate behavior is at 90% without a warning being present, then it will be very difficult to show a statistically significant increase in compliance because the maximum increase that can be produced is 10%. We suggest that a pilot study with a limited set of participants be conducted to get an idea whether the compliance levels are near floor or ceiling levels, and if so, to make some adjustments to achieve a more moderate base rate of compliance, say 35–65%. Some nonresearchers have misinterpreted the results of studies showing moderate levels of compliance. Citing these levels, they suggest that warning experts are unable to design effective warnings because the best compliance levels in some experiments are not at or near 100%. However, they fail to understand that the research is designed and intended to delineate factors that make a difference. The absolute levels of compliance in research should not be assumed to be the maximum levels that can be achieved.

Another consideration is statistical versus practical significance. For example, suppose some research study shows that some factor produces a statistically significant effect. Suppose further that the factor is font type where the warning in one font produced greater compliance than another font. The statistical results, however, describe only whether the difference between the two conditions is not likely to be zero and that such a difference is repeatable (with a small margin of error). It does not give information on the variable's importance. With large samples, the size of the difference can be very small and still be statistically significant. Therefore the effect's practical importance can be minuscule relative to the effects of other influential factors. Nevertheless, it is difficult to determine how small a beneficial effect must be to lack practical significance.

In trying to determine the effects of an independent variable, we recommend that researchers use several measures of compliance. As we have already noted, some measures are more sensitive than others. For example, consider the study mentioned earlier (Wogalter *et al.*, 1995) in which the warning directed participants to turn off a computer, eject the transport disk and touch a connector plug to release static electricity while connecting one external disk drive. The particular compliance measure, involving whether participants do or do not turn off the computer, might be relatively insensitive to the warning manipulation. A possible reason why this measure might not show an effect is that people installing the disk drives might tend to turn the computer off regardless of *whether or not they see a warning* directing them to do so. The base rate of turning off the computer is high (near ceiling level) without the warning being present. Or in other words, the warning has little or no value to add in this case. However, other compliance measures (such as ejecting a transport disk and touching the connectors to limit static electricity) might show effects of the warning manipulation because their base rates are relatively low. Thus, if an experiment uses only one dependent variable and that measure turns out to be relatively insensitive to the warning feature being manipulated (e.g., its placement), the potential benefit of that feature might go undetected. The likelihood of such situations occurring and their potential repercussions on warning design criteria should not be underestimated. This illustration also points out why it is important to be cautious in interpreting null (nonsignificant) effects in research studies. With null findings, it is difficult to know whether the feature actually has no effect or whether the experiment was not sufficiently powerful to detect the feature's effect. Given that most behavioral experiments are costly to conduct, it is worthwhile to make the experiment as sensitive as possible.

Behavioral compliance studies usually are much less sensitive at detecting small differences between conditions as compared to studies using rating scales (e.g., those assessing behavioral intentions). Commonly, rating studies show differences that are not found in compliance studies (and, interestingly, it is hardly ever the other way around). This noncorrespondence has been mistakenly interpreted by some nonresearchers as indicating

ratings are of limited value with respect to measuring warning effectiveness. The greater sensitivity of ratings to subtle warning design differences is attributable partially to heightened statistical power derived from the measures essentially being composed of scores having a wider range of values (e.g., a 9-point rating scale) as compared to the dichotomous measures (e.g., yes/no scores) used in compliance studies. Furthermore, the sensitivity difference is attributable also to the heightened attention to small differences between warning designs in an intentional exposure rating study versus the much more subtle manipulation in an incidental exposure compliance study. To be as sensitive as ratings, compliance research would need to employ substantially larger sample sizes than is commonly used in studies of this type, making them even more expensive in terms of labor and time commitment. Rating studies often produce results similar to compliance studies with respect to the basic patterns of scores exhibited between conditions. The problem is when the rating and compliance results exhibit very different patterns of results between similar warning conditions. As we have said before, you should probably give greater credence to the results of a well designed behavioral compliance study compared to those of a rating study.

4.5.2 Other Observational Measures of Compliance

In this section we describe several other kinds of observational technique to measure behavioral compliance. One potential measure that has been mentioned in the warning literature (e.g., Wogalter *et al.*, 1987) is how many people decide to discontinue their participation in the study. This is an interesting measure because it might indicate how risky the situation appears to be. Different interpretations can be inferred from the different points at which participants decide to quit. For example, if some participants quit early in the experimental procedures (such as during the consent form phase when they are given preliminary instructions about the study) then this might indicate that participants believe that the situation has some believable level of risk. Presumably, something in the situation is causing people to decide that it is 'just not worth taking a chance of getting hurt.' In our experience the number of individuals who decide to discontinue participation is extremely low, making it difficult to show statistically significant differences. However, as we discussed earlier, the costs of quitting may be too high. Participants may worry that to do so would jeopardize their receiving course credit for their participation or some other incentive offered to them earlier. Additionally, they might hold the belief that scientists and their employer would not let injurious events occur to volunteer participants.

There are two other kinds of potentially useful behavioral indicants that have received relatively little attention in the research literature: (a) task sequencing, and (b) latency or speed of compliance. In some situations, it is important to perform certain actions in a particular order and to do them quickly to protect against injury or property damage. The logic is that a person who puts on protective equipment before getting involved with a potential hazard is acting more safely than a person who puts on the protective equipment while they are actually at risk. Most compliance studies count compliance as adequate only if it is performed *before* the performance of particular acts. Conversely, a person who dons a piece of protective equipment *after* initiating a risky act would not be counted as having complied. While responding in due haste is important in some situations, in others a more deliberate approach may be more appropriate. In these cases, longer latencies before engaging a potential hazard could indicate greater safety. A lock-out tag-out warning that is located to protect a disengaged power switch from being improperly engaged

(e.g., while maintenance or repair work is being performed) is an example where the appropriate response is to wait until there is assurance that there is no potential danger to anyone (or to the equipment) before the equipment is serviced and restarted.

4.5.3 Laboratory studies

In this section, we will describe some of the methods used to measure behavioral compliance in the laboratory. Most of this work has been done at university-based laboratories under highly controlled conditions.

Chemical hazard

The chemistry laboratory paradigm has been employed in numerous studies since the mid-1980s. The basic methodology has proven to be successful in demonstrating the influence of numerous factors on warning compliance, including the effect of location of the warning in a set of instructions (Wogalter *et al.*, 1987), cost of compliance and social influence (Wogalter, Allison, and McKenna, 1989), video modeling (Racicot and Wogalter, 1995), message personalization (Wogalter *et al.*, 1994), voice (Wogalter *et al.*, 1993b; Wogalter and Young, 1991), clutter (Wogalter *et al.*, 1993a), pictorials (Jaynes and Boles, 1990), color (Braun and Silver, 1995; Rodriguez, 1991), shape (Jaynes and Boles, 1990; Rodriguez, 1991), container label design (Wogalter and Young, 1994), and time stress (Wogalter, Magurno, Rashid, and Klein, 1998). Many of these effects using the chemistry paradigm have been supported using other methodologies, giving at least some indication that results from experiments using the chemistry paradigm can be generalized to other situations. Because of its appreciable use in investigating various warning-related factors, the basic methodology of this technique will be presented in more detail than other techniques that we review.

At the outset of the chemistry task procedure, participants are told that the research is an engineering psychology study designed to determine how people perform a set of steps involving the measuring and mixing of chemicals. The opening description is actually accurate (i.e., not really deceptive), but it does not refer to warnings being the real purpose of the study. In other words, participants are incidentally exposed to the warning as part of the overall task of using the chemicals.

In the initial overview, participants are told: (1) that they will be mixing and weighing a set of chemical substances and solutions, (2) that they should complete the laboratory task as quickly and as accurately as possible, (3) that they will have a limited amount of time to complete the task, and (4) that the final product will be evaluated for accuracy. All participants are then shown how to use a triple-beam balance on a nearby desk top.

A variety of chemistry equipment including: beakers, flasks, graduated cylinders, stirring rods, measuring spoons, disposable vinyl gloves and paper surgical masks are located on a laboratory table in an adjacent room. The substances and solutions are disguised to make them appear somewhat novel and potentially hazardous. For example, food coloring is combined with water to make solutions of different colors. Other containers hold substances of different colors and graininess, e.g., pink table sugar, corn meal, and yellow powdered sugar. Some studies have added a small amount of 'chemical'-type odor (e.g., ammonia) to help make the situation more believable to participants that they were mixing potentially hazardous chemicals. Figure 4.1 shows a typical chemistry laboratory set up.

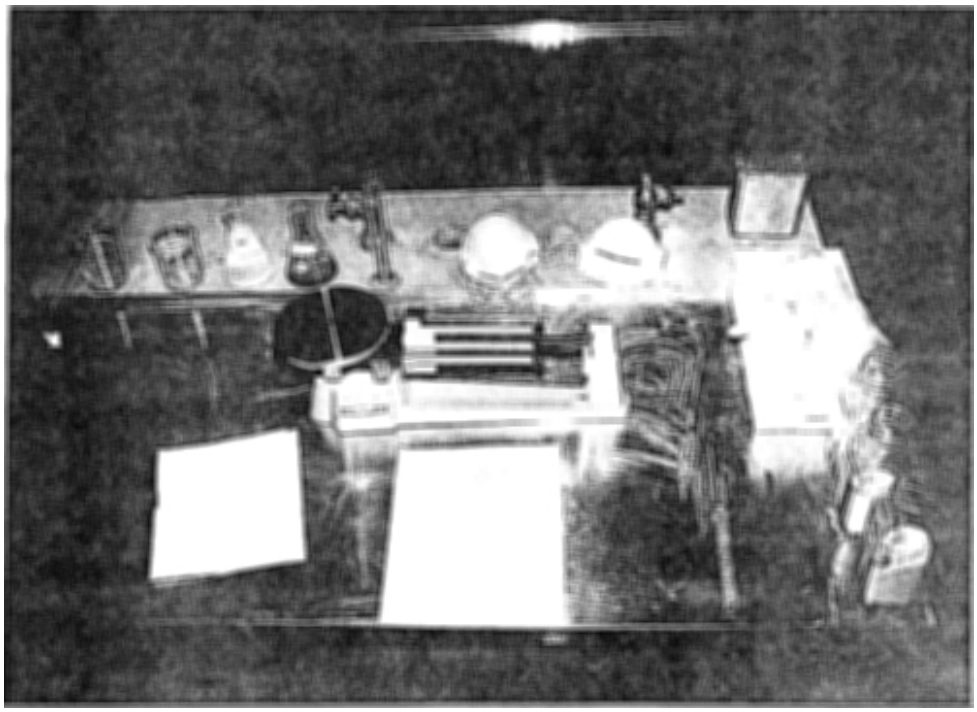

Figure 4.1 Typical chemistry laboratory set up (from Wogalter *et al.*, 1994).

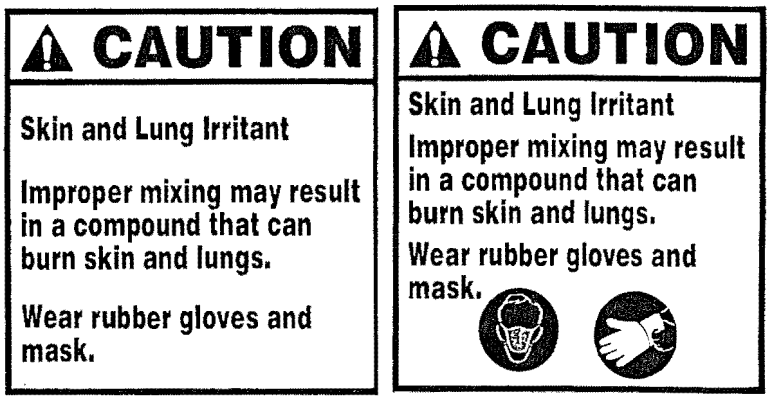

Figure 4.2 Example warning sign used in a chemistry laboratory experiment (from Wogalter *et al.*, 1993a).

At some point in the procedure participants are exposed to a warning (e.g., in the instructions, as a separate posted sign, from an audio tape player, or in a video, etc.) that might say the following: 'WARNING: Wear gloves and masks while performing the task to avoid irritating fumes and possible irritation of skin.' A depiction of an example sign is shown in Figure 4.2. The instruction sheet contains several steps describing how to measure and mix certain quantities of substances and solutions. The primary behavioral compliance measure is whether participants put on the mask and gloves before beginning to handle the chemicals. Some studies have also recorded the use of goggles and a lab coat.

Chemical products are particularly good 'hazards' to use in warnings research because participants cannot easily tell, just by looking at them, the true extent of risk. The difficulty of discriminating the potential hazardousness of chemical products can enhance participants' belief that they may indeed be working with a product that could be dangerous. Several studies have had participants use different kinds of chemically based consumer products in incidental exposure tasks. One popular item in warning research is glue. Glue is a good product because its fundamental purpose can be used as a foundation of the subterfuge task that participants are asked to perform. Strawbridge (1985) looked at the effect of several warning variables, including the embedded placement of warnings on labels of a glue product. The warning label stated that the glue contained acid and had to be shaken to avoid severe burns. The behavioral compliance measure was whether participants shook the container before using it. Hatem and Lehto (1995), while exploring the possible use of odor as a potential hazard cue, used a glue container that had the warning 'Danger: Toxic fumes may cause respiratory problems. Open all windows and doors before using. Turn on a fan if available.' The behavioral compliance measure was whether participants attempted to open a nearby window or turn on an accessible fan.

Other chemical products have also been employed. The chemicals used by Friedmann (1988) were a drain cleaner and wood cleaner. Participants' safety was protected by stopping the experiment after participants removed the lids (so they never actually worked with the substances). The primary measure was whether participants put on the provided safety equipment before removing the lid. Frantz (1993, 1994) measured participants' compliance while using a drain cleaner and water repellent sealer when examining the positioning of warnings with respect to the directions for use (integrated versus separated) and the explicitness of the procedures described. Compliance to the labeled warnings was measured and compared between conditions. Dingus, Wreggit, and Hathaway (1993) measured the use of gloves and mask for a 'newly formulated' household cleaner that participants were asked to try out at their home. Participants returned later with the materials, and compliance was measured according to whether there was a disturbance in the condition of the gloves (whether they were stretched) and the mask (whether a loose knot was untied).

An experiment by Chy-Dejoras (1992) involved a floor tile adhesive remover. The study was described as a marketing survey and participants watched a video of a person using adhesive remover. The behavior of the actor(s) in the video was manipulated. In one video, there was a single actor who used the product without gloves, and in another video, there were two actors, one of whom wore gloves and one who did not. Also, the aversive effects of the product were manipulated. In one video, the effects were benign. In a second condition, the depicted event was slightly aversive showing a person spilling the adhesive remover and vocally expressing pain. In the third condition, the video depicted a highly aversive experience showing the spilling of adhesive remover and pictures of a burned hand. After viewing the tape, participants prepared to perform a floor tiling task that required the use of the adhesive remover. The behavioral measure was whether or not they outfitted themselves with gloves before starting the tiling task.

Mechanical hazard

Several behavioral compliance studies have involved participants in tasks that require the use of tools and devices having mechanical injury risks. One tool that has been used is the power saw. Obviously such mechanical implements are dangerous, but also they are

somewhat familiar (at least to some participants). Setting up the situation so that participants are at very little or no risk requires careful planning.

Otsubo (1988) assigned participants to one of eight experimental conditions which corresponded to a factorial combination of two levels of product danger (high danger represented by a circular saw and low danger represented by a jigsaw), and four warning label formats (words only, pictograph only, words plus pictograph, and no warning). The study measured whether users donned gloves as directed by the warning label. Participants were stopped just before they used the saws as a precaution against possible injury.

Zeitlin (1994) arranged for four groups of college students differing in tool-using experience and exposure to safety training to use an electric chain saw to perform several tasks. Compliance with safety warnings contained in the chain saw operating instructions was measured. Participants actually used the saw in this study; but the author describes that adequate safety precautions were taken.

Research has involved other products that have a risk of mechanical injury. Frantz and Rhoades (1993) asked participants to unpack and arrange office furniture and supplies in a room. The office furniture and supplies included a file cabinet that displayed a label warning of a tipping hazard. Warning placement was manipulated: it was printed on the shipping carton, placed inside the bottom surface of the top drawer, placed on the front of the file cabinet as well as on the bottom of the top drawer, and placed in the top drawer on a piece of cardboard as well as on the surface of the drawer. The compliance measure was whether participants first placed materials in the bottom drawer to prevent the cabinet from tipping over.

Two other studies have examined the effectiveness of warnings for mechanical hazards. Dorris and Purswell (1977) employed a hammer with a cracked handle with or without a warning present that said the hammer should not be used if the handle is cracked. Another study, by Kalsher, Wogalter, and Silver (1998), had participants use a drill to construct parts of a bird house. The presence or absence of a tactile (raised border) warning telling users to wear gloves, a mask, and goggles was manipulated.

Electrical hazard

Several studies have involved electrical hazards. Duffy, Kalsher, and Wogalter (1995) led participants to believe that they were helping the experimenter set up some video and recording equipment for the 'real' study. In this context, participants connected the electrical cords to power outlets in which they were incidentally exposed to one of three warnings on the available extension cords. The resulting connection can be used to determine compliance, eliminating the need for direct observation. In this study, the effectiveness of an interactive label that required physical manipulation before it could be used was compared to a standard label. Compliance was based on whether the electrical cords were properly connected.

Gill, Barbera, and Precht (1987) had participants perform a series of tasks, one of which required them to use an extension cord to connect to an electric heater. The warning attached to the heater was the standard warning printed on the back of the unit, a color 'ski pass' label attached to the cord or a color-coded interactive label attached to the plug. The warning directed users not to use anything but a heavy-duty extension cord, whereas the only extension cord available in the room was light-duty. The number of people who used the inappropriate extension cord was recorded.

Wogalter *et al.* (1995) asked participants to connect an external disk drive to a computer. In this study, a short safety directive label was placed at various locations (on the

cover page of the manual, on the shipping box, in an accompanying leaflet, on the disk drive cable, and on the front of the drive). The label requested that installers first read the second page of an accompanying owner's manual which instructed them to take three precautions: (a) to turn off the computer, (b) to eject a shipping disk, and (c) to touch metallic plugs to discharge any static electricity. A similar study by Conzola and Wogalter (1999) using the same task manipulated the presentation of voice versus print warning/directives. Unlike most warning compliance research, the risk in this study was product damage, not personal injury.

4.5.4 Field and Quasi-field Studies

Several studies have observed compliance to warnings in field settings. As previously discussed, field studies have the advantage of increased external validity, but often at the expense of experimental control.

A common field research technique is to collect data from large numbers of consumers in a shopping mall. For example, Venema (1989) studied 330 participants visiting a home exhibition. Participants were asked to perform tasks involving methylated spirits (methyl alcohol). In one task, participants were asked to assume they were having a fondue and needed to refill the burner. In a second task, they were to assume they needed to remove paint from a table in their house using the spirits. Three versions of labels were studied: a neutral label with no safety information, the current label used for each product, and an improved layout constructed in accordance with recommendations found in labeling standards. The degree to which participants read and followed precautions stated in the label was observed.

Wogalter and Young (1991) observed 531 shoppers as they approached a simulated slippery-floor hazard near the entrance of the shopping center. Placed in the area were a set of orange cones and a mop inside of a bucket. There were four warning conditions: (a) none, (b) voice only, (c) print only, and (d) voice and print combined. Both the print and voice warnings stated 'Warning! Wet Floor. May be Slippery.' The voice (when present) emanated from a tape recorder inside a nearby mop bucket. Compliance was based on the proportion of individuals who walked through the area avoiding a specific section of the floor near the cones.

Other studies have employed similar, unobtrusive observational techniques in public areas of buildings. Wogalter *et al.* (1989) examined the use of stairs in a college dormitory where a warning was posted indicating that the elevator was broken. In another elevator-use study, Wogalter, Begley, Scancorelli, and Brelsford (1997) measured compliance to various signs directing individuals to use the stairs if they were only going up one floor or down two floors so that elevator users traveling between more distant floors would have better service. The researchers rode the elevators of six multi-storey buildings for specified intervals and recorded the numbers of persons who failed to comply with the signs.

Wogalter *et al.* (1987) describe several field studies utilizing unsuspecting 'participants' in public buildings. These studies measured: (a) the use of telephones and copy machines when warnings stating that the machines were broken were present or absent; (b) the use of a water fountain having an enhanced versus unenhanced contaminated-water warning; and (c) the use of exit doors in the presence or absence of broken-door warnings. In all of these, the researchers recorded the numbers of people who complied or did not comply.

Reisinger and Williams (1978) conducted a hospital study testing three educational/persuasive programs designed to increase the crash protection of infants in cars by increasing the use of infant car seats. The behavior of the women participating in the programs (all of whom were new mothers) was then compared to new mothers who received no crash protection education. A total of 1200 babies were observed during the study; the three program groups, as well as the control group, had about 300 subjects in each condition. Compliance was measured in terms of whether the infant car seats were positioned properly and securely in the car.

Field studies have been conducted also in work environments to test the effectiveness of a variety of safety programs. Zohar, Cohen, and Azar (1980) administered hearing tests to selected workers in a noisy metal fabrication plant and gave some of them feedback that they had incurred noise-induced shifts in hearing sensitivity during their work shifts. Over a period of five months, the use of hearing protectors by workers receiving this feedback was compared to a matched control group who did not receive the feedback.

Gomer (1986) conducted a field study in the context of litigation that was directed toward measuring the effectiveness of a label which warned about the risk of delayed lung disease. The study attempted to reconstruct the conditions and labeling requirements corresponding to the state-of-the-art in the mid-1960s. Seventeen employees handled bags of limestone in a dusty environment over a period of two days. On the second day, strong warnings of the hazard of limestone dust were placed on the bags that recommended respirators be worn. The number of workers who saw the warning and who requested respiratory protection was recorded.

Summala and Pihlman (1993) describe a safety campaign in which all 30 000 truck drivers in Sweden were sent a music tape that provided information about driving in work zones. The tape emphasized the concerns of workers in work zones regarding large vehicles that pass by too closely at excessive speeds. The study was conducted over a period of four months. Drivers were unobtrusively observed by camera, and vehicle speed and lane position in the work zone were measured. Figure 4.3 shows one of the scenes at a work zone.

Field research has been conducted successfully also in recreational settings. Hathaway and Dingus (1992) conducted a study investigating the effects of cost of compliance and warning information content in a racquetball venue. Cost of compliance consisted of two levels: in the high cost condition, no eye protection was provided, whereas in the low cost condition, eye protection was provided in a salient location just outside the court area. The warning information factor was comprised of three levels: (a) no warning provided, (b) an ANSI standard warning, and (c) an ANSI standard warning plus specific consequence information. The proportion of the 420 racquetball players who wore eye protection was observed unobtrusively.

Lehto and Foley (1991) conducted a field study of ATV operator behavior in six states that did or did not have helmet laws in 1988 and 1989. The use of helmets and other personal protective equipment was observed. Also recorded was: (a) the presence of warning labels; (b) the presence and enforcement of state regulations governing ATV use; (c) whether operator training courses had been taken; (d) self-reported reading of owner's manuals; and (e) operator attitudes.

Another study observed participants at a university automotive repair garage. Wogalter, Glover, Magurno, and Kalsher (1999) measured the effectiveness of warnings on battery booster cables to convey the proper procedure of connecting them to jump-start an automobile with a dead battery. In the context of several car-related service tasks, participants were asked to perform the jump-start procedure (both cars had realistic-appearing

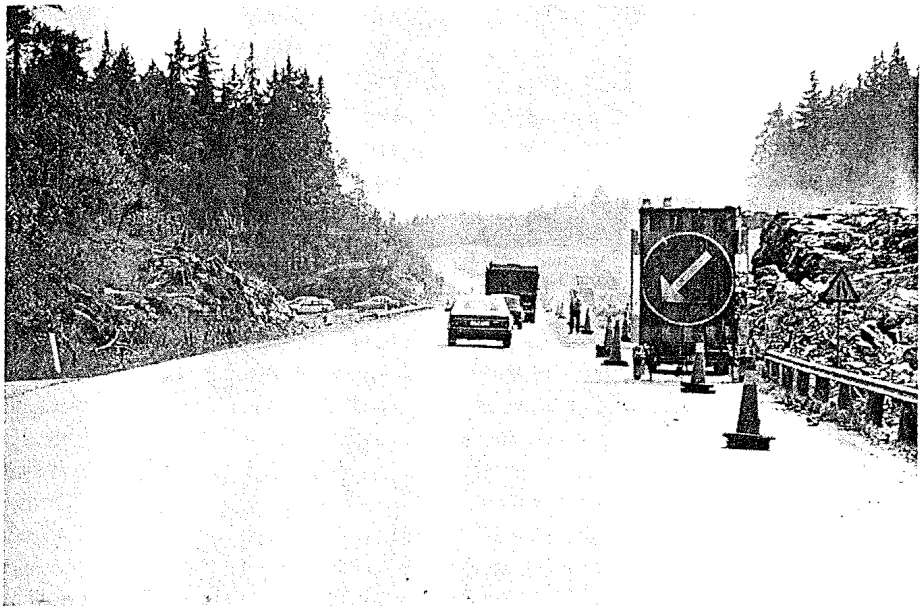

Figure 4.3 Swedish truckers were sent a cassette audio tape that included safety information about hazardous driving at work zone areas. Compliance was assessed by measuring the truck speed and distance of the vehicle to the side of the work zone (from Summala and Pihlman, 1993).

fake batteries). The warning (when present on the cables) contained verbal and pictorial information that described the hazards associated with car batteries as well as a pictorial diagram showing the sequence of steps that should be performed in the jump-start procedure. The number of people who properly connected the cables in the warning present and absent conditions was assessed.

4.6 PHYSICAL TRACE MEASURES OF COMPLIANCE

In most behavioral compliance studies, participants are observed more or less unobtrusively by the experimenter, because the presence of observers can influence compliance levels (Wogalter *et al.*, 1989). The best type of measurement would have no apparent observer and take place in a natural environment (i.e., where the product or equipment usually is used, such as in people's own homes).

One way to measure natural compliance behavior is to measure physical trace data. Physical trace data refer to any change or 'signature' in the environment associated with the compliance situation. An early example of this approach was to use 'glue-sealed' pages in magazines to assess advertising exposure (Politz, 1958 as cited by Ramond, 1976). Between each pair of pages in a magazine, a small glue spot was placed inconspicuously near the binding. The glue was configured such that it would not re-adhere once broken. Advertising exposure was then measured by counting the percentage of pages with a broken seal.

A primary advantage of physical trace measurement is that it can be used in field settings where direct observation is not feasible. A study by Hunn and Dingus (1992) illustrates this value. Compliance involved the use of protective gloves while using a cleaning product in a common household spray bottle. Participants were told that the

study dealt with the comparison of products of different strengths and qualities to their normal brand. The participants were told that they were testing a new cleaning formulation and could use as much of it as they wished. This study tested several factors including: two levels of information type and two levels of compliance cost (high cost: no gloves provided; low cost: gloves provided). The participants were instructed to take the packaged product home (in one half of the cases gloves were included in the package) and to use it for a week, at which time they were to return and complete a questionnaire asking about their experiences with the product. When they returned the package, they were given a questionnaire which contained a variety of product quality and marketing distracter questions, as well as label memory and compliance questions. In addition to questionnaire responses, physical trace measures were taken to verify whether the participants had worn the gloves and to ensure that they used the product on at least one occasion. Glove use was apparent from deformities at the finger-tips that occur after a very short period of use. Physical trace data of product use was assessed by the placement of a small paint dot on the threads of the spray bottle. If the paint dot was intact, it meant that the participant had not turned the bottle to the 'on' position, and therefore could not have used the product as intended. In all but a very few instances, the questionnaire responses indicating glove use were in agreement with the physical trace data.

In a follow-up study by Dingus *et al.* (1993), a different product was used and both gloves and respirator masks were provided as part of the packaging. The gloves and masks were prominently displayed in the consumer package so that participants would know of their availability. As with the Hunn and Dingus study, the consumers were informed that they would have to bring the contents of the package back to the same location after approximately one week had passed. A convenient time for the participant was noted and a majority of the participants did return to the mall. If a participant could not return to the mall at the specified time an alternative time was scheduled or the researcher arranged a time to pick up materials at the person's place of residence. In addition to the glove physical trace data, the mask elastic straps were tied in such a way that it was necessary for the participants to untie a simple knot to use the mask, thus giving a physical trace measure of use.

Trace measurements are in common use in health-related compliance (or adherence) research. Taylor (1991) describes the use of pill counts, that is, the amount of medication left in a bottle when the course of medication is supposed to be completed, to measure nonadherence. However, Taylor states that, despite their objective nature, pill counts are subject to several forms of bias. Patients may remove some pills from the bottle, or they may have pills left over from a previous treatment that they take instead. In addition, pill counts only estimate how many pills were removed from the dispenser, and not whether they took them at the correct times (Meichenbaum and Turk, 1987).

Brannon and Feist (1992) describe a number of automated devices to facilitate pill counting and to determine whether or not medication is taken at the prescribed time. Cramer, Mattson, Prevey, Scheyer, and Ouellette (1989) describe the use of a microprocessor in the pill caps to record every bottle opening and closing. The microprocessor yields information concerning the time of day that the bottle is opened, but does not detect the number of pills removed with each opening. Thus, this procedure provides more data than the pill-count technique, but still it cannot ascertain the exact rate of adherence. Brannon and Feist (1992) also examined biochemical evidence as a physical trace measure of compliance. Biochemical indices are detected through blood or urine samples. However, problems exist with the technique, including individual differences in absorption and metabolism, and the reliability, accuracy, and cost of the assays.

4.7 EPIDEMIOLOGICAL COMPLIANCE DATA

Epidemiology involves the distribution and determinants of disease or injury in a population. Epidemiological techniques are particularly valuable for assessing the value of injury control interventions (like warnings) in a population. Commonly, very large samples (or even entire populations) are analyzed in conjunction with naturally occurring changes or formal interventions. The data used for this type of analysis can be either archival (e.g., sales or accident records) or observational (e.g., the number of people seeking medical advice after a public service announcement). Epidemiological studies can be prospective or retrospective.

There have been a number of epidemiological studies which have provided valuable insight into the effectiveness of warning interventions by taking advantage of a change in legislative mandate. For example, Schucker, Stokes, Stewart, and Henderson (1983) evaluated the impact of the Saccharin Study and Labeling Act, passed by Congress in 1977, which required, among other things, that manufacturers place a warning on labels of products containing saccharin stating that: 'Use of this product may be hazardous to your health. This product contains saccharin which has been determined to cause cancer in laboratory animals.' With the enactment of the saccharin labeling requirement, an objective means of testing label effectiveness was created. That is, by monitoring the sales of soft drinks containing saccharin, the rate of warning compliance could be calculated for the entire population of diet soft drink customers. However, as is typical in this type of research, there is no opportunity for tight experimental control. For example, in addition to the warning labels, there were concurrent news reports and other information sources that were providing saccharin information. Thus, it became very difficult to assess the exact causal factors associated with any change in soft drink sales.

To evaluate the impact of the warning labels, Schucker and his colleagues developed a model that specified soft drink sales as a dependent variable and the presence of the warning, price of the product, news reporting and diet-drink advertising as independent variables. Seasonal sales trends were taken into account also. Orwin, Schucker, and Stokes (1984) used an auto-regressive moving-average modeling approach to evaluate the effect of the saccharin warning label on the sales of diet soft drinks. This technique allowed the authors to attribute any change in sales to specific causes.

A similar type of study was conducted in response to a national anti-smoking campaign. Warner (1977) evaluated the effects of the campaign on annual US per capita cigarette consumption. The anti-smoking campaign was a collection of mostly uncoordinated, educational activities by a variety of organizations including the government, private voluntary agencies, and for-profit business firms. To evaluate the effects, current cigarette consumption was compared to projections based on cigarette consumption prior to enactment of the anti-smoking campaign.

Several studies using large samples of the population have measured the impact of the alcohol warning label that is required on all containers in the USA since November, 1989. The measures have included behavioral intentions, awareness and memory of the label and its contents and changes in attitudes and beliefs (Greenfield and Kaskutas, 1993; Hankin, Sloan, Firestone, Ager. Goodman, Sokol, and Martier, 1996; Greenfield, Graves, and Kaskutas, 1999; Nohre, MacKinnon, Stacy, and Pentz, 1999).

Epidemiological methods have been applied to the motor vehicle domain in a number of instances. For example, Edwards and Ellis (1976) evaluated the effects of a driver improvement training program implemented by the Texas Department of Public Safety. The research studied the effect of the program on driving records and developed a

method for predicting the frequencies of violations and accidents for the 12 months following training. Robertson (1975) investigated the effectiveness of interlock and buzzer-light systems on the use of safety belts. The study was conducted at 138 sites in the cities of Baltimore, Houston, and Los Angeles, as well as suburbs of New York City, Richmond, VA, and Washington, DC. Use or non-use of safety belts by drivers was observed at each of these sites. To reduce any potential bias during the data collection phase, observers were 'blind' to (unaware of) the fact that buzzer-light and interlock systems were being compared.

Voevodsky (1974) studied the effectiveness of a center-mounted brake light as a means for preventing collisions under normal driving conditions. A total of 343 taxis operating with deceleration warning lights were compared to a control group of 160 taxis operating without the lights. After the light-equipped taxis had traveled a total of 12.3 million miles, rear-end collision rates were assessed.

Preusser, Ulmer and Adams (1976) studied compliance of drivers convicted of drinking and driving. Compliance in this case was the lack of a repeat offense. A program called the Nassau County Alcohol Safety Action Project Driver Rehabilitation Countermeasure ran from February, 1971 through June, 1973. The program's objective was to reduce the recidivism rate of drivers convicted of alcohol-related offenses. Random assignment of drivers to treatment and control groups was permitted by legislation. The experimental group consisted of approximately 3200 drivers who completed the rehabilitation program. The control group consisted of approximately 2600 drivers. The number of repeat offenses was measured and compared between these two groups.

4.8 SUMMARY AND IMPLICATIONS

In this chapter we reviewed studies that have measured behavioral intentions and compliance to warnings. In behavioral intention studies, participants make judgments of whether they would comply in a particular situation (or how careful or cautious they would be, etc.). Sometimes behavior intention studies are the best kind of warning assessment that can be obtained given the fact researchers cannot ethically expose participants to any type of substantial risk, and given practical considerations such as cost and time pressures. But when feasible, the best method of assessment is actual behavioral compliance. This chapter described various methods which have been used to conduct behavioral compliance research in laboratory and field settings. Most of these studies use some level of deception and use an incidental exposure paradigm where participants perform one or more tasks without being told that the study concerns warnings. Also described are techniques that examine physical trace indicators of compliance in naturalistic settings where direct observation is difficult or impossible. Finally, studies that have used epidemiological techniques are described. Because behavioral compliance is the ultimate measure of warning effectiveness, we hope that researchers will employ this measure more frequently in future research, and that this review will assist them in setting up future investigation, whether they make use of existing methods or create new ones.

Throughout this chapter, we have made recommendations with respect to the collection and analysis of warning compliance data. Several of the more general recommendations are worth summarizing here. We recommend that researchers collect data on several response measures in their studies whenever practical. Other kinds of measure, including subjective opinions, are valuable additions to the research literature because they can aid interpretation of compliance data.

Although direct behavioral measurement provides the most valid measure of warning compliance, collection of behavioral measures does not ensure that a study will provide valid and meaningful results. Critical aspects of the research design must be addressed to ensure successful evaluation of warning circumstances. These include unobtrusive measurement, an environment free of demand characteristics, and use of a scenario that does not contain inherent floor or ceiling effects. Generally a pilot study is recommended to ensure that critical features of the study are in order prior to actual data collection. Often even the most seasoned researchers are surprised by a particular study outcome while utilizing a new method or exploring a new content domain. In addition to the issues of measurement, the study must be carefully designed and must allow the determination of causes of compliance to be attained. Aspects of concern include: the use of proper baseline or control conditions and fair manipulation of selected variables, among others.

We reviewed some of the reasons why compliance measurement cannot be employed under certain circumstances. Behavioral intentions data can be substituted for measures of compliance, but specific research on the predictive significance of intentions in warnings applications is needed. Nevertheless, research in social psychology and other domains strongly suggests prediction is greater when the behavioral intentions data are assessed in situations that are similar to the actual compliance situation.

There are several important areas in the behavioral compliance research area that are likely to unfold in the next decade or so. One is the prediction of behavioral compliance, and the others are related to the rapid transition to powerful computers and people's interaction with them.

We expect that research will move towards more powerful models that predict behavioral compliance. With the variables that are discussed in this book, we already are able to predict and enhance compliance better than we were some 15 years ago, and we expect this trend to continue. Part of this will come from research started by Purswell, Schlegal, and Kejriwal (1986). They developed a questionnaire that was intended to measure risk-taking propensity which included items such as the percentage of time individuals used seat belts, whether they would use lifejackets when boating, and their reported tendency to cross a street against a light. In a set of tasks, participants were observed using a chemical drain opener, electric carving knife, sabre saw, and router. The researchers found that the questionnaire had significant value in predicting safe or unsafe behavior.

A second group of trends for research in this area will involve computer-based situations in which persons make risk decisions. These situations can be very lifelike considering that computers comprise a substantial portion of many people's lives. Thus, warnings during actual computer use procedures can be extremely real in the situation that they present to users. A substantial amount of work may be involved if a wrong decision is made (e.g., see Cox, 1995).

Perhaps the most exciting trend will be in the use of multimedia simulations of actual life events (other than computers) that can put participants in a 3D-like environment using a 2D computer or television screen (see Glover and Wogalter, 1997). In these simulations, individuals participate in a virtual environment (like those shown in sophisticated adventure games or in architectural design programs). Such programs, and even more sophisticated lifelike virtual reality environments, can put individuals into seemingly real risk environments without actual exposure to hazards (although it may appear that way). Using these programs, researchers will be able to place people into realistic hazard situations where warnings are present (in various conditions) and measurement can be made on whether they comply with them.

REFERENCES

ADAMS, A.S. and EDWORTHY, J. (1995) Quantifying and predicting the effects of basic text display variables on the perceived urgency of warning labels: tradeoffs involving font size, border weight and colour. *Ergonomics*, 38, 2221–2237.

AJZEN, I. and FISHBEIN, M. (1977) Attitude–behavior relations: a theoretical analysis and review of empirical research. *Psychological Bulletin*, 84, 888–918.

BARLOW, T. and WOGALTER, M.S. (1991) Increasing the surface area on small product containers to facilitate communication of label information and warnings. In *Proceedings of Interface 91*. Santa Monica, CA: Human Factors Society, pp. 88–93.

BRANNON, L. and FEIST, J. (1992) *Health Psychology*, 2nd Edn. Belmont, CA: Wadsworth.

BRAUN, C.C. and SILVER, N.C. (1995) Interaction of signal word and color on warning labels: difference in perceived hazard and behavioral compliance. *Ergonomics*, 38, 2207–2220.

CHY-DEJORAS, E.A. (1992) Effects of an aversive vicarious experience and modeling on perceived risk and self-protective behavior. In *Proceedings of the Human Factors Society 36th Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 603–607.

COCHRAN, W.G. (1950) The comparison of percentages in matched samples. *Biometrika*, 37, 256–266.

CONZOLA, V.C. and WOGALTER, M.S. (1999) Using voice and print directives and warnings to supplement product manual instructions. *International Journal of Industrial Ergonomics*, in press.

COX III, E.P. (1995) Compliance to product warnings. In *Proceedings of the Marketing and Public Policy Conference*. Atlanta: Department of Marketing, Georgia State University.

COX III, E.P., WOGALTER, M.S., STOKES, S.L., and MURFF, E.J.T. (1997) Do product warnings increase safe behavior? A meta-analysis. *Journal of Public Policy and Marketing*, 16, 195–204.

CRAIK, F.I.M. and LOCKHART, R.S. (1972) Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684.

CRAMER, J.A., MATTSON, R.H., PREVEY, M.L., SCHEYER, R.D., and OUELLETTE, V.L. (1989) How often is medication taken as prescribed? *Journal of the American Medical Association*, 261, 3273–3277.

DEJOY, D.M. (1989) Consumer product warnings: review and analysis of effectiveness research. In *Proceedings of the Human Factors Society 33rd Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 936–940.

DERSHEWITZ, R.A. (1979) Will mothers see free household safety devices? *American Journal of Diseases of Children*, 133, 61–64.

DINGUS, T.A., WREGGIT, S.S., and HATHAWAY, J.A. (1993) An investigation of warning variables affecting personal protective equipment use. *Safety Science*, 16, 655–673.

DORRIS, A.L. and PURSWELL, J.L. (1977) Warnings and human behavior: implications for the design of product warnings. *Journal of Products Liability*, 1, 255–263.

DUFFY, R.R., KALSHER, M.J., and WOGALTER, M.S. (1995) Interactive warning: an experimental examination of effectiveness. *International Journal of Industrial Ergonomics*, 15, 159–166.

EAGLY, A.H. and CHAIKEN, S. (1993) *The Psychology of Attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

EAGLY, A.H. and CHAIKEN, S. (1996) Attitude structure and function. In GILBERT, D., FISKE, S., and LINDZEY, G. (eds), *The Handbook of Social Psychology*, 4th Edn. New York: McGraw-Hill.

ECKES, T. and SIX, B. (1994) Fakten und Fiktionen in der Einstellungs-Verhaltens-Forschung: Eine Meta-Analyse [Fact and fiction in attitude-behavior research: a meta-analysis]. *Zeitschrift fuer Sozialpsychologie*, 253–271.

EDWARDS, M.L. and ELLIS, N.C. (1976) An evaluation of the Texas driver improvement training program. *Human Factors*, 18, 327–334.

EDWORTHY, J. and ADAMS, A. (1996) *Warning Design:A Research Prospective*. London: Taylor & Francis.

FARID, M.I. and LIRTZMAN, S.I. (1991) Effects of hazard warning on workers' attitudes and risk-taking behavior. *Psychological Reports*, 68, 659–673.

FRANTZ, J.P. (1993) Effect of location and presentation format on user processing of and compliance with product warnings and instructions. *Journal of Safety Research*, 24, 131–154.

FRANTZ, J.P. (1994) Effect of location and procedural explicitness on user processing of and compliance with product warnings. *Human Factors*, 36, 532–546.

FRANTZ, J.P. and RHOADES, T.P. (1993) A task analytic approach to the temporal placement of product warnings. *Human Factors*, 35, 719–730.

FRIEDMANN, K. (1988) The effect of adding symbols to written warning labels on user behavior and recall. *Human Factors*, 30, 507–515.

GELLER, E.S., CASALI, J.G., and JOHNSON, R.P. (1980) Seat belt usage: a potential target for applied behavior analysis. *Journal of Applied Behavior Analysis*, 13, 669–675.

GILL, R.T., BARBERA, C., and PRECHT, T. (1987) A comparative evaluation of warning label designs. In *Proceedings of the Human Factors Society 31st Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 476–478.

GLOVER, B.L. and WOGALTER, M.S. (1997) Using a computer simulated world to study behavioral compliance with warnings: effects of salience and gender. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, pp. 1283–1287.

GOMER, F.E. (1986) Evaluating the effectiveness of warnings under prevailing working conditions. In *Proceedings of the Human Factors Society 30th Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 712–715.

GREENFIELD, T.K., GRAVES, K.L., and KASKUTAS, L.A. (1999) Long-term effects of alcohol warning labels: findings from a comparison of the United States and Ontario, Canada. *Psychology & Marketing*, 16, 261–282.

GREENFIELD, T.K. and KASKUTAS, L.A. (1993) Early impacts of alcoholic beverage warning labels: national study findings relevant to drinking and driving behavior. *Safety Science*, 16, 689–708.

HANKIN, J.R., SLOAN, J.J., FIRESTONE, I.J., AGER, J.W., GOODMAN, A.C., SOKOL, R.J., and MARTIER, S.S. (1996) Has awareness of the alcohol warning label reached its upper limit? *Alcoholism: Clinical and Experimental Research*, 20, 440–444.

HATEM, A.T. and LEHTO, M. (1995) Effectiveness of glue odor as a warning signal. *Ergonomics*, 38, 2250–2261.

HATHAWAY, J.A. and DINGUS, T.A. (1992) The effects of compliance cost and specific consequence information on the use of safety equipment. *Accident Analysis and Prevention*, 24, 577–584.

HUNN, B.P. and DINGUS, T.A. (1992) Interactivity, information and compliance cost in a consumer product warning scenario. *Accident Analysis and Prevention*, 24, 497–505.

JAYNES, L.S. and BOLES, D.B. (1990) The effect of symbols on warning compliance. In *Proceedings of the Human Factors Society 34th Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 984–987.

JOHNSON, R.P. and GELLER, E.S. (1984) Contingent versus noncontingent rewards for promoting seat belt usage. *Journal of Community Psychology*, 12, 113–122.

KALSHER, M.J., WOGALTER, M.S., and SILVER, N.C. (1998) Behavioral compliance to tactile warnings using a bird house construction project scenario. Unpublished manuscript, Rensselaer Polytechnic Institute, Troy, NY.

KIM, M. and HUNTER, J.E. (1993) Attitude–behavior relations: a meta-analysis of attitudinal relevance and topic. *Journal of Communication*, 43, 101–142.

KRAUS, S.J. (1995) Attitudes and the prediction of behavior: a meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin*, 21, 58–75.

LAUGHERY, K.R., VAUBEL, K.P., YOUNG, S.L., BRELSFORD, J.W., and ROWE, A.L. (1993) Explicitness of consequence information in warnings. *Safety Science*, 16, 597–614.

LEHTO, M.R. and FOLEY, J. (1991) Risk taking, warning labels, training, and regulation: are they associated with the use of helmets by all-terrain vehicle riders? *Journal of Safety Research*, 22, 191–200.

LEHTO, M.R. and MILLER, J.M. (1988) The effectiveness of warning labels. *Journal of Products Liability*, 11, 225–270.

MACKINNON, D.P. (1995) Review of the effect of the alcohol warning label. In WATSON, R.R. (ed.), *Drug and Alcohol Abuse Reviews*, Vol. 7, *Alcohol, Cocaine, and Accidents*. Totowa, NJ: Humana Press, pp. 131–161.

MAYER, R.N., SMITH, K.R., and SCAMMON, D.L. (1991) Evaluating the impact of alcohol warning labels. *Advances in Alcohol Research*, 18, 706–714.

MEICHENBAUM, D. and TURK, D.C. (1987) *Facilitating Treatment Adherence: A Practitioners Guidebook*. New York: Plenum.

NOHRE, L., MACKINNON, D.P., STACY, A.W., and PENTZ, M.A. (1999) The association between adolescents' receiver characteristics and exposure to the alcohol warning label. *Psychology & Marketing*, 16, 245–259.

ORWIN, R.G., SCHUCKER, R.E., and STOKES, R.C. (1984) Evaluating the life cycle of a product warning: saccharin and diet soft drinks, *Evaluation Review*, 8, 801–822.

OTSUBO, S.M. (1988) A behavioral study of warning labels for consumer products: perceived danger and use of pictographs. In *Proceedings of the Human Factors Society 32nd Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 536–540.

PLANEK, T.W., SCHUPACK, S.A., and FOWLER, R.C. (1972) *An Evaluation of the National Safety Council's Defensive Driving Course in Selected States*. Chicago: National Safety Council Research Department.

POLITZ, A. (1958) *The Readers of the Saturday Evening Post*. New York: Curtis Publishing.

PREUSSER, D.F., ULMER, R.G., and ADAMS, J.R. (1976) Driver record evaluation of a drinking driver rehabilitation program. *Journal of Safety Research*, 8, 98–105.

PURSWELL, J.L., SCHLEGAL, R.E., and KEJRIWAL, S.K. (1986) A prediction model for consumer behavior regarding product safety. In *Proceedings of the Human Factors Society 30th Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 1202–1205.

RACICOT, B.M. and WOGALTER, M.S. (1995) Effects of a video warning sign and social modeling on behavioral compliance. *Accident Analysis and Prevention*, 27, 57–64.

RAMOND, C. (1976) *Advertising Research: The State of the Art*. New York: Association of National Advertisers.

REISINGER, K.S. and WILLIAMS, A.F. (1978) Evaluation of programs designed to increase the protection of infants in cars. *Pediatrics*, 62, 280–297.

ROBERTS, D.S. and GELLER, E.S. (1994) A statewide intervention to increase safety belt use: adding to the impact of a belt use law. *American Journal of Health Promotion*, 8, 172–174.

ROBERTSON, L.S. (1975) Safety belt use in automobiles with starter-interlock and buzzer-light reminder systems. *American Journal of Public Health*, 65, 1319–1325.

RODRIGUEZ, M.A. (1991) What makes a warning label salient? In *Proceedings of the Human Factors Society 35th Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 1029–1033.

SCHUCKER, R.E., STOKES, R.C., STEWART, M.L., and HENDERSON, D.P. (1983) The impact of the saccharin warning label on sales of diet soft drinks in supermarkets. *Journal of Public Policy and Marketing*, 2, 46–56.

SILVER, N.C., LEONARD, D.C., PONSI, K.A., and WOGALTER, M.S. (1991) Warnings and purchase intentions for pest-control products. *Forensic Reports*, 4, 17–33.

STAELIN, R. (1978) The effects of consumer education on consumer product safety behavior. *Journal of Consumer Research*, 5, 30–40.

STRAWBRIDGE, J.A. (1985) The influence of position, highlighting, and imbedding on warning effectiveness. In *Proceedings of the Human Factors Society 30th Annual Meeting*. Santa Monica, CA: Human Factors Society, pp. 716–720.

SUMMALA, H. and PIHLMAN, M. (1993) Activating a safety message from truck drivers' memory: an experiment in a work zone. *Safety Science*, 16, 675–687.

TAYLOR, S.E. (1991) *Health Psychology*, 2nd Edn. New York: McGraw-Hill.

URSIC, M. (1984) The impact of safety warnings on perception and memory. *Human Factors*, 26, 677–684.

VENEMA, A. (1989) Research Report No.69. *Product Information for the Prevention of Accidents in the Home and During Leisure Activities: Hazard and Safety Information on Non-durable Products*. The Netherlands: Institute for Consumer Research, SWOKA.

VOEVODSKY, J. (1974) Evaluation of a deceleration warning light for reducing rear-end automobile collisions. *Journal of Applied Psychology*, 59, 270–273.

WARNER, K.E. (1977) The effects of the anti-smoking campaign on cigarette consumption. *American Journal of Public Health*, 67, 645–650.

WOGALTER, M.S., ALLISON, S.T. and MCKENNA, N.A. (1989) Effects of cost and social influence on warning compliance. *Human Factors*, 31, 133–140.

WOGALTER, M.S., BARLOW, T., and MURPHY, S. (1995) Compliance to owner's manual warnings: influence of familiarity and the task-relevant placement of a supplemental directive. *Ergonomics*, 38, 1081–1091.

WOGALTER, M.S., BEGLEY, P.B., SCANCORELLI, L.F., and BRELSFORD, J.W. (1997) Effectiveness of elevator service signs: measurement of perceived understandability, willingness to comply, and behaviour. *Applied Ergonomics*, 28, 181–187.

WOGALTER, M.S., FORBES, R.M., and BARLOW, T. (1993a) Alternative product label designs: increasing the surface area and print size. In *Proceedings of Interface 93*. Santa Monica, CA: Human Factors and Ergonomics Society, pp. 181–186.

WOGALTER, M.S., GLOVER, B.L., MAGURNO, A.B., and KALSHER, M.J. (1999) Safer jumpstarting procedures with an instructional tag warning. In ZWAGA, H.J.G., BOERSEMA, T., and HOONHOUT, H.C.M. (eds), *Visual Information for Everyday Use: Design and Research Perspectives*. London: Taylor & Francis.

WOGALTER, M.S., GODFREY, S.S., FONTENELLE, G.A., DESAULNIERS, D.R., ROTHSTEIN, P.R., and LAUGHERY, K.R. (1987) Effectiveness of warnings. *Human Factors*, 29, 599–612.

WOGALTER, M.S., KALSHER, M.J., and RACICOT, B.M. (1993b) Behavioral compliance with warnings: effects of voice, context, and location. *Safety Science*, 16, 637–654.

WOGALTER, M.S., MAGURNO, A.B., RASHID, R., and KLEIN, K.W. (1998) The influence of time stress and location on behavioral warning compliance. *Safety Science*, 29, 143–158.

WOGALTER, M.S., RACICOT, B.M., KALSHER, M.J., and SIMPSON, S.N. (1994) The role of perceived relevance in behavioral compliance in personalized warning signs. *International Journal of Industrial Ergonomics*, 14, 233–242.

WOGALTER, M.S., and YOUNG, S.L. (1991) Behavioral compliance to voice and print warnings. *Ergonomics*, 34, 79–89.

WOGALTER, M.S., YOUNG, S.L. (1994) Enhancing warning compliance through alternative product label designs. *Applied Ergonomics*, 25, 53–57.

ZEITLIN, L.R. (1994) Failure to follow safety instructions: faulty communication or risky decisions? *Human Factors*, 36, 172–181.

ZOHAR, D., COHEN, A., and AZAR, N. (1980) Promoting increased use of ear protectors in noise through information feedback. *Human Factors*, 22, 69–79.