Do Product Warnings Increase Safe Behavior? A Meta-Analysis

Eli P. Cox III, Michael S. Wogalter, Sara L. Stokes, and Elizabeth J. Tipton Murff

In a meta-analysis of warnings literature, the authors examine empirical studies containing no-warning control groups to address the question of whether the presence of on-product warnings increases the safe behavior of product users. The major findings of the study are that (1) warnings increase safe behavior and (2) this increase is found for both nonstudent and student subjects.

An effective free-market system requires that consumers be enabled to make informed product purchase and usage decisions. Product warnings are an especially important information source designed to protect consumers and their property from physical harm. Warnings provide a proactive public policy alternative to reliance on the tort–liability system for the redress of consumer grievances or government intervention in which the ultimate action could be to recall or ban a product.

Warnings now appear on thousands of products as a result of manufacturers' concerns for user safety, fear of litigation, legal requirements, and industry standards. However, surprisingly little is known about their effectiveness in preventing accidents and injuries. Part of the problem is that only a few dozen empirical studies have addressed the behavioral effectiveness of on-product warnings. In addition, the seven qualitative reviews that have been published cover different portions of the literature and draw conflicting conclusions (Ayres et al. 1992; DeJoy 1989; Haddon 1986; Hardie 1994; Lehto and Miller 1988; McCarthy et al. 1984; Stewart and Martin 1994).

The first published studies evaluating on-product warnings yielded conflicting findings, and controversy over the behavioral effectiveness of on-product warnings continues today. Dorris and Purswell (1977) find that college and high school students uniformly ignored hammer warnings, but Schneider (1977b, p. 73), in his study of warnings directed toward preschool children, concludes that "package and label design can be effectively utilized to reduce attraction toward harmful substances."

Some writers continue to question the usefulness of warnings as safety mechanisms, arguing that in "a review of approximately 400 published articles ... no scientific evidence was found to support the contention that on-product warning labels measurably increase the safety of any product" (McCarthy et al. 1984, p. 81); "studies ... provide evidence that on-product warnings have not been effective in preventing injuries" (Horst et al. 1986, p. 111); "warning labels ... provide no guarantee that people will respond in the manner required by the label" (McCarthy et al. 1987, p. 483); "most warnings have little or no effect on safety; some compromise it" (Barnett and Switalski 1988, p. 11); and "identifying such situations and products [in which warnings are effective] has proven difficult, given the usual ineffectiveness of warnings" (McCarthy, Ayres, and Wood 1995, p. 2169). Together, these statements suggest that on-product warnings have little utility in the production of safe behavior.

Others have concluded that "a well designed warning can be effective in increasing warning compliance" (Strawbridge 1986b, p. 720); "warnings can be effective in modifying user behavior" (Friedmann 1988, p. 515): "product warnings can have a substantial effect on product use" (Frantz and Rhoades 1993, p. 729); "the use of mandated warnings on hazardous products seems to improve safety behavior over time" (Edworthy, Stanton, and Hellier 1995, p. 2153). Together, these statements support the argument that on-product warnings have utility in facilitating safe behavior.

Warnings research must answer two important questions. The first is whether an on-product warning is an effective communication medium in increasing the safe behavior of product users. If on-product warnings are effective, the second question is, What factors combine to produce an effective warning? We address the first and more fundamental of these two questions by conducting the first meta-analysis of the existing empirical research base. The results of this study are important to managers, policymakers, and researchers. On the one hand, if the results indicate that warnings are effective communication media, then managers and policymakers will be supported in their use of warnings as safety tools, and researchers will be encouraged to conduct the additional studies needed to understand the factors influencing the effectiveness of warnings in their thousands of applications. On the other hand, if the indication is that warnings are ineffectual, then managers, policymakers, and researchers must seek more effective alternatives for protecting consumers from potentially hazardous products.

First, we discuss the procedures followed in conducting the study, including the rationale behind the use of meta-analysis. Second, we discuss the analysis and results of the

Eli P. Cox III and Sara L. Stokes are professors, and Elizabeth J. Tipton Murff is a doctoral student, University of Texas at Austin. Michael S. Wogalter is Associate Professor of Psychology, North Carolina State University. The authors thank the four reviewers for their constructive comments on previous drafts of this article.

meta-analysis. We conclude with a discussion of the results and their implications.

Designing the Study

The procedures consisted of (1) deciding to use meta-analysis, (2) defining the type of study to be included in the analysis, (3) identifying and acquiring all studies meeting the specified criteria, (4) selecting the unit of analysis and coding the studies, and (5) identifying the most suitable meta-analysis procedures.

Choosing Meta-Analysis

First, we wished to conduct a review of warnings research that would avoid problems encountered in previous reviews by defining strict criteria for deciding which studies to include in the review, including all of the relevant studies, and evaluating them systematically using common criteria. These considerations all pointed to the use of meta-analysis.

Meta-analysis is a set of procedures for integrating findings from several empirical studies that examine a central research question and share a common criterion serving as the dependent variable. The dependent variable is typically a correlation coefficient (r) or the measure of experimental effect (d). The independent variables can be the design characteristics of the individual studies, the independent variables employed in them, or the experimental conditions of these studies.

Meta-analyses tend to take one of two approaches. The first approach, represented by Wolf (1986) and Hunter and Schmidt (1990), involves a detailed examination of the dependent variable, in which observations for sampling error and characteristics of study design that may have introduced error are adjusted. The goal is to obtain an "error-free" global estimate of the dependent variable. The second approach, outlined by Farley and Lehmann (1986), employs either regression or fixed- or random-effects ANOVA to examine the impact of the independent variables employed in the individual studies on a shared dependent variable. This approach enables the assessment of the relative importance of individual or categories of independent variables. The ability to combine within-study variation with between-study variation enables the researcher to examine larger sets of independent variables in more complex ways than is feasible in a single empirical study.

Meta-analysis is a quantitative alternative to qualitative literature reviews and has several advantages. First, there is a formal means of integrating the results found in the various studies, because independent variables can be judged in terms of the percentage of variation they explain in the dependent variable. Second, because a meta-analysis consists of a series of formal steps that can be specified in advance, it is replicable. This shifts the focus of discussion to recognized standards for conducting empirical research (e.g., adequate sample size) and away from the subjective assessment of findings characteristic of qualitative research reviews. Third, meta-analysis imposes discipline on researchers that forces them to judge the individual studies using common criteria. Such criteria often reveal omissions and inconsistencies in the individual studies that might not have been observed in a qualitative review.

Thus, meta-analysis can highlight gaps in the literature, direct further research, and examine mediating or interactional relationships that cannot be hypothesized or tested in individual studies (Wolf 1986, p. 55). Viewed from this perspective, meta-analysis is an important complement to qualitative research reviews and can make significant contributions in evaluating the state of the art and pointing to future directions of research. We believe that our meta-analysis, which examines the behavioral effectiveness of on-product warnings, makes an important addition to the seven qualitative reviews cited previously.

Defining the Studies to Be Included

Miller, Lehto, and Frantz's (1994) annotated bibliography of the warnings literature contains 785 entries dating from 1941. Although many of these articles and books make important contributions to the knowledge of warnings, most are not directly relevant to the issue of interest here. Many of the publications are qualitative and discuss legal and social dimensions of warnings, their psychological foundations, and industry warning standards and systems for designing warnings. The bulk of the empirical studies examine (1) other media, such as instructions, manuals, and signs; (2) self-reported compliance that cannot be verified; (3) other warning effectiveness criteria, such as noticing and reading warnings, attitude changes, and behavioral intentions, which may be necessary but are not sufficient conditions for consequent behavior; and (4) alternative forms of on-product warnings without including a no-warning control condition. This last category of studies aids our understanding of the factors influencing warning effectiveness but does not test whether a warning is better than no warning at all. Thus, only empirical studies comparing conditions with and without on-product warnings and using warning compliance as a criterion have been included in this analysis.

Acquiring the Relevant Studies

The search for studies meeting these strict criteria proceeded on several fronts. We searched electronic databases and bibliographies, examined the annotated bibliography and qualitative reviews mentioned previously, reviewed the citations of relevant publications, and contacted researchers. Published studies led to unpublished theses and dissertations. The results of this search are the 14 studies presented in Table 1. Venema (1989) reports two separate experiments, and thus 15 experiments constitute the 14 studies.

To illustrate the information contained in the table, the study by Schneider (1977b), which examines whether the presence of writing, a picture, color, the shape of the warning, and the product's fragrance altered the number of preschool children who opened a liquid-filled bottle, can be examined. The percentages in the table following the "C" in the Experimental Conditions column indicate the subjects who behaved safely in the absence of a warning, whereas the percentages following the "E" indicate those who complied with the warning(s) being tested. Schneider (1977b) finds that 44% of the children did not open the bottle when no warning was present, and those who complied in an experimental condition ranged from 33% to 89%. Thus, warning effectiveness varied across a range of 56 percentage points, with the least effective experimental condition resulting in a compliance rate 11 percentage points lower than the control and the most effective experimental condition producing a compliance rate 45 percentage points

Table 1. **Studies Evaluating On-Product Warnings Using a Behavioral Criterion and a No-Warning Control Condition**

Authors	Observations	Products	Hazards	Safe Behaviors	Experimental Conditions	Independent Variables	Main Effects > Control
Schneider (1977a)[a]	81 Children	Hazardous Chemicals	Poisoning	Not Opening Container	C: 44% E: 33–89%	Writing Picture Color Shape Fragrance	0 of 2[d] 2 of 2 n.a.[e] n.a.[e] 0 of 2
Strawbridge (1986a)	195 Students	Adhesive	Skin Burns	Shake Before Opening	C: 33% E: 20–60%	Position Imbeddedness[b] Highlighting	3 of 3 1 of 2 1 of 2
Gomer (1986)	17 Workers	Limestone	Lung Disease	Wear Respirator	C: 18% E: 35%	Warning Present	1 of 1
McCarthy et al. (1987)	50 Expectant Mothers	Infant Auto Restraint	Injury to Child	Proper Installation	C: 70% E: 48%	Warning/Instruction	0 of 1
Thyer and Geller (1987)	1722 Individuals	Automobile	Physical Injury	Wear Seat Belts	C: 34–41% E: 70–78%	Warning Present[b]	2 of 2
Otsubo (1988a)	131 Students	Circular Saw Jigsaw	Cut to Hand	Wear Gloves	C: 0% E: 13–50%	Hazardousness[b] Words/Pictographs	2 of 2 3 of 3
Venema (1989)	330 Adults	Paint Remover	Skin Burns	Wear Gloves	C: 87–96% E:.81– 94%	Warning Design	1 of 2
		Fondue Fuel	Flammable	Extinguish Flame Close Bottle	C: 30% E: 21–37%	Warning Design	0 of 4
Wreggit (1991)	224 Adults	Tile Cleaner	Skin Burns Harmful Fumes	Wear Gloves and Mask	C: 6–81% E: 0–100%	Format Compliance Cost[b] Interactivity	4 of 4 n.a.[e] 4 of 4
Lehto and Foley (1989)	269 All-Terrain Vehicle Riders	All-Terrain Vehicles	Head Injury	Wear Helmet	C: 79% E: 66%	Warning Present	0 of 1
Duffy, Kalsher, and Wogalter (1993)	120 Students	Extension Cord	Shock Fire	Do Not Overload	C: 0% E: 0–60%	Task Load Interactivity[b]	2 of 2 3 of 3
Hatem (1993)	38 Students	Glue	Respiratory Damage	Ventilate Area	C: 0% E: 0–9%	Text and Odor	1 of 3
Wogalter, Barlow, and Murphy (1994)	24 Students	Jumper Cables	Explosion, Shock, and Fire	Correct Connection	C: 0% E: 0–50%	Warning Present[b]	1 of 2
Wogalter and Kalsher (1994)	84 Students	Computer Disk Drive	Shock Self Damage Drive	Turn Off Ground Self Eject Disk	C: 50–58% E: 83–100%	Supplemental Directive[c]	6 of 6
Cox, Hoyer, and Krshna (1995)	47 Students	Computer Program	Caught in Program	Remember Command	C: 39% E: 71%	Redundant Warning[b]	1 of 1

[a] When a study appears in print more than once, all of the publications are listed in the bibliography and the publication with the earliest date is listed here.
[b] All results were significant at $\alpha = .05$.
[c] Some results were significant at $\alpha = .05$.
[d] "0 of 2" means that none of the two levels of the main effect resulted in compliance rates greater than that in the control condition.
[e] There was no null condition for these independent variables.

higher than the control. The last column indicates how many of the treatment levels for each independent variable examined in a study produced compliance rates greater than the control condition. Thus, both of the pictures evaluated by Schneider (1977b) increased warning compliance over the control condition where no picture was present.

Coding the Studies

The next stage involved selecting the unit of analysis and coding the individual studies. The two alternatives considered in choosing the unit of analysis for the studies were their experimental effects (main effects and interactions) and the individual experimental conditions. For the experimental-effect alternative, a study looking at the effect of color and shape on a warning would contribute three units of analysis: one for color, one for shape, and one for the interaction between the two variables. Using these experimental effects as the units of analysis is particularly helpful when the objective of the meta-analysis is to ascertain the relative importance of factors contributing to warning effectiveness.

The isolation of a single factor is not helpful, however, because the research question addressed here is whether the presence of a warning is better than no warning at all. Consider again the example of the study examining the effects of warning color and shape; it is possible that a positive main effect for color might be shown even if all test warnings were found to be inferior to the no-warning control because the main effect of color was offset by a negative main effect for shape or a negative interaction between color and shape.

Thus, experimental conditions were chosen as the unit of analysis for this study because they show the combined results of all main effects and interactions and can be compared directly with the no-warning control conditions. They relate directly to the objective of this meta-analysis, because each experimental condition is a specific set of circumstances in which a warning is present and its effectiveness can be examined.

For the three studies that employed more than one dependent variable (Venema 1989; Wogalter and Kalsher 1994; Wreggit 1991), we treated the additional measures of behavior as experimental conditions. This resulted in 79 experimental conditions for the 15 experiments, in which the dependent variable was expressed as the marginal compliance rate. For example, an experimental condition producing a compliance rate of 50%, where the rate of safe behavior when no warning was present was 70%, would yield a marginal compliance rate of –20 percentage points. The coding of the studies was done separately by two of the researchers. A few discrepancies were found revealing coding errors or judgment differences and these were eliminated before the research continued.

In Table 2, we present a summary of the 15 experiments. Twelve of 15 experiments were controlled; one employed an inequivalent control group design, and two utilized an interrupted time-series design (Cook and Campbell 1979). Less than half involved college students, and the combined sample size of these individual studies was 3229. Most employed only one independent variable. Twelve of the independent variables examined warning design alternatives, and two involved product characteristics and the environment in which the product is used. Differences among product users were not examined (except as a covariate in some studies). In only one instance was the safety information presented in an instruction format (which served as the no-warning control). In total, the 15 experiments comprise 24 control conditions and 79 experimental conditions.

Table 2. Classification of the Studies

Research Design	
Experiments	12
Quasi-Experiments	3
Subjects	
College Students	7
Others	8
Number of Independent Variables	
One	9
Two	3
Three or More	3
Type of Independent Variables	
Warning	21
Product	2
Product User	0
Usage Environment	2
Hazard Information Presentation	
Warning Format	15
Instruction Format	1
Observations	
Experimental Conditions	
Control	24
Experimental	79
Treatment Levels	
Control	24
Experimental	54
Sample Size	3229

Several general observations can be made regarding the 15 experiments. First, the percentage of subjects following the proper product usage procedures in the absence of a warning ranges from 0% (Otsubo 1988a) to 96% (Venema 1989). Second, warning compliance varies widely, with the marginal compliance rate ranging from –22% (McCarthy et al. 1987) to 60% (Duffy, Kalsher, and Wogalter 1993). Third, the absolute level of behavioral compliance in the presence of warnings ranges from 0% (Dorris and Purswell 1977) to 100% (Dingus, Hunn, and Wreggit 1991). Finally, these studies support Stewart and Martin's (1994) view that warning effectiveness is determined by characteristics of (1) the warning, (2) the product, (3) the usage situation, and (4) the user.

Designing the Meta-Analysis

We decided to combine the two alternative approaches to meta-analysis described in the previous subsection. In addressing the general question of whether on-product warnings are behaviorally effective, we chose to account for the differing sample sizes among the experimental conditions using a random-effects model. In this analysis, the effect size itself is viewed as a random variable, and the goal of the analysis is to estimate the mean and variance of the warning effect-size distribution. In attempting to identify study artifacts contributing to the variation in warning study results, we allowed for a portion of the variability in the observed effect sizes to be due to these artifacts and used weighted regression to estimate their impact.

Analysis and Results

Are Warnings Effective?

Descriptive Statistics

We present descriptive statistics of the marginal compliance rates before the meta-analysis because of the ease with which these statistics can be interpreted. To estimate the population effect of the marginal compliance rates, the 79 experimental conditions were examined. In Figure 1, we present a frequency distribution of these rates. These rates range from –21.4 percentage points to 60 percentage points, which means that the level of safe behavior was 21.4 percentage points lower than the control for the least effective condition and 60 percentage points higher than the control for the most effective condition. In 15 of the 79 instances, the presence of a warning was worse than no warning at all. In 11 instances, the presence of a warning failed to increase safe behavior, whereas for the remaining 53 experimental conditions, the presence of a warning improved safe behavior. The mean of this distribution is .157% with a variance of .037. Because the standard error is .022, the population mean falls within the interval of .114 to .200 at a confidence level of .95, which indicates that, in general, the presence of a warning results in an increase in safe behavior.

Meta-Analysis

The effect size of study i is defined generally as

$$\delta_i = \frac{\mu_i^E - \mu_i^C}{\sigma_i},$$

where μ_i^E and μ_i^C are the means of the dependent variable under experimental and control conditions, and σ_i is the within-group standard deviation. In our case, the means are the compliance rates for the experimental and control groups. In our random-effects approach, we assume that δ_i is a random variable. Thus, we assume that there exists a population of compliance rate effect sizes, of which the 79 experimental conditions constitute a sample. This population has a mean of $\bar{\Delta}$ and variance of $\sigma^2(\Delta)$, which are the parameters of interest.

Figure 1. Marginal Compliance Rates for the 79 Experimental Conditions

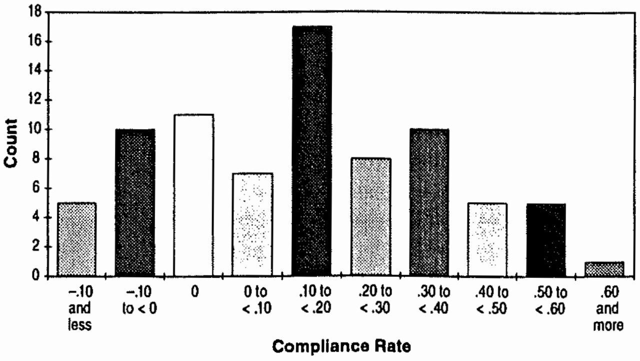

Table 3. Comparison of Meta-Analysis, Simulation Experiment, and Descriptive Statistics Results

Estimate	Meta-Analysis	Simulation	Descriptive Statistics
$\hat{\Delta}$	.311	.265	.157
$\sigma^2(\Delta)$	.055	.125	.037
$\hat{\pi}_{\delta < 0}$	.093	.208	.190

Applying the estimation procedures described in the Appendix to the data yields the estimate of $\hat{\Delta}$ as .311 and its estimated variance as .002. Because $\hat{\Delta}$ is approximately normally distributed, an approximate 95% confidence interval for $\hat{\Delta}$ is .220 to .402. We also estimated that 29% of the variation in the observed effect sizes was due to variation in effect sizes themselves, whereas the remaining 71% of the variation was due to sampling error.

The positive mean of the distribution of effect sizes does not indicate that all effect sizes are positive, as is evidenced by Figure 1. If this distribution has a mean of .311, a standard deviation of .220, and is normally distributed, then the estimate for the percentage of studies with negative compliance rates ($\hat{\pi}_{\delta < 0}$) equals approximately 9%.

One of the more serious limitations of meta-analysis is that by taking more than one observation from each study, the assumption of the independence of observations is violated. To address this problem, a simulation was run in which one experimental condition was drawn from each of the 14 studies, and the statistics $\hat{\Delta}$, $\hat{\sigma}^2(\Delta)$, and $\hat{\pi}_{\delta<0}$ were calculated. This procedure was repeated for a total of 120 simulations.

Table 3 compares the means for the three statistics with those obtained using all of the observations in a single analysis. The lack of independence among observations caused only a small overestimation of the mean compliance rate. In contrast, the variation among effect sizes was estimated to be approximately 2.3 times greater for the simulation than the estimate for the analysis of the full sample, as is expected because of the reduction in the number of observations. The estimate of the percentage of studies with negative compliance rates was approximately 21% for the simulation as opposed to the previous estimate of 9%, as is expected because of the greater variance for the simulation.

Several conclusions can be drawn by comparing the descriptive statistics with the meta-analysis and the simulation (also see Table 3). First, these three analyses are consistent in indicating that the mean marginal compliance rate for on-product warnings is positive. Second, the mean marginal compliance rate was substantially lower for the descriptive statistics than for the meta-analysis or the simulation (.16 as opposed to .31 and .27), which indicates that some of the experimental conditions with low compliance rates also had low sample sizes and thus should not have received equal treatment with the studies with larger sample sizes. Third, the estimated variance of the marginal compliance rates was greater for the simulation than for the descriptive statistics or the meta-analysis (.125 versus .037 and .055), which reveals that the lack of independence among the 79 experimental conditions did influence the results. Fourth, the lack of independence also resulted in the underestimation of the percentage of experimental conditions with negative compliance for the meta-analysis compared with the simulation and the descriptive statistics (.09 as opposed to .21 and .19).

A second limitation of meta-analysis is that academic journals tend to publish only those studies giving positive results. The literature review discussed previously indicates that this may be less of a problem for warnings research than for other areas, because several studies with negative results were found in masters theses and conference proceedings. Nevertheless, the possibility of this bias was addressed by estimating the number of studies with negative results required to undermine the conclusion that warnings are generally effective in increasing safe behavior (Hedges and Olkin 1985). This number is estimated by

$$k_0 = k(\bar{d} - d_c)/d_c,$$

where k is the number of studies included in the analysis, $\bar{d}$ is their average effect size, and d_c is the effect size small enough to be considered negligible. If $d_c = .1$, $k_0 = 171.96$. Thus, it is unlikely that there is a sufficient number of unpublished studies with negative results to undermine the conclusion that the presence of warnings generally increases safe behavior.

What Factors Contribute to the Variation in Warning Study Results?

The previous analysis employing a random-effects model establishes that effect sizes differ from one experimental condition to another. Our second approach to meta-analysis of these data attempts to explain the differences in the effect sizes using characteristics of the studies themselves. A regression model is fit to the effect sizes using characteristics of the studies as predictors. The model fit is $d_i = \alpha + \beta\chi_i + \varepsilon_i$, where χ_i is a characteristic of experimental condition i, and $\varepsilon_i = d_i - \delta_i$ is the residual.

The fitting of a regression model for meta-analysis requires the use of weighted least squares, in which the weight is the inverse of the variance of the effect size. (Estimates of these weights are obtained using Equation 4 in the Appendix.) From this analysis, we obtain estimates of the coefficients α and β, their standard errors, a test statistic for significance of the model, and a test statistic for testing model specification (Hedges and Olkin 1985, Chapter 8). In general, the test statistic for significance of the model, denoted by Q_R, has an approximate chi-square distribution with p degrees of freedom if the slope and intercept are simultaneously zero, where p is the number of parameters to be estimated in the regression model.

The test statistic for correct model specification, denoted by Q_E, has an approximate chi-square distribution with k – p degrees of freedom when the model is correctly specified, where k is the number of effect sizes. If this hypothesis of correct specification is rejected, we can draw the conclusion that there remain real differences in effect sizes that are not explainable by the model, while acceptance of this hypothesis finds no evidence of remaining differences beyond random variation.

Analysis failed to isolate meaningful relationships between warning compliance and the following study characteristics: (1) the experimental design of the studies, (2) the sample sizes of the studies, (3) the number of independent variables employed in the studies, and (4) the year the studies were conducted. Additionally, a survey was administered to obtain measures of the likelihood of encountering the hazard, the severity of the hazard, perceived risk, and the likelihood of reading and following the warnings for each of the experiments included in this meta-analysis. These measures also were employed in weighted regressions, but no statistically significant relationship with warning compliance was found.

Student Subjects

The possible impact on marginal compliance rates of using student subjects also was addressed in the analysis, because some researchers (e.g., DeJoy 1989) have questioned the validity of their use. A regression analysis was run in which the predictor involved a classification of the studies based on whether or not the subjects were students (i.e., $x_i = 1$ if students were employed as subjects and 0 otherwise). The estimates of the model parameters (and their standard errors) from the weighted regression were $\hat{\alpha} = .236(.059)$ and $\hat{\beta} = .214(.096)$. There is evidence of explanatory power in the model ($Q_R = 5.00$, $p = .03$), and the model is properly specified ($Q_E = 83.19$, $p = .29$).

That this analysis shows that the marginal compliance rates were higher for studies using students as subjects raises the possibility that the positive effect for warnings found previously could be attributable to this artifact. Consequently, the sample was divided between students and nonstudents, and separate meta-analyses using the random-effects model were conducted for the two subsets.

For the nonstudent sample of 43 experimental conditions, $\hat{\Delta} = .24$ and $\hat{\sigma}^2(\hat{\Delta}) = .003$. An approximate 95% confidence interval for $\bar{\Delta}$ is .14 to .35. Because zero lies outside this range, we again can conclude that the mean of the distribution of warning effect sizes exceeds zero. If the distribution of these effects is normal with a mean of .24 and a standard deviation of .17, then 7% of its values would be below zero.

For the student sample of 36 experimental conditions, $\hat{\Delta} = .43$ and $\hat{\sigma}^2(\hat{\Delta}) = .006$. Because $\hat{\Delta}$ is approximately normally distributed, an approximate 95% confidence interval for $\bar{\Delta}$ is .28 to .59. Because zero lies outside this range, we can conclude that the mean of the distribution of warning effect sizes exceeds zero. If the distribution of these effects is normal with a mean of .43 and a standard deviation of .27, then 6% of its values would be below zero.

There are two main limitations of this meta-analysis. First, its results can be generalized only to the population of which the 15 studies are representative. To the extent that products, product users, and usage situations are not included in the sample, the results are limited. For example, none of the experiments employs geriatric subjects. Second, several of the studies were susceptible to demand effects, because subjects knew they were in some sort of study and did not use the products in their natural habitats. Researchers must strive to ensure the realism of their experimental settings, and additional field studies are needed. These limitations notwithstanding, this meta-analysis indicates that (1) the mean compliance rate is higher for students than nonstudents, (2) the variance for students is also higher, and (3) the positive effect of on-product warnings on behavior exists for students and nonstudents alike.

Conclusions and Discussion

The 15 warning experiments examined individually and collectively through meta-analysis lead to several conclusions. First, conformity with warning instructions varies remarkably. The absolute level of compliance in the presence of warnings was found to range from 0% to 100%, and the marginal compliance varied from –22 percentage points to 60 percentage points. This suggests that the phenomena underlying warning effectiveness are dynamic and probably reflect the complex interaction of the (1) warning, (2) product, (3) usage situation, and (4) user (Stewart and Martin 1994).

Second, the descriptive statistics, the meta-analysis, and the simulation indicate that warnings increase safe behavior in general in the studies examined. This means that managers and policymakers should consider the use of on-product warnings to be a potentially effective method of increasing safe behavior. It also indicates that the examination of warnings is a profitable area of research by academics and practitioners. However, eliminating product hazards through design and using some type of guarding mechanism to protect users from hazards remain the preferred methods of protecting product users.

Third, in a small but significant number of instances, the addition of a warning actually reduces safe behavior from the level achieved when no warning is present. Whether this finding applies only to instances in which a warning is poorly designed or extends to instances in which a warning is inappropriate cannot be determined without additional research. In either case, it is imperative that the effectiveness of all warnings be assessed through testing a group of individuals representative of the population of product users; subjective evaluations alone are inadequate.

Fourth, though warning compliance rates tend to be higher among studies employing students as subjects, warnings generally were found to be effective with both students and nonstudents, and the majority of subjects employed in the 15 experiments were not students. Whether this difference in compliance rates is attributable to characteristics of the students themselves or to the conditions in which they usually are studied cannot be determined at this point. In any case, the examination of the behavioral effectiveness of on-product warnings among different population segments is an important area for further research.

Fifth, more studies with no-warning control conditions are needed. Although the typical factorial designs can indicate the relative effectiveness of alternative warning formulations, they are not capable of determining whether any of these formulations are superior to no warning at all. In the future, these studies also should include a no-warning control condition (see Otsubo 1988a), because this meta-analysis makes it clear that there are instances in which the presence of a warning actually reduces safe behavior.

The qualitative reviews cited previously were reexamined upon the completion of the meta-analysis. This reexamination revealed that these reviews differ significantly from the meta-analysis presented here in the experiments included in their analyses. Of the fifteen experiments included in this meta-analysis, McCarthy and colleagues (1984) and Lehto and Miller (1988) each cited one, Stewart and Martin (1994) cited two, and Ayres and colleagues (1992) cited six. This reexamination also revealed that these qualitative reviews shared an attribute with this meta-analysis: All of the studies address the general question of warning effectiveness, but they do not provide a detailed and systematic review of the myriad factors contributing to warning effectiveness. Qualitative and quantitative reviews that examine the nature and impact of the various characteristics of the warning, product, product user, and usage environment are needed badly.

Public Policy Implications

It is paradoxical that though warnings are viewed as important social instruments for protecting consumers, relatively little research evaluates their behavioral effectiveness. There are no published results of any manufacturer's attempts to test its warnings, and government mandated warnings are typically developed through administrative or legislative processes without the aid of empirical testing. An exception is the alcoholic beverage warning legislation, which provided funds for evaluating the effectiveness of the warnings after they were introduced into the marketplace (Alcohol Beverage Labeling Act of 1988).

As was suggested previously, it is imperative that the effectiveness of warnings be tested in each instance in which they are employed. Stewart and Martin (1994) suggest that strategic research and copy testing used commonly in the development of other communication tools could be used in the design of product warnings.

Some might argue that the existence of evidence revealing the empirical testing of alternative warning designs only would serve to aid the prosecution in product liability lawsuits. However, using empirical research to improve warning effectiveness could reduce accidents and avoid these lawsuits. Furthermore, such evidence would demonstrate that the manufacturer is not negligent and has made every effort to design effective warnings. Finally, a manufacturer could defend itself in a lawsuit by claiming that the plaintiff would need to support its claim that the warning is defective by providing empirical evidence that an alternative warning design is superior in effectiveness. The alternative commonly employed today in lawsuits is to have the plaintiff's and defendant's warnings experts offer conflicting subjective opinions regarding a warning's effectiveness (Henderson and Twersky 1990, p. 305).

Additionally, government mandated warnings should be evaluated empirically for their effectiveness. In Wilkie's (1982, 1983) review of the Federal Trade Commission's (FTC) affirmative disclosure cases between 1970 and 1977, 34% involved warnings. In his recommendations for improving disclosures, he indicates that the FTC should specify the objectives of the disclosure in terms of the desired cognitions and behaviors of consumers and enable the respondent in the case to determine the means of achieving those objectives. Wilkie (1982) also states that either pretesting (Recommendation 7) or post-testing (Recommendation 8) should be used to evaluate the effectiveness of the disclosure (i.e., warning).

Basic research is needed to develop a comprehensive theory of warning effectiveness that can direct the efforts of applied researchers charged with developing a warning for a specific product. Although the behavioral effectiveness of product warnings is an issue of great social significance, the funds supporting warnings research are negligible at best. Manufacturers, foundations, and government agencies could make funds available to researchers either directly or through a funding organization such as the Marketing Science Institute. Working with a funding organization would enable manufacturers to learn more about the design of effective warnings even if they were concerned about the liability exposure of funding such research directly.

The availability of such funds would increase greatly the amount of research on product warnings. Perhaps more important, it might provide an opportunity for collaboration between academic researchers and industry and government practitioners, which would result in studies outside the laboratory employing samples representative of the general population.

We present here a meta-analysis of studies of the behavioral effectiveness of on-product warnings and provide an addition to the contributions of qualitative literature reviews published previously. Additional meta-analyses employing experimental conditions as the unit of analysis can enable researchers to evaluate systematically the relative importance of the many factors contributing to warning effectiveness. For individual empirical studies to contribute significantly to the literature and future meta-analyses, it is imperative that the researchers give complete descriptions of their warnings and the settings in which they are being tested. As well, compliance frequencies and sample sizes should be published fully even when the results are not statistically significant, so that compliance rates can be calculated for all main effects and experimental conditions.

Appendix

Meta-Analysis Estimation Procedures

The effect size for each experimental condition was estimated using Hedges's bias-corrected estimator:

$$\text{(1)} \qquad d_i = \left(1 - \frac{3}{4N_i - 9}\right)\left(\frac{p_i^E - p_i^C}{s_i}\right),$$

where n_i^E and n_i^C are the sample sizes, and p_i^E and p_i^C are the observed compliance rates for the experimental groups, $s_i = \sqrt{\left\{\left[n_i^C p_i^C(1 - p_i^C) + n_i^E p_i^E(1 - p_i^E)\right] / \left[n_i^C + n_i^E - 2\right]\right\}}$ and $N_i = n_i^E + n_i^C$. These effect-size estimates range from –.43 to 1.48 for the 79 experimental conditions.

Under the random-effects model, the estimator d_i is afflicted by two sources of variation: one due to the sampling variance of d_i, and the other to the variance of the true effect sizes in the population of warning effects. This partitioning can be described by the identity

$$\text{(2)} \qquad \sigma^2(d_i) = \sigma^2(d_i|\delta_i) + \sigma^2(\Delta).$$

The components of variance on the right-hand side of Equation 2 can be estimated by partitioning into two parts the estimate of total variance in the d_i's:

$$\text{(3)} \qquad s^2(d) = \frac{1}{k-1}\sum_{i=1}^{k}\left(d_i - \bar{d}\right)^2,$$

where $\bar{d}$ is the unweighted mean of the estimated effect sizes, and k is the number of observations (Hedges and Olkin 1985). These component estimators are

$$\text{(4)} \qquad \hat{\sigma}^2(d_i|\delta_i) = \frac{1}{n_i^E} + \frac{1}{n_i^C} + \frac{d_i^2}{2N_i}$$

and

$$\text{(5)} \qquad \hat{\sigma}^2(\Delta) = s^2(d) - \frac{1}{k}\sum_{i=1}^{k}\hat{\sigma}^2(d_i|\delta_i).$$

The estimates of these components of variance for these data are $\Sigma\hat{\sigma}^2(d_i|\delta_i)/k = .14$ and $\hat{\sigma}^2(\Delta) = .055$. A test of the hypothesis that the effect-size variance component is 0 (i.e., that $\delta_1 = \delta_2 = ... = \delta_k = \bar{\Delta}$) is provided by the test statistic

$$Q = \sum_{i=1}^{k}\frac{\left(d_i - d_+\right)^2}{\hat{\sigma}^2\left(d_i|\delta_i\right)},$$

where $d_+ = \Sigma[d_i/\hat{\sigma}^2(d_i - \delta_i)]/\Sigma[1/\hat{\sigma}^2(d_i|\delta_i)]$. If there are no differences among the δ_i's, Q will have an approximate chi-square distribution with k – 1 degrees of freedom. For our data, Q = 210.68, which leads to the rejection of the hypothesis of equal effect sizes, because the 99th percentile point of a chi-square distribution with 78 degrees of freedom is 109.95. This means that a portion (estimated to be .055/(.14 + .055) = 29%) of the variability in the d_i's can be attributed to nonconstant effect sizes across the observations.

The next step is to estimate the mean of the distribution of warning effect sizes $\bar{\Delta}$, which is efficiently estimated by a weighted average of the estimates of the individual observation effect sizes, where the weights are inversely proportional to the estimated variances of the d_i's. That is,

$$\hat{\bar{\Delta}} = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i-1}^{k} w_i},$$

where $w_i = 1/[\hat{\sigma}^2(\Delta) + \hat{\sigma}^2(d_i|\delta_i)]$. An appropriate variance estimate for this estimator is then $\hat{\sigma}^2(\hat{\bar{\Delta}}) = (\Sigma w_i)^{-1}$. Applying this estimation procedure to the data yields the result $\hat{\bar{\Delta}}$ = .311 and $\hat{\sigma}^2(\hat{\bar{\Delta}}) = .002$. Because $\hat{\bar{\Delta}}$ is approximately normally distributed, an approximate 95% confidence interval for $\hat{\bar{\Delta}}$ is .220 to .402. Because zero lies outside this range, it can be concluded that the mean of the distribution of warning effect sizes is nonzero.

Note that this does not mean that there are no zero, or negative, warning effect sizes. Suppose that the warning effect size distribution is normally distributed, which is a reasonable assumption based on the appearance of Figure 1.

A distribution of warning effects with a mean of .311 and a standard deviation of $\sqrt{(.055)} = .220$ therefore would have less than 10% ($\hat{\pi}_{\delta<0} = \Phi\{[0 - \hat{\Delta}]/[\hat{\sigma}(\Delta)]\} = .093$) of its values less than zero.

References

Alcohol Beverage Labeling Act of 1988 (1988), Public Law 100,690, 27 U.S.C. §201–211 (November 18).

Ayres, T. J., M. M. Gross, D. P. Horst, and J. N. Robinson (1992), "A Methodological Taxonomy for Warning Research," in *Proceedings of the Human Factors Society 36th Annual Meeting*. Santa Monica, CA: The Human Factors Society, 499–503.

Barnett, R. L. and W. G. Switalski (1988), "Principles of Safety Behavior," *Safety Brief*, 5 (February), 1–15.

Cook, T. D. and D. T. Campbell (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Cox, E. P., III, W. D. Hoyer, and K. Krshna (1995), "The Behavioral Effectiveness of On-Product Warnings in Low Hazard Conditions," in *Marketing and Public Policy Conference Proceedings*, Pam Scholder Ellen and Patrick J. Kaufman, eds. Atlanta: Georgia State University, 1–10.

DeJoy, D. (1989), "Consumer Product Warnings: Review and Analysis of Effectiveness Research," in *Proceedings of the Human Factors Society 31st Annual Meeting*. Santa Monica, CA: The Human Factors Society, 936–40.

Dingus, T. A., B. P. Hunn, and S. S. Wreggit (1991), "Two Reasons for Providing Protective Equipment as Part of Hazardous Consumer Product Packaging," in *Proceedings of the Human Factors Society 35th Annual Meeting*. Santa Monica, CA: The Human Factors Society, 1039–42.

———, S. S. Wregitt, and J. A. Hathaway (1993), "Warning Variables Affecting Personal Protective Equipment Use," *Safety Science*, 16, 655–73.

Dorris, A. L. and J. L. Purswell (1977), "Warnings and Human Behavior: Implications for the Design of Product Warnings," *Journal of Products Liability*, 1, 255–64.

Duffy, R. R. (1993), "The Effectiveness of an Interactive Warning Label: An Examination of Warning Effectiveness, Task Load and Hazard Perception," unpublished master's thesis, Rensselaer Polytechnic Institute.

———, M. J. Kalsher, and M. S. Wogalter (1993), "The Effectiveness of an Interactive Warning in a Realistic Product-Use Situation," in *Proceedings of the Human Factors Society 37th Annual Meeting*. Santa Monica, CA: The Human Factors Society, 935–39.

Edworthy, J., N. Stanton, and E. Hellier (1995), "Editorial: Warnings in Research and Practice," *Ergonomics*, 38 (November), 2145–54.

Farley, J. U. and D. R. Lehmann (1986), *Meta-Analysis in Marketing: Generalization of Response Models*. Lexington, MA: Lexington Books.

Frantz, J. P. and T. P. Rhoades (1993), "A Task-Analytic Approach to the Temporal and Spatial Placement of Product Warnings," *Human Factors*, 35 (4), 719–30.

Friedmann, K. (1988), "The Effects of Adding Symbols to Written Labels on User Behavior and Recall," *Human Factors*, 30 (August), 507–15.

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: The Human Factors Society, 712–15.

Haddon, S. G. (1986), *Read the Label: Reducing the Risk by Providing Information*. Boulder, CO: Westview Press.

Hardie, W. H. (1994), "Critical Analysis of On-Product Warning Theory," *Product Safety & Liability Reporter*, 22 (February 2), 145–63.

Hatem, A. T. (1993), "The Effect of Performance Level on Warning Compliance," unpublished master's thesis, Purdue University.

——— and L. Lehto (1995), "Effectiveness of Glue Odour as a Warning Signal," *Ergonomics*, 38 (November), 2250–61.

Hedges, L. V. and I. Olkin (1985), *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Henderson, J. A., Jr. and A. D. Twersky (1990), "Doctrinal Collapse in Products Liability: The Empty Shell of Failure to Warn," *New York University Law Review*, 6 (May), 265–327.

Horst, D. P., G. E. McCarthy, J. N. Robinson, R. L. McCarthy, and S. Krumm-Scott (1986), "Safety Information Presentation: Factors Influencing the Potential for Changing Behavior," in *Proceedings of the Human Factors Society 30th Annual Meeting*. Santa Monica CA: The Human Factors Society, 111–15.

Hunn, B. P. and T. A. Dingus (1992), "Interactivity, Information, and Compliance Cost in a Consumer Product Warning Scenario," *Accident Analysis and Prevention*, 24 (5), 497–505.

Hunter, J. E. and F. L. Schmidt (1990), *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage Publications.

Lehto, M. R. and J. P. Foley (1989), "The Influence of Regulation, Training and Product Information on Use of Helmets by ATV Operators: A Field Study," in *Proceedings of INTERFACE '89—The Sixth Symposium on Human Factors and Industrial Design in Consumer Products*. Santa Monica, CA: The Human Factors Society, 107–13.

——— and ——— (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 (Winter), 191–200.

——— and J. M. Miller (1988), "The Effectiveness of Warning Labels," *Journal of Products Liability*, 11 (3), 225–65.

McCarthy, G. E., D. B. Horst, R. R. Beyer, J. N. Robinson, and R. L. McCarthy (1987), "Measured Impact of a Mandated Warning on User Behavior," in *Proceedings of the Human Factors Society 31st Annual Meeting*. Santa Monica, CA: The Human Factors Society, 479–83.

McCarthy, R. L., T. J. Ayres, and C. T. Wood (1995), "Risk and Effectiveness Criteria for Using On-Product Warnings," *Ergonomics*, 38 (November), 2164–75.

———, J. P. Finnegan, S. Scott-Crum, and G. E. McCarthy (1984), "Product Information Presentation, User Behavior and Safety," in *Proceedings of the Human Factors Society 28th Annual Meeting*. Santa Monica, CA: The Human Factors Society, 81–85.

Miller, J. M., M. H. Lehto, and J. P. Frantz (1994), *Instructions and Warnings: The Annotated Bibliography*. Ann Arbor, MI: Fuller Technical Publications.

Otsubo, S. M. (1988a), "Effectiveness of Warning Signs on Consumer Products," unpublished master's thesis, California State University, Northridge.

——— (1988b), "A Behavioral Study of Warning Labels for Consumer Products: Perceived Danger and Use of Pictographs," in

Proceedings of the Human Factor Society 32nd Annual Meeting. Santa Monica, CA: The Human Factors Society, 536–40.

Schneider, K. C. (1977a), "An Experimental Evaluation of Relationships Between Product Package and Label Characteristics and Preschool Childrens' Awareness of and Behavior Toward Package Contents: A Study in Prevention of Childhood Poisoning," unpublished doctoral dissertation, University of Minnesota.

——— (1977b), "Prevention of Accidental Poisoning Through Package and Label Design," *Journal of Consumer Research*, 4 (September), 67–74.

Strawbridge, J. A. (1986a), "The Influence of Position, Highlighting and Imbedding on Warning Effectiveness," unpublished master's thesis, California State University, Northridge.

——— (1986b), "The Influence of Position, Highlighting, and Imbedding on Warning Effectiveness," in *Proceedings of the Human Factors Society 30th Annual Meeting*. Santa Monica, CA: The Human Factors Society, 716–20.

Stewart, D. W. and I. M. Martin (1994), "Intended and Unintended Consequences of Warning Messages: A Review and Synthesis of Empirical Research," *Journal of Public Policy & Marketing*, 13 (Spring), 1–19.

Thyer, B. A. and E. S. Geller (1987), "The 'Buckle-Up' Dashboard Sticker: An Effective Environmental Intervention for Safety Belt Promotion," *Environment and Behavior*, 19 (July), 484–94.

Venema, A. (1989), "Product Information for the Prevention of Accidents in the Home and During Leisure Activities," Research Report No. 69, Institute for Consumer Research, SWOKA, The Netherlands.

Wilkie, W. L. (1982), "Affirmative Disclosure: Perspectives on FTC Orders," *Journal of Marketing and Public Policy*, 1, 95–110.

——— (1983), "Affirmative Disclosure at the FTC: Theoretical Framework and Typology of Case Selection," *Journal of Marketing and Public Policy*, 2, 3–15.

Wogalter, M. S., T. Barlow, and S. A. Murphy (1994), "Compliance to Owner's Manual Warnings: The Influence of Familiarity and Placement of a Supplemental Directive," *Ergonomics*, 38 (June), 1081–91.

——— and M. J. Kalsher (1994), "Increasing the Correct Connection of Car Battery Jumper Cables with an Enhanced Tag Warning," in *Proceedings of Public Graphics*. Lunteren, The Netherlands, (September 26–30).

Wolf, F. M. (1986), *Meta-Analysis: Quantitative Methods for Research Synthesis*. Beverly Hills, CA: Sage Publications.

Wreggit, S. S. (1991), "The Effects of Cost of Compliance, Information, and Physical Activity on Consumer Product Warning Compliance," unpublished master's thesis, University of Idaho.