Comprehension of Pictorial Symbols: Effects of Context and Test Method

Jennifer Snow Wolff, Georgia Institute of Technology, Atlanta, Georgia, and **Michael S. Wogalter**, North Carolina State University, Raleigh, North Carolina

This research examined two factors involved in the evaluation of pictorial symbol comprehension: context (absence vs. presence of photographs depicting the probable environments where a symbol would be seen) and test method (multiple-choice with less vs. more plausible distractor alternatives vs. open-ended). We tested 33 pictorial symbols from various sources. The results showed that the multiple-choice test with less plausible distractors inflated comprehension scores by an average of 30% compared with the other two tests, which did not differ. The presence of context increased symbol comprehension in the open-ended test and in the multiple-choice test that had more plausible distractors. Extensive preliminary procedures demonstrated the difficulty of forming a multiple-choice test with plausible distractor alternatives. This fact, combined with multiple-choice tests' low ecological validity in reflecting the real-world task of symbol comprehension, suggests that this test should be avoided in favor of an open-ended testing procedure. It is suggested that context provides ecologically valid cues that limit the range of possible constructs that the pictorial symbol could be, raising comprehension scores. The use of context may help reduce the costs (money, time, effort) of producing pictorial symbols with acceptable, above-criterion comprehension levels.

INTRODUCTION

With increasing attention to multiculturalism and worldwide trade, pictorial symbols are increasingly being used to convey important messages. Well-designed symbols have the ability to communicate large amounts of information at a glance. They can also be useful in conveying information to persons who cannot read a printed verbal message, either because they have vision problems (e.g., older adults), lower-level verbal skills, or inadequate knowledge of the language being used in the warning (Boersema & Zwaga, 1989; Collins, 1983; Laux, Mayer, & Thompson, 1989; Lerner & Collins, 1980; Zwaga & Easterby, 1984). Clear communication is particularly critical when a pictorial symbol conveys safety-related information, as the lack of understanding or misinterpretation could lead to injury.

Given pictorial symbols' potentially important role in communicating hazards, national and international standards have been established that describe how to evaluate their comprehensibility, such as the American National Standard Institute's ANSI Z535.3 (ANSI, 1991) and the Organization for International Standardization's ISO 3864 (ISO, 1984). ANSI and ISO advise that symbols must reach a criterion of at least 85% or 67% correct, respectively, in a comprehension test to be considered acceptable. Despite the existence of these standards, pictorial symbols are routinely placed on signs, labels, and other materials without any evaluation as to whether they communicate the intended concepts to

Requests for reprints should be sent to Michael S. Wogalter, Psychology Department, 640 Poe Hall, CB 7801, North Carolina State University, Raleigh, NC 27695-7801. **HUMAN FACTORS**, Vol. 40, No. 2, June 1998, pp. 173–186.

the targeted audience. Moreover, very little research has evaluated the methods of testing comprehension. The present research examines some of the factors that might influence the results of symbol comprehension tests.

Study Goals

The present study had three major objectives. The first was to compare comprehension performance using two test methods commonly employed to assess symbol comprehensibility: the multiple-choice test and the open-ended test. In multiple-choice tests, respondents are asked to choose the answer that best expresses the symbol's meaning from several alternative answers. In open-ended tests, participants are shown a symbol and are asked to give its meaning in their own words. In the current version of the ANSI (1991) Z535.3 symbol standard, either kind of test is allowed, although preference is given to the open-ended test method.

Dewar (1994) and others have expressed concerns about multiple-choice tests by pointing out that the quality of the distractor alternatives (wrong answers) could greatly influence comprehension scores. An obvious example is a pedestrian crossing symbol in a test that includes distractors such as "keep refrigerated" or "no left turn." These distractor answers could be easily ruled out by respondents, enabling them to choose the correct answer and unfairly inflating the symbol's purported comprehension level compared with a test with more plausible distractors.

Although the lack of plausibility can be obvious (as in the previous example), distractor quality in actual multiple-choice tests can be subtle and difficult to spot and may influence test performance (Katz & Lautenschlager, 1994; Wogalter & Marwitz, 1987; Wogalter, Marwitz, & Leonard, 1992).

Thus because of the importance of having plausible distractor alternatives, multiple-choice tests might require considerable development work in the test construction stage. Open-ended tests are easier to develop because they require only the set of symbols to be placed on test sheets with blank spaces. Nevertheless, multiple-choice tests are generally much easier to score than open-ended tests. Indeed, because of its straightforward quantification (the simple counting of responses), multiple-choice tests might appear more scientific. The scoring of open-ended responses is more difficult and less clear-cut. One must establish criteria for the kinds of answers that will be counted as correct; there is usually (if not always) at least some subjective judgment of the correctness of participants' responses. There should also be an assessment of reliability, requiring the responses to be scored by more than one judge. Dewar (1994) pointed out that the extra effort is worthwhile in terms of information gained about the types of errors and confusions people make and might assist in any subsequent redesign work if comprehension scores fall below some acceptable criterion level.

Another important issue concerns the ecological validity of the tests. The cognitive operations involved in taking a multiple-choice test do not reflect the processing operations that people ordinarily perform when encountering pictorial symbols in the real world. Generally, people do not select from a set of alternative answers; rather, they generate meaning from a symbol (and any associated words) in ways that more closely mirror the cognitive processes involved in the open-ended test. Neisser (1987) suggested that free-recall-type evaluations (such as open-ended tests) are less likely to produce constrained and distorted participant reports compared with cued-recall and recognition tests (such as multiple-choice tests).

Therefore, a second goal of this research was to examine the influence of distractor plausibility in multiple-choice tests. We compared two tests, both having the same correct referent answer, with three distractors that were either low or high in plausibility (less vs. more plausible). Also described is the extensive preliminary work that went into developing the distractor sets used in these tests – precursor aspects that have generally not been articulated in previous pictorial symbol research.

In real-world environments, symbols generally exist in contexts that are likely to assist in the comprehension of their meaning. However, most symbol-testing protocols, which are often done in laboratory and classroom settings, evaluate comprehension with little or no information about the context in which the symbol might actually appear (Dewar, 1994). Without contextual cues, low test comprehension scores would falsely indicate that additional, often costly, design and test work is necessary. If the participants had known where the symbol would be located, there might have been much better comprehension. Also, without apparent context, participants might supply their own implicit context (mental set), which might or might not reflect the real-world context in which the symbol would appear.

For example, a symbol depicting a boot might produce two or more interpretations depending on the context inferred (e.g., that safety shoes must be worn or that a shoe store or repair shop is present). As a consequence of priming from other thought processes, one concept might come to mind more easily than another. If the individual was recently thinking about industrial work, the symbol might be more likely to be interpreted as a directive to wear safety shoes. If the individual had been thinking of old shoes, it might be interpreted as indicating a shoe store or repair shop. However, had a context been provided (e.g., a photograph of a construction site vs. a marketplace), it is likely that the number of incorrect responses would be substantially reduced. Thus the practical issue for symbol comprehension testing is that without context, the results might not reflect real-world understandability and the symbol might fail to reach ANSI's (1991) or ISO's (1984) criterion level.

The third goal of the present research was to examine the effects of context on comprehension scores. Context was defined as information relating to the probable environments in which the symbol would appear. Context was manipulated as the absence versus presence of location-appropriate photographs concurrent with the presentation of the symbols during testing.

Numerous studies in the basic cognitive literature suggest that context would be beneficial for symbol comprehension. For example, participants identified words better under degraded conditions when they were previously given part of a sentence that provided related information (Tulving, Mandler, & Baumal, 1964). Biederman, Glass, and Stacy (1973) showed that participants were more likely to correctly identify objects in a coherent scene than the same objects in a jumbled scene. Also, Palmer (1975) showed that parts of faces (e.g., an ear) were more likely to be identified if they were shown in the context of a complete face than if shown separately.

However, with regard to pictorial symbol comprehension, only a few studies have examined the effect of context. Vukelich and Whitaker (1993) investigated the influence of verbal context by supplying participants with a more elaborate description, partial description, or no description. Comprehension was highest with the more elaborate verbal description. Cahill (1975) tested 10 graphic symbols in context (a drawing plus verbal instructions) or in isolation. Symbols were more frequently identified when they appeared in context compared with no context. However, two other studies showed no effect of context. Wogalter, Sojourner, and Brelsford (1997) were unable to demonstrate an effect of explicit verbal consequence information on symbol comprehension. Silver et al. (1995) found no comprehension enhancement for symbols accompanied by a photograph and a verbal description of an environmental scene compared with the same symbols without this information.

Thus across these studies, the effect of context in symbol comprehension tests is equivocal, and the reasons why some studies show a benefit and others do not are not entirely clear. Possibilities include the particular set of stimuli tested and the kind of context provided. Some studies used complex symbols containing many details in the symbol itself. This visual detail might provide enough information about where the symbol would be placed that additional contextual information might have minimal value. Also, all of the previous studies used a verbal description to provide all or part of the context, and they provided different amounts of verbal cues. In the present research detailed color photographs were used to depict the environment in which the symbol would be placed. Although the use of photographs has been mentioned as a good method to provide context (Dewar, 1994), no other study to date has used nonverbal, photographic context exclusively. Given that one purpose of pictorial symbols is to convey information to people who have less proficient language skills, a purely pictorial context would seem appropriate.

Finally, in this study we also examined the effect of context on errors. If context aids in limiting what the symbol might be, then it should also limit the range of responses and the number of confusions produced.

The present research consisted of two major stages. The first involved preparing materials for the main experiment, including symbol selection, development and selection of the multiple-choice distractor alternatives, evaluation of distractor plausibility, and a preliminary comprehension test. This rarely described stage is critical in the construction of multiple-choice tests, the purpose of which are to assess an absolute criterion of comprehension. Through its more explicit treatment here, it will be seen that the development of a set of adequately plausible distractors might not be achieved even with extensive preliminary work. The second stage, described later, is the main experiment, which examined symbol comprehension performance as a function of test method and context.

PRELIMINARY STAGES: SELECTING, DEVELOPING, AND TESTING MATERIALS

Selection of Symbols and Distractors

We used 33 pictorial symbols (7 pharmaceutical, 21 industrial safety, and 5 from various categories). Most had been tested in previous research. Various kinds were included to foster generalizability to a broad range of safety-related symbols. The list of the symbols, referents, and sources is given in Figure 1 and Table 1.

Preliminary Selection of Distractor Alternatives

The multiple-choice tests used in the main experiment (to be described later) consisted of items with four alternatives: one correct choice and three distractors. Thus to examine the effect of plausibility, six distractors (three of high plausibility and three of low plausibility) were needed.

The pharmaceutical pictorial symbols were tested in earlier research using open-ended tests (Magurno, Wogalter, Kohake, & Wolff, 1994; Wolff & Wogalter, 1993). These data provided a large set of incorrect answers that might be used as distractor alternatives. The industrial safety symbols were tested by Collins (1983) in a mine hazard study using a multiple-choice test, with each item having three distractor alternatives along with the correct answer. For these symbols at least three more distractors were needed. The third group of five other symbols were obtained from various published articles: (a) "No Boating, Undertow"; (b) "No Fishing, Rising Water"; (c) "Shallow Water, No Diving, You Can Be Paralyzed"; (d) "Remain clear of lift when raising or lowering"; and (e) "Do Not Dig." The first two were created and tested by Dewar and Arthur (1994). The others (Eberhard & Green, 1989; Goldhaber & DeTurck, 1988; Grism, 1993; Silver et al., 1995) had unknown amounts of prior testing, and there were no materials to assist in forming the distractors.

A preliminary set of distractor items was assembled from the materials described earlier (Collins, 1983; Magurno et al., 1994; Wolff & Wogalter, 1993). In cases in which there were no distractors or an insufficient number of existing distractors, a set of potential answers was written to fit the more versus less plausible categories. The more plausible distractors were operationally defined as incorrect answers that potentially reflected what the symbol could mean. Less plausible distractors were defined as incorrect answers that more remotely reflected what the symbol could mean but were not completely implausible, and frequently named resemblances to visual forms in the symbol.

Participants

A total of 225 individuals participated across the various parts of the preliminary phases. They ranged in age from 17 to 56 with an average age of 24.0 (*SD* = 6.5). These volunteers were approached at various locales in the Atlanta, Georgia, area (described later) and were offered snack food and sodas for their participation.

Preliminary Plausibility Ratings

An initial set of 274 candidate distractors and 33 correct answers (i.e., the referent names) were evaluated on the dimension of plausibility. Each verbal item was printed on a separate page below its associated symbol. The pages were randomly assigned to different test booklets with the constraint that no symbol appeared twice in a booklet. The order of pages was randomized.

The initial plausibility test was conducted at the Georgia Tech Student Center with 75 student, staff, and faculty volunteers. The plausibility ratings were made based on the question, "How well does the verbal description match the pictorial?" on a 7-point Likert-type scale with the following numerical and verbal anchors: 1 = *does not fit at all*; 2 = *fits very poorly*; 3 = *fits poorly*; 4 = *fits somewhat*; 5 = *fits well*, 6 = *fits very well*; and 7 = *fits perfectly*. Each verbal item was rated from 8 to 14 times by different evaluators. Only 5 of the 33 symbols yielded a complete set of three more-plausible distractors – distractors with means above the middle of the rating scale (i.e., 4.0).

Preliminary Comprehension Test

To obtain additional plausible distractors, an open-ended comprehension test was conducted with a new group of 100 participants attending two advertised, nonuniversity community events (a dance and a circus). Test booklets each containing approximately one-half of the 33 symbols (16 or 17) were assembled. Symbols were randomly assigned to and randomly ordered within the booklets. Participants wrote their interpretations of the symbols on numbered response sheets. From these responses 78 additional candidate distractors were obtained.

Second Set of Plausibility Ratings

A second plausibility rating test was conducted on the new distractors using 50 participants who approached a table set up near the entrance of an urban public park. These data plus those obtained from earlier plausibility ratings provided the remaining distractors used in the main experiment. Although we tried to compile a complete set of distractors for the experiment, a fully adequate set of more plausible distractors was not achieved. Of the symbols, 7 had three distractors with mean plausibility ratings over 4.0, 10 had only two distractors over 4.0, 10 had only one distractor over 4.0, and 6 had none.

In selecting the final distractor set, an attempt was made to avoid multiple answers with overlapping concepts. The mean fit rating of the entire set of more plausible distractors was 4.04 (SD = 0.67), whereas the mean of all of the less plausible distractors was 2.09 (SD = 0.43). The mean of the correct answers was 4.91 (SD = 0.08). A one-way repeated-measures analysis of variance (ANOVA), $F(2, 64) = 158.03$, $MSE = 0.435$, $p < .0001$, followed by Tukey's Honestly Significant Difference (HSD) test, showed that all differences between these means were significant.

MAIN EXPERIMENT

Method

Design. The experiment was a 2 (Context: absent vs. present) × 3 (Test Method: less vs. more plausible multiple-choice distractors vs. open-ended) factorial design.

Materials. Multiple-choice tests were constructed based on the criteria described earlier. The order of the answers (the correct answer and the three distractor items) was randomized for each question. The location of the correct answer in the set of alternatives for each question was held constant between the two plausibility conditions.

For the two multiple-choice tests, booklets were constructed, with each page having a different symbol and four numbered alternative answers. For the open-ended test the materials were identical, except the multiple-choice alternative answers were replaced with blank spaces.

Context Materials

When context was present, color photographs corresponding to each symbol were included in the test. The photographs were obtained from various tool catalogs and magazines (architectural, automotive, boating, news, scientific, and sports). Additional color prints were obtained by photographing various sites on the Georgia Tech campus (e.g., an eyewash and first aid station, chemistry apparatus, exit doors, machine tools, cylinders,

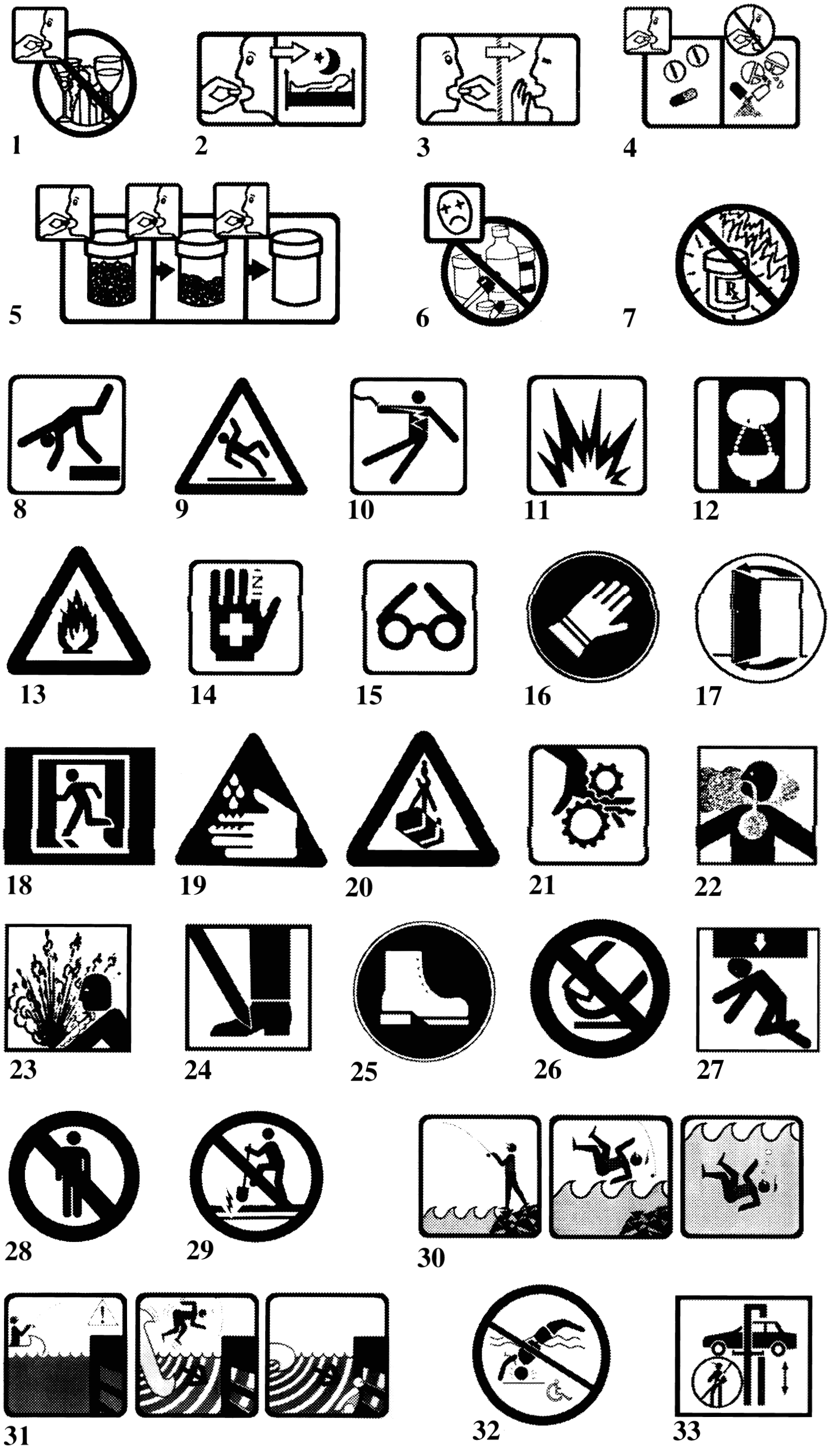

Figure 1. The pictorial symbols. Permission to reprint is gratefully acknowledged to the following sources: for images 1–7 (in Wolff & Wogalter, 1993), U.S. Pharmacopeial Convention, Inc., abstracted from *USP DI®,* copyright 1993; for images 8–28 (in Collins, 1983), National Institute of Standards and Technology; for image 29 (in Silver et al., 1995), Electromark; for images 30 and 31 (in Dewar & Arthur, 1994), the authors; for image 32 (in Goldhaber & deTurck, 1988), the authors; and for image 33 (in Eberhard & Green, 1989), the Automotive Lift Institute.

TABLE 1: The Referents, Scores, and Source List

			MC/ More Plaus/ No Context	MC/ Less Plaus/ No Context	Open Ended/ No Context	MC/ More Plaus/ Context	MC/ Less Plaus/ Context	Open Ended/ Context
Pharm	1	Do not drink alcohol when taking this medication[a]	75%	100%	65%	93%	100%	68%
Pharm	2	Take this medicine at bedtime[b]	57%	100%	38%	57%	96%	45%
Pharm	3	This medicine may make you drowsy[b]	93%	100%	77%	100%	93%	84%
Pharm	4	Do not break or crush tablets or open capsules[b]	47%	93%	39%	50%	89%	50%
Pharm	5	Take until gone[b]	86%	96%	68%	79%	96%	63%
Pharm	6	Do not take other medicines with this medicine[c]	18%	52%	4%	29%	61%	20%
Pharm	7	Do not store near heat or sunlight[a]	54%	86%	27%	50%	82%	28%
Industr	8	Fall from Elevation Hazard[d]	43%	86%	30%	93%	100%	73%
Industr	9	Slip Hazard[d]	100%	100%	91%	96%	96%	94%
Industr	10	Electrical Hazard Present[d]	61%	96%	77%	86%	100%	86%
Industr	11	Explosion Hazard[d]	82%	93%	83%	86%	82%	81%
Industr	12	Eyewash Location[d]	75%	83%	34%	75%	86%	59%
Industr	13	Flammable Hazard[d]	46%	45%	83%	71%	100%	88%
Industr	14	First Aid Location[d]	86%	100%	80%	89%	100%	84%
Industr	15	Eye Protection Required[d]	57%	90%	77%	100%	96%	97%
Industr	16	Hand Protection Required[d]	93%	100%	74%	100%	100%	98%
Industr	17	Keep Door Open[d]	11%	100%	3%	21%	86%	3%
Industr	18	Exit[d]	43%	97%	43%	46%	93%	44%
Industr	19	Corrosive Hazard[d]	82%	90%	60%	93%	82%	81%
Industr	20	Overhead Hazard[d]	39%	90%	74%	54%	89%	82%
Industr	21	Entanglement Hazard[d]	86%	90%	86%	93%	93%	88%
Industr	22	Poison Hazard[d]	36%	48%	59%	75%	96%	78%
Industr	23	Sudden Pressure Release Hazard[d]	36%	76%	3%	71%	96%	25%
Industr	24	Sever Hazard[d]	36%	90%	37%	54%	93%	64%
Industr	25	Foot Protection Required[d]	75%	90%	77%	100%	100%	96%
Industr	26	Do Not Touch[d]	79%	90%	67%	79%	93%	78%
Industr	27	Crush Hazard[d]	50%	86%	23%	61%	86%	16%
Industr	28	Do Not Enter[d]	54%	100%	52%	68%	100%	75%
Misc	29	Do Not Dig[e]	29%	90%	82%	25%	96%	34%
Misc	30	No Fishing, Rising Water[f]	50%	100%	47%	29%	96%	35%
Misc	31	No Boating, Undertow[f]	50%	90%	20%	21%	86%	21%
Misc	32	Shallow Water, No Diving, You Can be Paralyzed[g]	79%	83%	67%	75%	96%	79%
Misc	33	Remain clear of lift when raising or lowering[h]	68%	79%	71%	86%	100%	84%

[a] United States Pharmacopeial Convention (USPC; 1993). [b] Wolff & Wogalter (1993). [c] Magurno et al. (1994). [d] Collins (1983). [e] Silver et al. (1995). [f] Dewar & Arthur (1994). [g] Goldhaber & DeTurck (1988). [h] Eberhard & Green, 1989.

pipes, and electrical boxes). One to four photographs were selected to represent a cross section of environments where a given symbol might be placed. When context was absent, no photographs accompanied the symbols.

The symbols were approximately 5.1 × 5.1 cm (2 × 2 inches). Each symbol, a corresponding number label (to match the answer sheet), and in the context-present condition, photographic material, were assembled onto 21.6 × 28.0-cm (8.5 × 11-inch) paper sheets, color copied, and inserted into plastic protectors. Six sets of these materials were reproduced to allow simultaneous participation by small groups of volunteers.

Procedure. Instructions were given in both oral and printed form. Participants were told either to select one answer from the four alternatives (multiple-choice tests) or to write in their own words (open-ended tests) the meaning conveyed by each symbol. A "No Smoking" symbol was used to illustrate the task. Participants were then told to proceed through their booklets in the order given and not to preview later pages or change earlier answers. They were told, when making their responses, to find the number on the response sheet that corresponded to the current symbol and to mark their answer at that point. Participants were not given any time restrictions.

Before testing, participants also completed a demographics questionnaire requesting gender, age, occupation, native language, racial/ethnic group, and educational level.

Participants. There were 211 participants (52% male; mean age = 34.5, *SD* = 21.1). They were solicited at various locales in the Atlanta, Georgia area including the Georgia Tech Student Center, a church gathering, and a senior citizens center. Of the participants, 91% reported that they were native English speakers and 55% classified themselves as students. In addition, 76% were white, 11% black, and the remaining classified themselves into other racial and ethnic groups; 84% reported having taken at least some college-level classes, and 45% reported attaining a four-year college degree. Participants were assigned randomly to conditions in approximately equal proportions from each locale.

Approximately 100 participants were assigned to the two main factors: (a) 113 in the multiple-choice test and 98 in the open-ended test, and (b) 104 in the context-absent condition and 107 in the context-present condition. Specifically, three of the four Multiple-Choice × Context conditions had 28 participants. The context-absent, less plausible distractor condition had 29 participants. For the open-ended test, there were 47 participants in the context-absent condition and 51 in the context-present condition.

Scoring of open-ended responses. The open-ended answers were scored by two independent judges. For every symbol the judges were provided with the correct answer and all of the participants' written responses from both the context and no-context conditions in a randomized order. The order was reversed for the second judge. The judges did not see the actual symbols before or during the scoring procedure in order to avoid bias by their personal interpretation of the images.

To assist the judges, a score sheet was prepared that had the referent name (correct answer) on top, with the six (more plausible and less plausible) distractors below it, followed by additional spaces for blank (missing), "I don't know," and "other" answers. The judges were asked to match the participants' responses to one of these alternatives on the score sheet and to itemize nonlisted answers in the "other" category. The purpose of this procedure was not only to assess the correctness of the responses but also to determine how frequently participants in the open-ended test mentioned the alternatives in the multiple-choice tests and to assess the range or variety of answers that was produced. Interrater reliability, which was determined by summing the number of agreements between judges divided by the total × 100, was 87.5%.

RESULTS

Comprehension

Correct answers were assigned a score of "1" and incorrect answers a score of "0." In cases in which the judges disagreed in the scoring of an open-ended test response, an average of the two scores was assigned (i.e., 0.5). The raw data were then collapsed across participants to form proportion mean scores, for which each of the 33 symbols had six scores corresponding to the experimental conditions. Table 1 shows these scores converted to percentages. Table 2 shows the proportion-correct means as a function of test method and context collapsed across the symbols.

A 3 × 2 repeated-measures ANOVA using symbols as the random variable showed a significant main effect of test method, $F(2, 64) = 41.81$, $MSE = 0.045$, $p < .0001$. Comparisons among these means using Tukey's HSD test showed that the multiple-choice test with less plausible distractors ($M = .90$) produced significantly higher comprehension scores ($p < .05$) than the tests using either the more plausible multiple-choice distractors ($M = .64$) or the open-ended items ($M = .59$). The latter two means did not significantly differ.

The ANOVA also showed a significant main effect of context, $F(1, 32) = 14.95$, $MSE = 0.025$, $p < .001$. When context was present ($M = .75$), comprehension was significantly greater than when context was absent ($M = .67$).

The interaction between context and test method was not significant at the conventional level of significance, $F(2, 64) = 2.71$, $MSE = 0.010$, $p < .07$. The means in Table 2 show that context had a positive effect across all test methods, but its effect was smaller for the less plausible multiple-choice test. Simple effects analyses corroborated this pattern: Context significantly improved comprehension performance for the multiple-choice test with the more plausible distractors and the open-ended test, but not for the multiple-choice test with less plausible distractors, which already had relatively high scores in the absence of context.

Critical Confusions

ANSI (1991) Z535.3 recommends that acceptable symbols have no more than 5% critical confusions. Critical confusions are a type of wrong answer that is opposite to the answer intended or suggests a behavior that could lead to an accident or injury. Because of the nature of the distractor selection procedure for the multiple-choice tests, alternatives indicating a critical confusion were not always present. However, in the open-ended test all 33 symbols could produce critical confusions.

Of the symbols, 13 had critical confusion levels above 5%. The symbols with the highest critical confusion rates were "Do Not take other medicines with this medicine," "Do Not Dig," "Exit," "Fall from Elevation Hazard," "No Fishing, Rising Water," "Take until gone," and "Do not store near heat or sunlight." The overall critical confusion rate across all symbols was .06 and .07 for the context-absent and context-present conditions, respectively, $p >. 05$.

Range of Responses

To determine whether the range of responses differed as a function of test and context, the number of unique or distinct answers was counted. The first two columns of Table 3 show the mean number of alternatives selected per symbol (from a total of four possible, including the correct answer) for the two multiple-choice tests. The open-ended tests were scored in two ways. The third column gives the mean number of distinct alternatives per symbol when only the correct answer and six possible multiple-choice answers are counted (from a total of seven possible choices), whereas the fourth column gives the mean number of distinct alternatives for which any

TABLE 2: Mean Proportion Correct as a Function of External Context and Test Method

	Test Method			
	Multiple Choice			
External Context	Less Plausible	More Plausible	Open Ended	Mean
Absent	.88	.57	.55	.67
Present	.93	.70	.64	.75
Mean	.90	.64	.59	

TABLE 3: Mean Range of Answers for the Multiple-Choice and Open-Ended Tests

	Test Method			
	Multiple Choice (out of 4)		Open Ended	
External Context	Less Plausible	More Plausible	Out of 7	All Counted
Absent	2.36	3.30	4.33	8.53
Present	2.09	2.94	3.84	7.62

unique answer was counted (no limit of possible answers).

Table 3 shows that the multiple-choice test with less plausible distractors had the most restricted (least varied) response range. As would be expected, the open-ended test produced the widest range of responses. It is also apparent that context reduced the range of responses in all three tests. However, a significant reduction in the range of responses attributable to the presence of context occurred only for the multiple-choice test with more plausible distractors, $t(32) = 2.10, p < .05$.

DISCUSSION

People's comprehension of safety symbols can be critical for avoiding accidents and injuries. Thus valid methods of testing symbol comprehensibility are crucial (Brugger, 1994). The present research examined several factors involved in the measurement of symbol comprehension.

Two multiple-choice comprehension tests were compared: one with less plausible distractors and the other with more plausible distractors. Across all 33 symbols tested, selection of the correct answer from among the less plausible distractors was 30% higher than with the more plausible distractors. We suggest that the comprehension scores provided by the less plausible distractor test give an inflated measure of the symbols' actual comprehensibility.

Although test developers might not intend to include low-plausibility distractors, the sometimes subtle nature of low-plausibility distractors might not be noticed. Collins (1983), one of the pioneers of symbol comprehension testing, obtained high comprehension scores for the "Exit," "Severe Hazard," and "Keep Door Open" symbols (89.2% across all three symbols). In the present research these symbols also received high comprehension scores with the same distractors (93.1%). However, in the present study's preliminary phase evaluations, the plausibility of Collins's distractors was found to be low ($M = 1.7$ on a scale of 7). Here the ratings for the newly derived, more plausible multiple-choice distractors for these three symbols were higher ($M = 3.9$). The presence of these more plausible distractors produced much lower comprehension performance (35%).

Despite extensive pretesting procedures, an adequate number of good distractors still could not be found for some of the symbols. For example, only one plausible distractor (beyond a rating of 4.0) was found for the symbols representing "Fall from Elevation Hazard," "Crush Hazard," "Sever Hazard," and "Sudden Pressure Release Hazard." With the resulting set of distractors, these symbols produced relatively high comprehension levels. Thus the present results suggest not only that distractor plausibility can substantially affect comprehension performance but also that the identification of plausible distractors is not a trivial task and, even with extensive work, might not be fully successful.

Performance on the two multiple-choice tests was compared with that on the open-ended test. Overall mean comprehension on the multiple-choice test with more plausible distractors (59%) was slightly above, but not significantly different from, the open-ended test (55%). Although these two tests provide nearly equivalent performance levels, interpretation requires some caution. They are different kinds of tests, and direct comparison between them cannot be done without an assumption of adequate similarity. This caution notwithstanding, the open-ended test is generally considered the "gold-standard" measure of symbol comprehension (Dewar, 1994), and it has been routinely used as a benchmark to compare with other symbol test methods (Brugger, 1994).

Five other concerns about the use of multiple-choice tests bear mentioning. The first concerns instances in which distractors may be carried over from an earlier symbol test. The consequence of this could be the inclusion of less plausible distractors. For example, Collins (1983) tested the "Slip Hazard" symbol with the distractors "keep area clean," "wear boots in area," and "dangerous poisonous snakes in area." These distractors were derived from the responses to an earlier version of the symbol that depicted a large boot and a curved line, but these old distractors were no longer appropriate for a newer, redesigned symbol (shown in Figure 1, Symbol 9). To avoid this potential problem, every symbol variant would need pilot testing to identify plausible distractors. The procedure is clearly labor intensive.

The second concern relates to the number of plausible multiple-choice alternatives in the comprehension test. In the preliminary phase the symbol "Keep Door Open" (Figure 1, Pictorial 17) was found to have four highly plausible distractors. Because the number of distractors per symbol was held to three, one of the distractors was discarded. The discarded distractor, "caution, swinging door," was mentioned (incorrectly) in the open-ended test more frequently than the plausible distractors that were actually in the multiple-choice test. Had this item been included (i.e., four distractors), it might have drawn selections away from the correct answer, lowering comprehension performance. A somewhat different situation is the rare case in which a symbol is so unequivocally clear that it generates no plausible alternative answers in an open-ended test. A fair multiple-choice test for this symbol cannot be constructed and is probably not worthwhile, given the results of the open-ended test.

The third concern is how to deal with guessing rates. With four or five alternatives, participants who have no idea what a symbol means will be able to guess the correct answer 25% or 20% of the time by chance alone. If a correction for guessing were strictly applied (simple subtraction), no symbol could exceed ANSI's (1991) 85% acceptability criterion for items with fewer than seven alternatives.

The fourth concern relates to the difficulty of assessing critical confusions in multiple-choice tests. Critical confusions are a type of incorrect response that is extremely important to detect because it could lead to inappropriate, unsafe behavior. Most of the critical confusions detected in open-ended tests are opposite of the concept to be conveyed. Inserting an opposite alternative into a multiple-choice test might give away the correct answer to individuals who might not otherwise have known the answer ("test-wiseness"). To obscure or make it more difficult for a test-wise participant to be able to detect the correct answer, one would have to double the number of opposite answers in the set of alternatives. However, this method would call unfair attention to opposites or negatives, which would affect one's ability to measure the level of critical confusions. Detection of critical confusions might be readily accomplished only in open-ended tests.

The fifth and perhaps most important concern is that multiple-choice tests do not realistically reflect the actual cognitive task that people perform with pictorial symbols in the real world. When confronted with a symbol in actual environments, people do not select from a set of alternative answers. Rather, they generate meaning in a way that reflects the cognitive processes involved in the open-ended test. Retrieval depends on a host of cues that might be present, such as information in the symbol itself, in the surrounding environment, and in the individual's head. The open-ended test is ecologically valid; the multiple-choice test is not.

Another factor investigated in the present research was context. Context was manipulated according to the presence or absence of color photographs of potential locations where the symbols might be viewed. The results showed that the presence of context facilitated comprehension on both the multiple-choice test with more plausible distractors and the open-ended test. One reason for context's beneficial effect is that it eliminates incorrect responses that otherwise might have been given without context. Without context, the symbol's meaning can be ambiguous. For example, photographs of an industrial environment make it clear that a symbol of a person wearing glasses is not denoting an eye-care firm. In real-world environments, ambiguity is less problematic because symbols are seen within an appropriate context. Context reduces erroneous interpretations by providing a mental set that restricts the symbol's possible interpretations (e.g., Biederman et al., 1973). It disambiguates and does so in an ecologically valid way.

Some symbols appeared to provide their own context. That is, they contained location information within the depictions themselves. In such cases it might be expected that additional photographic context would not benefit comprehension performance as much as it would for symbols without this information. To check this possibility, a group of North Carolina State University undergraduates were asked to rate the symbols on the extent to which the symbols inherently provided context information. The symbols were then divided into two sets based on an approximate median split using these ratings. This factor was then added to an ANOVA that also included test method and context.

The results indicated that symbols that contained less inherent context (e.g., the symbols showing simply a glove, a boot, or glasses) were benefited by the presence of photographic context compared with its absence. However, symbols that had more inherent context (e.g., the symbols showing a pharmaceutical bottle, a car-repair shop environment, or fishing area) were benefited less by the additional photographic context. This pattern was shown for both the multiple-choice test with more plausible distractors and the open-ended test, but not for the multiple-choice test with less plausible distractors. Performance on the latter test had high (near ceiling) performance levels regardless of photographic context or inherent location information. These results might help to explain why some studies have found effects of context and others have not. Additional information on these data can be obtained on request (Wogalter & Wolff, 1998).

All the previously published studies investigating the effect of context on symbol comprehension used some sort of verbal description, whereas the present study exclusively used a visual (photographic) context. Because ANSI (1991) recommends that symbols be designed as simply as possible, it is particularly important to provide appropriate contextual reference points during testing. Without context, a simpler symbol might fail to meet a criterion level of comprehension that would otherwise be met or exceeded if a contextual frame had been provided.

In the process of conducting the present research, some initial guidelines for choosing photographs for context were identified. The photographic images used as context can affect people's interpretations. The images can emphasize certain objects over others, and what is seen can be influenced by cropping, lighting, angle, and focal length. A photograph should show an environment rather than a person. If a person is shown, the photograph should not show someone engaging or not engaging in the prohibited or suggested behavior, as it could unfairly bias or cue the test participant. For example, a "Do Not Dig" symbol (Figure 1, Symbol 29) might be accompanied by photographs of a construction site or a residential lawn, but should not show a person digging, as it might suggest to participants that digging is allowed and is demonstrating something about how or where to dig. In fact, a photograph like this was erroneously included as one of the pictures shown to participants in the context-present condition. This was the only symbol for which the scores were actually lower when context was present compared with when context was absent.

The context used in a test might limit the applicability to other contexts. That is, a symbol might be understandable in one context but not understood (or understood differently) in another. For example, the symbol depicting "Flammable Hazard" (Figure 1, Symbol 13), if viewed with industrial workplace photographs (i.e., flammable/combustible materials present), is likely to be interpreted differently than when combined with photographs of an outdoor camping ground (e.g., campfires allowed in area). Symbols should be distributed with information on what kind of context was provided so that warning designers will know the limits of its testing history.

When scoring open-ended symbol tests, consideration should be given to procedures that reduce the likelihood that bias will be introduced into the scores. We offer seven guidelines:

1. Have more than one judge score the answers so that a reliability measure can be calculated. To reduce time and resources, a second judge might score a random sample or subset of scores to get an estimate of reliability.
2. Judges should be familiarized with the referent concept(s) so that they know what idea is actually intended and should be conveyed by the symbol(s).
3. Use independent judges who have not engaged in cross-discussion during the scoring process and who have no stake in the outcome.
4. Decide on the scoring criteria and what kinds of answers are acceptable ahead of time. A lenient gist-criterion is probably more appropriate than a strict verbatim criterion because people will use different wording to convey synonymous answers.
5. If possible, the judges should score the answers blindly; that is, they should not know which particular pictorial is being answered. Preferably the judges should see only the participants' written answers and compare them with the content of an answer key.
6. Avoid extraneous demand characteristics that might unfairly benefit some pictorials over others (e.g., judges should not know which is the "favorite" pictorial symbol).
7. Have the judges look for and record the kinds of errors people make, with particular attention to critical confusions.

When a symbol suggests the wrong or opposite concept, the likelihood of unsafe behavior is greater than if it is simply not understood. In other words, it is one thing not to know what a symbol is, and it is an entirely different matter for a symbol to suggest the wrong concept. The ANSI (1991) Z535.3 standard on safety symbols allows no more than 5% critical confusions (and no more than 15% total errors) for acceptable symbols. Indeed, the goal of reducing critical confusions is probably more important than raising a symbol's comprehension level.

The aforementioned guidelines are not a complete list of procedures that should be considered when scoring open-ended symbol comprehension answers. The overall point is that the scoring should be conducted in a fair, unbiased way.

CONCLUSIONS

In evaluating the comprehensibility of symbols, we recommend using open-ended testing and appropriate context showing the environment in which the symbol will likely be placed. Adding context is an ecologically valid method of raising comprehension performance and consequently might save development costs in trying to reach ANSI's (1991) 85% or ISO's (1984) 67% criteria.

A number of concerns were raised against the use of multiple-choice tests. Without elaborate measures to obtain plausible distractors (such as conducting preliminary evaluations), distractor quality in typical multiple-choice tests could be poor, producing misleadingly high comprehension performance scores. Moreover, multiple-choice tests lack the ecological validity of open-ended tests.

ACKNOWLEDGMENTS

The authors thank Blair Brewster, William Evans, and Neff Walker for their assistance in this research. The helpful comments by Russell J. Sojourner and the two referees were appreciated. Portions of this research were supported by Electromark Co., Lab Safety Inc., and Georgia Institute of Technology. Additional material on this research can be found in Wolff (1995).

REFERENCES

American National Standards Institute (ANSI). (1991). *Accredited standard on safety colors, signs, symbols, labels, and tags, Z535.1–5.* Washington, DC: National Electrical Manufacturers Association.

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real world scenes. *Journal of Experimental Psychology, 97,* 22–27.

Boersema, T., & Zwaga, H. J. G. (1989). Selecting comprehensible warning symbols for swimming pool slides. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 994–998). Santa Monica, CA: Human Factors and Ergonomics Society.

Brugger, C. (1994). Public information symbols: A comparison of ISO testing procedures. In *Proceedings of Public Graphics* (pp. 26.1–26.10) Utrecht, Netherlands: University of Utrecht, Department of Psychonomics.

Cahill, M. C. (1975). Interpretability of graphic symbols as a function of context and experience factors. *Journal of Applied Psychology, 60,* 376–380.

Collins, B. (1983). *Use of hazard symbols/pictorials in the minerals industry* (Report NBSIR 83-2732). Washington, DC: National Institute of Standards and Technology.

Dewar, R. (1994). Design and evaluation of graphic symbols. In *Proceedings of Public Graphics* (pp. 24.1–24.18). Utrecht, Netherlands: University of Utrecht, Department of Psychonomics.

Dewar, R., & Arthur, P. (1994). Warning of water safety hazards with cartoon images. In *Proceedings of Public Graphics* (pp. 7.1–7.10). Utrecht, Netherlands: University of Utrecht, Department of Psychonomics.

Eberhard, J. & Green, D. (1988). The development and testing of warnings for automotive lifts (Tech. Report UMTRI-89-26). Ann Arbor: University of Michigan Transportation Research Institute.

Goldhaber, G. M., & DeTurck, M. A. (1988). Effectiveness of warning signs: Gender and familiarity effects. *Journal of Products Liability, 11,* 271–284.

International Standards Organization (ISO). (1984). *International standard for safety colours and safety signs: ISO 3864.* Geneva, Switzerland: Author.

Katz, S., & Lautenschlager, G. J. (1994). Answering reading comprehension items without passages on the SAT, the ACT, and the GRE. *Educational Assessment, 2,* 295–308.

Laux, L. F., Mayer, D. L., & Thompson, N. B. (1989). Usefulness of symbols and pictorials to communicate hazard information. In *Proceedings of Interface '89* (pp. 79–83). Santa Monica, CA: Human Factors and Ergonomics Society.

Lerner, N. D., & Collins, B. L. (1980). *The assessment of safety symbol understandability by different testing methods* (PB81-185647). Washington, DC: National Institute of Standards and Technology.

Magurno, A., Wogalter, M. S., Kohake, J., & Wolff, J. S. (1994). Iterative test and development of pharmaceutical pictorials. In *Proceedings of the 12th Triennial Congress of the International Ergonomics Association* (Vol. 4, pp. 360–362). Toronto, Ontario, Canada: Human Factors Association of Canada.

Neisser, U. (1987). *The present and the past.* Paper presented at the Second International Conference on the Practical Aspects of Memory, Swansea, Wales.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3,* 519–526.

Silver, N. C., Wogalter, M. S., Brewster, B. M., Glover, B. L., Murray, L. A., Tillotson, C. A., & Temple, T. L. (1995). Comprehension and perceived quality of warning symbols. In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 1057–1061). Santa Monica, CA: Human Factors and Ergonomics Society.

Tulving, E., Mandler, G., & Baumal, R. (1964). Interaction of two sources of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5,* 381–391.

United States Pharmacopoeial Convention. (1993). "Pharmaceutical icon symbols." Rockville, MD: Author [http://www.usp.org].

Vukelich, M., & Whitaker, L. A. (1993). The effects of context on the comprehension of graphic symbols. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp. 511–515). Santa Monica, CA: Human Factors and Ergonomics Society.

Wogalter, M. S., & Marwitz, D. B. (1987). The effect of selecting multiple-choice distractor items around a single target alternative. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 378–381). Santa Monica, CA: Human Factors and Ergonomics Society.

Wogalter, M. S., Marwitz, D. B., & Leonard, D. C. (1992). Suggestiveness in photospread lineups: Similarity induces distinctiveness. *Applied Cognitive Psychology, 6,* 443–453.

Wogalter, M. S., Sojourner, R., & Brelsford, J. W. (1997). Comprehension and training of safety pictorials. *Ergonomics, 40,* 531–542.

Wogalter, M. S., & Wolff, J. S. (1998, August). *Pictorial symbols: Influence of testing technique and context on comprehension.* Paper to be presented at the American Psychological Association meeting, San Francisco, CA.

Wolff, J. S. (1995). *A study of context and test method in evaluating safety symbols* (Tech. Report GIT-GVU-96-07). Atlanta: Georgia Institute of Technology, Graphics, Visualization and Usability Center [http://ftp.gvu.gatech.edu/pub/gvu/tech-reports/96-07.ps.Z].

Wolff, J. S., & Wogalter, M. S. (1993). Test and development of pharmaceutical pictorials. In *Proceedings of Interface '93* (pp. 187–192). Santa Monica, CA: Human Factors and Ergonomics Society.

Zwaga, H. J. G., & Easterby, R. S. (1984). Developing effective symbols for public information. In H. J. G. Zwaga & R. S. Easterby (Eds.), *Information design: The design and evaluation of signs and printed material* (pp. 277–297). New York: Wiley.

Jennifer Snow Wolff recently left the human-computer interaction program at Carnegie-Mellon University to consult in San Francisco, California. She is also a graphic information and interaction design consultant in Pittsburgh. She received an M.S. in information design and technology from the Georgia Institute of Technology in 1995.

Michael S. Wogalter is an associate professor of psychology at North Carolina State University. He received his Ph.D. in psychology from Rice University in 1986.

Date received: March 13, 1997
Date accepted: December 9, 1997