PREDICTORS OF PICTORIAL SYMBOL COMPREHENSION

Stephen L. Young
Liberty Mutual Research Center for Safety & Health
Hopkinton, MA 01748 USA

Michael S. Wogalter
North Carolina State University
Raliegh, NC 27695 USA

Open-ended comprehension testing is a commonly-recommended form of evaluation for safety symbols, but such testing can be costly in terms of time, effort and expense. The present study examines two alternative rating methods that can be used to approximate open-ended comprehension results. The first method, used previously in the literature, had participants estimate the percentage of the population that would correctly interpret the symbol's meaning. The second method involved providing participants with the symbol and its meaning and having them provide a rating of the correspondence between the two. Results demonstrated that both ratings correlated highly with participants' open-ended comprehension results. The present study suggests the utility of alternatives to open-ended testing, especially in the early stages of a symbol's development cycle.

INTRODUCTION

Pictorial symbols are warning components that can be used to attract attention and convey information. A symbol's utility in conveying information is proportional to the extent to which it is comprehensible (understandable) in the population to which it is directed. There are several different methods that can and have been used to assess the comprehensibility of a safety symbol, but one of the more common methods is open-ended evaluation. This procedure has been promoted by ANSI Z535.3: Criteria for Safety Symbols (1998).

Annex B of the ANSI Z535.3 standard presents a methodology for assessing comprehension of symbols that includes an open-ended evaluation. This procedure entails collecting short definitions for symbols using either written responses or oral interviews. A verbal or pictorial context is provided with the symbol to assist users in providing a definition. It is suggested that this open-ended comprehension test be conducted with a sample of 50 people (that are presumed to be representative of the target population). The open-ended responses are then scored as correct or incorrect by some number of judges.

Assuming good inter-rater reliability between judges, a pictorial is deemed, by the ANSI standard, to be "acceptable" if a test of at least 50 people shows the symbol to be comprehended by 85% of the sample with no more than 5% critical confusions (i.e., responses that have the opposite meaning of that intended by the symbol). According to ANSI, any symbol that meets these criteria can be displayed on warnings or signs without any additional verbal information. Symbols that fail to meet the criteria "should be either rejected, modified and retested, used with a supplementary word message, or be supplemented by specialized training. (pg. 30)"

There is a relatively high cost associated with conducting formal, open-ended comprehension tests such as the ones outlined in the ANSI Z535.3 standard. These costs can include

- developing/producing the symbol and any alternatives
- developing/producing data collection materials
- developing/producing contextual descriptions and/or graphics
- recruiting participants
- administering the tests
- compensating participants for their time
- scoring the open-ended responses by two or more judges
- assessing inter-rater reliability and dealing with disagreements between judges to determine comprehension scores

These costs can be substantial—especially when the comprehensibility of a symbol is unknown and potentially uncertain.

Because of these costs, researchers and practitioners have attempted to find more efficient methods for evaluating symbols, especially in the formative stages. For example, Zwaga (1989) had participants provide data about the meaning of different symbols (open-ended comprehension) and an estimate of the percentage of the population that they expected would understand the meaning of the symbol. Except for a few errors, the estimates of comprehension in the population were consistent (and highly correlated) with the results of the open-ended comprehension test. Brugger (1994) also demonstrated the utility of population estimates compared to an open-ended comprehension test.

The present study evaluated two alternative methods to open-ended comprehension tests. The first was an estimation of population comprehension like that employed by Zwaga (1989) and Brugger (1994). The second method involved having participants look at a symbol and its intended meaning (which was provided in written form next to the pictorial) and then having them provide a rating of the correspondence between the two. In both cases, participants gave ratings on scales with anchors of 0% to 100%. It was expected that both alternative assessments would yield a high positive linear relationship with comprehension. If so, they could serve as surrogates for open-ended comprehension testing. In the early phases of symbol development, such procedures might be used to evaluate symbols in an efficient and cost-effective manner.

METHOD

Participants

Fifty participants from the central Massachusetts area were recruited through advertisements in local newspapers. Participants were monetarily compensated for their participation.

Materials

Fifty pictorial symbols were selected from a wide variety of sources including prescription drug label stickers, consumer product labels, industrial safety signs, safety-related clip-art databases, and instruction manuals. Symbols were selected to represent a wide range of situations.

Fifty booklets were created for the open-ended comprehension test. Each page in the booklets contained a symbol, an identifying number (from 1 to 50), and a brief verbal description of the context in which the symbol might appear (e.g., "This symbol might appear on a piece of heavy industrial machinery."). The order of the pages in each of the 50 booklets was randomized so that no participant saw the same order of symbols as another participant.

A response sheet for the open-ended comprehension consisted of a five-page booklet with ten numbered spaces per page. To the right of each number was an open space where participants wrote definitions for the symbols with the corresponding number.

Participants also rated the symbols on two questions. One asked: "What percentage of the general population do you think would correctly interpret the meaning of this symbol?" For the other question, a different booklet was used to present the symbols. In this booklet, each symbol was accompanied by its actual verbal referent definition (i.e., they symbol's meaning). Participants gave ratings to each symbol-definition pair: "To what extent does the symbol convey the meaning of the text?" Below both rating questions was a scale which ranged from 0% to 100%, with increments of 10%. Participants recorded the answers to these two rating questions on a rating sheet, with fifty numbers each followed by two spaces.

Procedure

Participants were run in groups of one to five. After providing consent to participate, participants were first administered the comprehension test. They were given the booklet containing only the numbered symbols and the open-ended response sheets. Participants were told to write the meaning for each symbol as specifically and completely as they could and to progress through the symbol sequence in the booklet until they finished the entire symbol booklet. After completing the open-ended comprehension test, participants were asked to provide an estimate of the population's comprehension for each symbol and then later to provide a rating of the correspondence between each symbol and its referent definition.

RESULTS

The open-ended comprehension data was scored by three different raters. Each rater scored each open-ended response as either 1 ("correct") or 0 ("incorrect"). To be cored as correct, participants had to demonstrate that they understood "the gist" or general meaning of the symbol. Inter-rater reliability was high (alpha = .94).

Open-Ended Responses

The mean comprehension score for all 50 symbols was 44.9% (SD = 28.9%) with a range of 0% to 91%. The distribution of comprehension scores demonstrated that the sample included a range of symbols that were quite evenly distributed across the entire range of comprehension. The distribution can be seen in Figures 1 and 2. Only 5 of the 50 symbols attained a level of 85% correct comprehension (or better). These symbols are shown in Table 1. There were no critical confusions associated with these five symbols. For comparison purposes, the five lowest-rated symbols are presented in Table 2.

Prediction of Population Comprehension

Participants were asked to predict the percentage of the population that would comprehend the meaning of the symbol. Correlation between this measure and lenient comprehension scores was high ($r = 0.79$, $p < .001$). Figure 1 shows the distribution of scores for both measures. This figure demonstrates two interesting characteristics. First, the somewhat flatter slope of the prediction scores (compared to the open-ended comprehension scores) suggests that people may tend to overestimate population comprehension when it should be low and underestimate it when it should be high. This particular type of bias is a common finding in other research domains (e.g., risk perception; e.g., Slovic, Fischhoff & Lichtenstein, 1979).

Second, the figure demonstrates that participants grossly misestimated population comprehension for some pictorials, especially at the lower end. It is possible that participants considered others in the general population

Table 1: The five symbols with greater than 85% correct comprehension

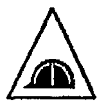
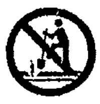
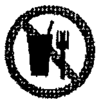
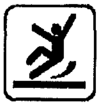
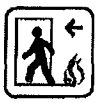
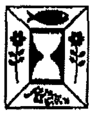
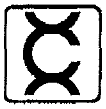
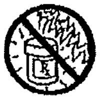
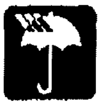
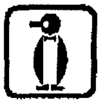
Symbol	Referent	Comp. Score	Rated Under.	Corresp. Rating
	Wear hard hat	85	60	81
	Do not dig	87	72	79
	No food or drink	91	88	83
	Slippery surface	91	74	85
	Fire exit	91	78	87

Table 2: The five symbols with the lowest correct comprehension scores

Symbol	Referent	Comp. Score	Rated Under.	Corresp. Rating
	Perishable food	0	16	7.4
	Carcinogen	0	13	13.5
	Keep drugs away from heat/sun-light	2	41	12.5
	Keep away from water or rain	3	37	23.9
	Keep frozen	3	32	24.2

more adept at interpreting symbols than they. However, it is also possible that people do not necessarily know when they are incorrect in interpreting the meaning of a symbol. Participants provided significantly higher predicted population comprehension scores when they correctly identified the meaning of a symbol in the open-ended comprehension test (m = 58%) than when they provided an incorrect answer (m = 42%), $F\,(1, 88) = 18.5, p < .001$. However, participants would not necessarily know whether they had correctly identified the meaning of a symbol in the open-ended comprehension test. Thus, it is possible that participants provided inflated population comprehension predictions based on an assumption that they correctly interpreted a symbol when, in fact, they did not actually know its meaning.

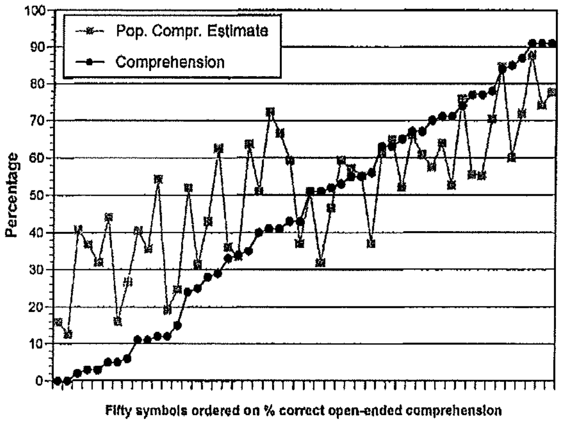
Correspondence Ratings

Participants were shown the intended referent definition of the symbol (along with the symbol) and asked to rate the extent to which the two corresponded (on a scale of 0% to 100%). Correlation between this correspondence measure and the comprehension scores was high ($r = 0.93, p < .001$). Figure 2 shows the distribution of scores for both measures. As this figure demonstrates, participants were generally more accurate in their assessments of correspondence than they were when attempting to predict population comprehension.

DISCUSSION

These results confirm and extend the findings from other studies examining the utility of alternatives to open-ended comprehension testing of symbols. This study demonstrated that participants were able, for the most part, to provide predictions of population comprehension that corresponded to performance on an open-ended test. However, like Zwaga (1989) and Brugger (1994), the correlations across the symbols demonstrate general correspondence between the two measures, while evaluation of the individual symbols shows somewhat more erratic behavior. One potential problem with this method is the fact that people may be incorrect in their interpretation of a symbol and not be aware of the fact that they have incorrectly interpreted it. This might lead to an overestimation of the number of people in the general population who would be able to correctly interpret the symbol. One method to prevent such problems might be to provide par-

Figure 1: Plot of comprehension and predicted population comprehension estimates



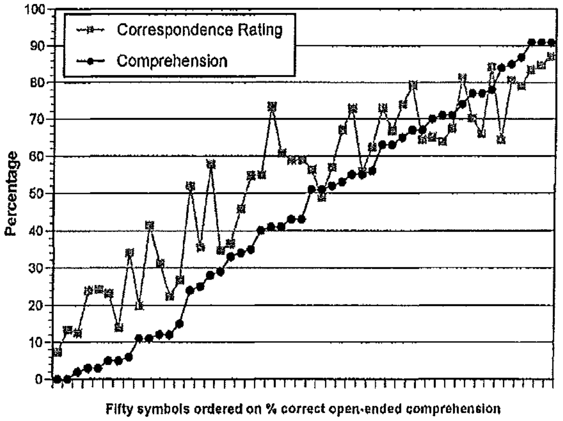
ticipants with the symbol's referent definition (as in the correspondence ratings).

Regarding the correspondence ratings, participants were able to predict open-ended comprehension scores with relative accuracy. One benefit of this procedure is that, in providing the referent definition, participants are less likely to provide responses that are based on an incorrect interpretation of the symbol. Participants could examine the symbol, come to an understanding of its meaning and determine the extent to which the verbal referent was consistent or inconsistent with their understanding.

Two potential drawbacks associated with this procedure include the possibility of demand characteristics in testing and interpretation of the absolute rating numbers. With regard to demand characteristics, it is possible that participants might report greater correspondence between a symbol and its definition than is appropriate. Since participants are not required to provide a definition, this procedure does not objectively demonstrate the true level of understanding possessed by any given participant. However, the high correlation between correspondence ratings and open-ended comprehension scores in this study suggests that participants were not providing inflated correspondence ratings in response to experimenter demand characteristics.

The other issue is that interpretation of the absolute correspondence numbers can be difficult. Unlike the predicted population percentages, it is difficult to predict how many people will correctly interpret a symbol in an open-ended test based solely on correspondence ratings. Thus, this measure can be considered a rough guide to potential performance for a symbol.

Figure 2: Plot of comprehension scores and correspondence ratings



While both rating procedures provided useful information about symbol comprehension, it should be noted that the same participants were used for both the open-ended and rating tests. Thus, we would expect the correlation between the different measures to be higher than we might expect if two different samples were used (one for the comprehension test and another for the ratings). The extent to which this is a significant problem is not yet known, but it is a concern that should be addressed in future research on this issue.

In conclusion, both of these rating methods provide a way of evaluating symbols without the time consuming examination and scoring procedures that formal comprehension testing requires. Such ratings are not yet adequate replacements for open-ended testing. However, they can be beneficial in the development/prototype stages, when multiple versions of the same symbol may require evaluation. It is suggested that these methods can improve and streamline the process of developing and evaluating symbols.

REFERENCES

ANSI (1998). American National Standard. Criteria for Safety Symbols. ANSI Z535.3–1998. Rosslyn, VA: National Electrical Manufacturers Association.

Brugger, C. (1994). Public information symbols: A comparison of ISO testing procedures. *Proceedings of Public Graphics*. pp. 26.1–26.11.

Slovic, P., Fischhoff, B. and Lichtenstein, S. (1979). Rating the risks. *Environment*, *21*, 14-39.

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