

Stephen L. Young and Michael S. Wogalter

Predictors of pictorial symbol comprehension

Keywords: Warnings, pictorials, symbols, labeling

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 several issues related to symbol testing. First, two alternative rating methods intended to approximate open-ended comprehension results were evaluated in both Study 1 and 2. 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. A second issue relates to the way in which people perceive various qualitative aspects of the symbols (e.g., quality of the drawing, clutter, legibility and the extent to which the symbol conveyed a sense of hazard or danger) and how these variables relate to one another. Implications for symbol evaluation are discussed.

Predictors of pictorial symbol comprehension

Safety 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 (e.g., focus group evaluation, rating data, multiple-choice tests, etc.), but one of the more common methods is open-ended evaluation. This type of evaluation method has been recommended by the safety symbol standards promulgated by the Organization for International Standardization's ISO 3864 (ISO, 1984) and the American National Standards Institute, ANSI Z535.3: Criteria for Safety Symbols (1998). Specifically, Annex B of the ANSI Z535.3–1998 standard (which is not technically part of the Standard) presents methodologies for assessing comprehension of symbols that includes open-ended evaluation. The particular procedure outlined in the Annex entails collecting short definitions for symbols from participants using either written responses or oral interviews. A verbal or graphic context can be provided with the symbol to assist users in providing a definition. The open-ended responses are then scored as correct or incorrect by some number of judges.

Assuming the scoring among the judges shows good reliability, a symbol is deemed, by the ANSI standard, to be

'acceptable' if the symbol is understood 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). These criteria assume a sample of 50 participants that are reasonably representative of the population to which the symbol is directed. According to ANSI Z535.3, any symbol that meets these criteria can be displayed on warning labels 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' (p. 30).

The present article evaluates several issues related to symbol comprehension testing in two studies. The first issue deals with two alternative testing methods to open-ended comprehension evaluation. Alternative methods are of interest because of the issue of cost. There is a relatively high cost associated with conducting formal, open-ended comprehension tests such as the ones outlined in ANSI Z535.3 Annex B. 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 there are many candidate symbols per concept to be evaluated or when the comprehensibility of a symbol or group of symbols is uncertain.

Because of these costs, researchers and practitioners have attempted to find more efficient methods of evaluat-

ing symbols, especially in the formative stages of development. 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 predictive value of population estimates in relation to open-ended comprehension scores. ANSI Z535.3 (Annex B) provides for this type of population estimation as a means to evaluate symbols, but only as 'preliminary informal' evaluation that might be conducted for the purpose of selecting only the best candidate symbols from a larger set to be evaluated in a 'final open ended test'. A population-estimation rating evaluation was employed in the present study as one alternative to open-ended testing.

A second alternative method of assessing symbol comprehension examined in the present research is a correspondence evaluation. This involves having participants look at a symbol and its linguistic definition or verbal meaning (which is provided in written form next to the pictorial) and then having them provide a rating of the correspondence between the two. It was expected that both alternative assessments (population estimates and correspondence ratings) would yield a high positive linear relationship with open-ended comprehension scores. If this is true, data from the alternative methods could serve as surrogates for open-ended comprehension testing. Such procedures might be used to evaluate symbols in an efficient and cost-effective manner, especially in the early phases of symbol development.

A second issue addressed in the current article is how people perceive various qualitative aspects of the symbols and how these variables relate to one another. Four qualitative symbol variables were evaluated in both Study 1 and 2 – quality of the drawing, the extent to which the symbol was graphically cluttered or busy, the legibility of

the symbol, and the extent to which the symbol conveyed a visual sense of hazard or danger. These factors were selected for investigation because they dealt with issues that are independent of the symbol's meaning or interpretability. Such issues have not generally been addressed in previous research on safety symbols.

Method

Study 1

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

Materials. Fifty symbols were selected from a wide variety of sources including those from prescription drug label stickers, consumer product labels, industrial safety signs, safety-related clip-art databases, and instruction manuals. These 50 were selected to reflect the wide variety of safety symbols that are present in the 'real world'. Fifty booklets were created for the open-ended comprehension test. Each page in the booklets contained a symbol and an identifying number (from 1 to 50). 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.

An additional 50 booklets of symbols were created for ratings of the percentage of the population that would understand the symbol (population estimates). On each page of the booklet was a symbol with its corresponding identification number. The order of the symbols in each booklet was randomized. A two-page rating sheet had the numbers 1 through 50 with spaces next to the number

where participants could record their ratings. When providing population estimates, participants used the following rating question:

What percentage of the US population would understand the meaning of this pictorial symbol? If you believe that virtually no one would understand it, then you should give a low percentage for your answer. If you believe that virtually everyone would understand what the pictorial symbol means, then you should give a high percentage for your answer. Use the intermediate percentages to reflect estimates in between these two extremes. In your estimates, you should consider all people comprising the US population, including individuals who have not attained high levels of education and non-native individuals and visitors to the US.

A scale of 0% to 100% was employed to represent the percentage of the population that would be expected to understand the meaning of the symbol.

For the correspondence ratings, participants were provided with a separate booklet of symbols. On each page in the booklet was a symbol, the symbol's identification number (1 through 50) and the symbol's referent meaning. The order of the symbols in these booklets was randomized for each participant. Participants provided their correspondence ratings on a two-page form with the numbers 1 through 50 with space provided next to each number for the rating.

Participants also provided ratings of symbol quality according to four rating questions:

- *Quality of the Drawing:* How well is the pictorial symbol drawn? Here, we would like you to make a judgment about the quality of the drawing or, in other words, how professional-looking it is. If the pictorial symbol resembles a child's stick-figure, then it should receive relatively low ratings on this dimension. If the pictorial looks like it was drawn by a professional graphic artist,

then it should receive relatively high ratings on this dimension.

- *Clutter*: How cluttered is the pictorial symbol? If the symbol is very 'busy' looking and has considerable detail (many separate ink markings), then the pictorial should receive relatively high ratings on this dimension. If the symbol is very plain, with relatively few simply-drawn objects, then the symbol should receive relative low ratings on this dimension.
- *Legibility*: How easy would it be to interpret the pictorial in a variety of environmental settings and by a variety of persons? If the relevant parts of the symbol would be legible under a variety of poor environmental (or degraded viewing) conditions (e.g., when viewed from a distance, in smoky or foggy conditions, when reduced in size, or by persons with poor eyesight), then it should receive relatively high ratings on this dimension. If the symbol appears as though it would not be legible when seen at a distance or when reduced, in smoky or foggy conditions or by persons with poor eyesight, then it should receive relatively low ratings on this dimension.
- *Danger*: To what extent does this pictorial symbol depict a dangerous situation? If the pictorial symbol specifically displays a threatening, harmful or injurious situation, then you should give a relatively high rating on this dimension. If it does not show a threatening, harmful, or injurious situation, then you should give a relatively low rating on this dimension.

The questions, with their associated rating scales (which ranged from 0 to 100, in increments of 10), were printed on individual sheets that were bound together in a random order for each participant. Participants recorded the answers to these four rating questions on a rating sheet, with 50 numbers each followed by five spaces.

Procedure. Participants were run in groups of one to five. After providing consent to participate, participants were

first administered the open-ended comprehension test. They were given the booklet containing the 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 symbols in their sequence of appearance (which was randomized) until the entire booklet was completed. After the open-ended comprehension test, participants were then asked to provide ratings of each symbol according to the four symbol quality rating questions. At this time, participants also provided the population estimate ratings. Participants provided ratings for all 50 symbols on a given question before providing ratings for the next question and so forth until all five randomly-ordered questions were answered. After providing these ratings, participants provided correspondence ratings using the booklet with the symbols and their associated referent meaning. This rating was always given last so that exposure to the meaning of the symbol would not influence the results of the previous ratings.

Study 2

Study 2 was a replication of Study 1, with the primary difference being that verbal context was provided with the symbols. This consisted of 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.'). In Study 1, no such contextual information was present. Study 2 was designed to determine if similar comprehension results would be obtained with context present compared to its absence. Fifty participants were recruited from the central Massachusetts area and were paid for their participation.

Results and discussion

In Study 1, the open-ended comprehension data was scored by three different raters. In Study 2, two different

raters were used. Each rater scored each open-ended response as either 1 ('correct') or 0 ('incorrect'). To be scored as correct, participants had to demonstrate that they understood 'the gist' or general meaning of the symbol. Inter-rater reliability was high in both Study 1 ($\alpha = .94$) and Study 2 ($\alpha = .93$). A one-way between-subjects analysis of variance (ANOVA), with responses collapsed across symbols, demonstrated no significant differences between open-ended comprehension scores in Study 1 ($m = 0.424$) vs. Study 2 ($m = 0.449$), $F(1, 98) = 0.19$, $p > .05$. Therefore, in the remaining analyses, the data from Study 1 and 2 are combined.

Open-ended comprehension

Mean comprehension scores for each symbol were developed by averaging the scores from the five raters over the two studies. The mean comprehension score for all 50 symbols was 43.9% ($SD = 28.6\%$) with a range of 0% to

90.4%. The distribution of comprehension scores (see Figure 1) demonstrated that the range of scores was fairly evenly distributed between the lowest and highest scores. Only 5 of the 50 symbols attained a level of 85% correct comprehension or better (see 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 open-ended comprehension scores was high ($r = 0.77$, $p < .001$). A one-way within-subjects ANOVA demonstrated that the estimated percentage of the population who would understand the symbols ($m = 51.9\%$; $SD = 16.5$) was significantly higher than the average percentage of this

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











Symbol	Referent	Comprehension Score (%)	Population Comprehension Estimate (%)	Correspondence Rating (0-100)
	Slippery surface	85	78	89
	Wear hard hat	85	60	79
	Do not dig	86	69	75
	No food or drink	90	86	79
	fire exit	90	79	88

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

Symbol	Referent	Comprehension Score (%)	Population Comprehension Estimate (%)	Correspondence Rating (0-100)
	Perishable food	0	23	20
	Carcinogen	0	16	13
	Keep frozen	2	39	40
	Keep away from water or rain	3	37	32
	Keep drugs away from heat/sunlight	4	36	17

experimental sample that understood their meaning ($m = 43.9\%$; $SD = 28.6$), $F(1, 49) = 8.90$, $p < .005$.

Figure 1 shows the distribution of scores for both measures. This figure demonstrates two interesting points. First, the figure demonstrates that participants grossly misestimated population comprehension for some individual symbols, especially those with low levels of comprehension. 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. 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 ($m = 42\%$), $F(1, 88) = 18.5$, $p < .001$.

Second, 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 open-ended comprehension scores are low and underestimate it when open-ended comprehension scores are high. This particular type

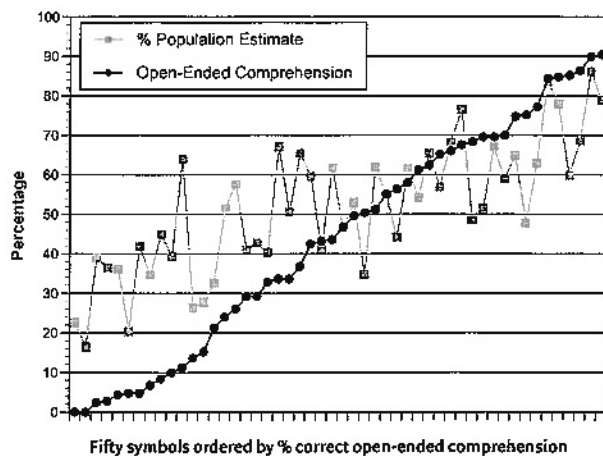


Figure 1. Plot of open-ended comprehension scores and predicted population comprehension.

of bias is a common finding in other research domains (e.g., risk perception; e.g., Slovic, Fischhoff & Lichtenstein, 1979). If anything, such bias is not problematic because its tendency is toward conservative scores. Specifically, overestimation (and even somewhat large deviations) at the lower levels of open-ended comprehension does not result in symbols incorrectly being deemed acceptable. Such overestimations would be an issue if they occurred at the higher end of the open-ended comprehension scores, but, in fact, slight underestimations were observed in these cases. Only one symbol ('No food or drink'; see Table 1) received a population estimate greater than 85% and this symbol was correctly interpreted by 90% of participants in the open-ended comprehension test. Thus, it is quite possible that observed biases in population estimates do not substantially affect the utility of this procedure in the evaluation of symbol comprehension.

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 ($r = 0.87$, $p < .001$). A one-way within-subjects ANOVA demonstrated that the correspondence ratings ($m = 54.4$; $SD = 20.4$) were significantly higher than the average open-ended comprehension scores ($m = 43.9$; $SD = 28.6$), $F(1, 49) = 25.1$, $p < .001$. Figure 2 shows the distribution of scores for both measures. As this figure demonstrates, there was a high degree of concurrence for the symbols that were better comprehended in the open-ended test. However, there was a slight bias toward overestimating correspondence when open-ended comprehension scores were low. As with the population estimates, this bias is conservative. Only two symbols (Fire Exit and Slippery Surface; see Table 1) received correspondence ratings

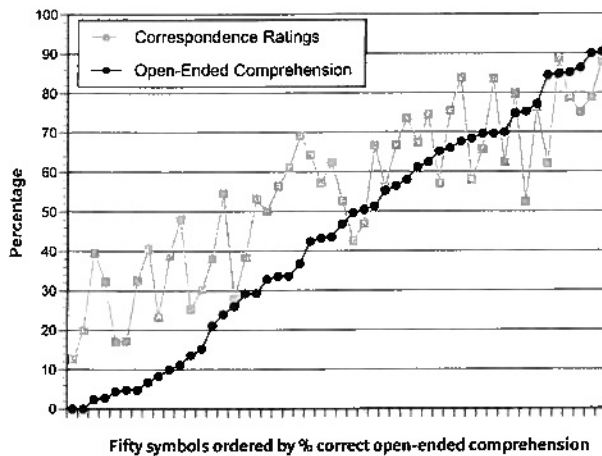


Figure 2. Plot of open-ended comprehension scores and correspondence ratings.

greater than 85%, and both of these symbols were correctly interpreted by over 85% of the participants in the open-ended comprehension test.

Symbol quality

An analysis of the four symbol quality factors was conducted. These factors included the quality of the pictorial drawing (quality), the extent to which the symbol appeared cluttered or busy (clutter), the legibility of the symbol (legibility), and/or the extent to which the symbol conveyed a sense of hazard or danger (danger). Correlations among these four variables and between open-ended comprehension scores showed several significant relationships. First, the more cluttered a symbol, the less it was deemed legible, $r(50) = -.837$. Also, the higher the quality of depiction, the higher the perceived legibility of the symbol, $r(50) = .775$. The quality of the pictorial's drawing ($r = .455$), its legibility ($r = .343$) and the depiction of danger ($r = .452$) were somewhat correlated with open-ended comprehension scores.

General discussion

These results confirm and extend the findings from other studies examining the utility of alternatives to open-ended comprehension testing of symbols. Both population estimates and correspondence ratings could be used to predict open-ended comprehension scores. While there were deviations (sometimes large deviations) between ratings for the alternative methods and open-ended comprehension scores, the results of these studies suggest that such deviations do not substantially affect their utility in evaluating the extent to which a symbol is likely to meet the 85% criterion in open-ended testing suggested by ANSI. If anything, these alternative methods are conservative and are less likely to incorrectly 'accept' a symbol that does not meet the 85% criterion.

This study demonstrated that participants were able, for the most part, to provide reasonably accurate predictions of population comprehension that corresponded to performance on an open-ended test. This finding was observed even though participants were not provided with the referent meaning of the symbol as is suggested in the ANSI Z535.3 annex and in previous research (e.g., Zwaga, 1989). However, like research that has provided referent meanings to participants, the correlations across the symbols demonstrate general correspondence between the two measures, while evaluation of the individual symbols shows somewhat more erratic results.

One issue with population estimation, as employed in the present study, is that participants may incorrectly interpret a symbol and not be aware of the fact that they have incorrectly interpreted it. If so, this could lead to an overestimation of the number of people in the general population who would be able to correctly interpret the symbol. Overestimations of population comprehension observed in the present study may be related to this issue. An obvious remedy to this problem would be to provide the referent meaning of the symbol during the rating process, as it has been done in previous research. Providing

the symbol's meaning has the advantage of reducing errors in ratings that result from misinterpretations of the symbol in the first instance. However, providing the referent meaning could also lead to several undesirable outcomes. First, participants could report greater correspondence between a symbol and its definition than is appropriate. Since participants would not be required to provide a definition, the rating process would no longer necessarily demonstrate the true level of understanding possessed by any given participant. Second, participants could be induced to provide inflated ratings as a result of demand characteristics.

While the correspondence ratings predicted open-ended comprehension scores with a relatively high degree of accuracy, interpretation of the absolute correspondence numbers can be difficult. Unlike the predicted population percentages, it is difficult to determine, from this value, how many people might correctly interpret a symbol in an open-ended test. Thus, this measure can be considered a guide to potential open-ended comprehension performance for a symbol and may be especially useful in preliminary evaluations where one wishes to eliminate candidate symbols from further consideration. It might also be possible to calibrate correspondence ratings using 'marker' symbols for which open-ended comprehension scores are known *a priori*.

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 an issue is not yet known and should be addressed in future research.

Another issue addressed in the present research was the way in which participants considered the symbols according to four qualitative variables (the quality of the drawing, legibility, symbol clutter, and symbol danger-

ousness). Not surprisingly, symbol clutter was negatively related with legibility, and legibility was positively related with drawing quality. Thus, under potentially degraded visual-presentation conditions (e.g., fire exit symbols), a symbol designer might consider a cleaner and less-cluttered presentation over one that might present greater detail at the expense of legibility. The analysis of correlations between these symbol attributes and open-ended comprehension scores do not allow for the provision of design guidelines, primarily because many other factors can influence comprehension scores besides these qualitative aspects.

One other finding of note in the present article is the fact that very few symbols met the 85% correct interpretation criterion recommended by ANSI. Specifically, only five of the 50 symbols would be deemed 'acceptable' to use without an associated verbal message according to the ANSI criterion. We do not claim that these 50 symbols are a representative sample of the population (or universe) of symbols that could be used in safety communications, but they are sufficiently representative to suggest at least one point – that it may be difficult, in general, to design symbols that meet the high standard of 85% correct comprehension. The rather even and monotonic distribution of comprehension scores for these 50 symbols suggests that the ability to abstract various hazards into symbolic form may lie along a continuum rather than being a dichotomous variable (i.e., hazard can be abstracted successfully vs. hazard cannot be abstracted successfully).

In conclusion, this article addresses several issues related to the design and evaluation of symbols. In part, this research demonstrates that there may be several different formal methods of testing or evaluating symbols that can provide designers and researchers with information about the quality and interpretability of pictorials. Factors that influence the selection of a particular method could include time and monetary constraints, the number of symbols being evaluated, the number of people in the sample or expected population, the stage of development

in the design of the symbol, etc. Alternative evaluation methods represent ways to streamline the process of developing and evaluating symbols compared to open-ended testing. Such alternative procedures have been used successfully in the design of symbols (e.g., Kalsher et al., 2000). It is suggested that these methods, along with others (e.g., focus groups, etc.) or even variations on the ones presented in this paper, could be used along with, or possibly apart from, open-ended testing to provide designers and researchers with methods of efficiently 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.
- ISO. (1984). *International standard for safety colours and safety signs*. ISO 3864. Geneva, Switzerland: ISO.
- Kalsher, M. J., Brantley, K. A., Wogalter, M. S., Snow-Wolff, J. (2000). Evaluating choking child pictorial symbols. In: *Proceedings of the IEA 2000/HFES 2000 Congress*. (pp. 4-790-4-793). Santa Monica, CA: Human Factors and Ergonomics Society.
- Slovic, P., Fischhoff, B., Lichtenstein, S. (1979). Rating the risks. *Environment*, 21, 14-39.
- Young, S. L., Wogalter, M. S. (2000). Predictors of pictorial symbol comprehension. In: *Proceedings of the IEA 2000/HFES 2000 Congress*. (pp. 4-294-4-297). Santa Monica, CA: Human Factors and Ergonomics Society.
- Zwaga, H. J. (1989). Comprehensibility estimates of public information symbols: Their validity and use. In: *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 979-983). Santa Monica, CA: Human Factors Society.

Acknowledgements

The authors would like to thank the individuals who performed the open-ended scoring for the symbols – Ilya Bezverkhny, Lyn Melli, and Heidi Wallace of the Liberty Mutual Research Center for Safety & Health, Kenya Freeman of North Carolina State University, and Jennifer Bell of the University of Dayton. Portions of this research were presented at the IEA 2000/HFES 2000 Congress (Young & Wogalter, 2000).

ABOUT THE AUTHORS

Stephen L. Young
Liberty Mutual Research Center for Safety & Health
Hopkinton MA 01748 USA
e-mail: syoung@appliedsafety.com

Michael S. Wogalter
Department of Psychology
North Carolina State University
Raleigh NC 27695 USA
e-mail: wogalterm@AOL.com