Oral and Written Symbol Comprehension Testing:
The Benefit of Cognitive Interview Probing

Kimberly A. Brantley and Michael S. Wogalter

Department of Psychology
North Carolina State University
Raleigh, North Carolina 27695-7801

Traditionally, symbol comprehension is tested using open-ended, written responses. However, responses are often so brief that they may fail to indicate a participant's true understanding of some symbols. In the present study, several test methods were compared to the standard written method to determine if they produce better symbol comprehension performance. The four alternative methods included: written test with probe questioning after all responses were provided, oral test without probe questioning, oral test with probe questioning after responses for all symbols were provided, and oral test with probe questioning after each partially correct or incorrect response. The probe or follow-up questioning technique is taken from the cognitive interview procedure used in eyewitness identification research to elicit more detailed responses. Participants reported their interpretations of 31 safety symbols in one of the 5 test method conditions. Results showed that the test methods that included follow-up questioning elicited more information from participants, and increased comprehension rates in both oral and written test formats. The results have implications for cost-effective symbol design and evaluation.

INTRODUCTION

Pictorial symbols often accompany the instructions and warnings associated with various consumer and industrial products. Potentially, well-designed pictorials can quickly communicate instructions, warnings, and other information to individuals from varied social and educational backgrounds. However, poorly designed symbols may produce dangerous comprehension errors. To reduce errors, symbols must be designed and tested, sometimes iteratively, so that satisfactory comprehension levels are reached.

Comprehension testing is a useful tool in developing effective pictorial symbols and determining which symbols must be revised. Various comprehension testing methods have been discussed in the literature (ANSI, 1998; Brugger, 1994; Dewar, 1994; Silver, Wogalter, Brewster, Glover, Murray, Tillotson, & Temple, 1995; Wolff

& Wogalter, 1993; Zwaga, 1989). Research shows that scores on comprehension tests are affected by factors other than the symbols' quality, including the context provided and test method (Wolff & Wogalter, 1998). Also, iterative testing and design can be a time-consuming and expensive process (Wolff & Wogalter, 1993). Pursuant to the objective of producing understandable symbols at a minimum cost, methods to conduct effective, unbiased symbol comprehension tests appear to be worthy of further investigation. The current experiment explores the potential benefits of non-traditional comprehension test methods.

Most current comprehension test procedures employ multiple-choice and open-ended written techniques. In a comparison study of these techniques, Wolff and Wogalter (1998) concluded that multiple-choice testing lacks ecological validity and inappropriately inflates comprehension scores. ANSI (1998) Z535.3 standards recommend a

procedure that includes comprehension estimation and open-ended testing. The comprehension estimation procedure attempts to identify which symbols will probably fail to reach acceptable levels of comprehension in formal testing, and which symbols will pass, by asking participants to estimate the percentage of the population who would understand a symbol (Brugger, 1994; Zwaga, 1989). The purpose of this procedure is to estimate comprehensibility in a low-cost way before employing more expensive testing procedures.

Most methods of open-ended testing require participants to write the meaning of a pictorial symbol. In addition, the addendum to the Z535.3 standard recommends that participants also be asked to report how they would act in response to the symbol (ANSI, 1998). However, the conventional open-ended written tests could fail to elicit complete answers from participants. Participants' responses are sometimes so brief that they may not indicate their full knowledge of the symbol's meaning; some people may not be able to fully convey their knowledge in writing and others may rush through the test. Incomplete responses create problems for judges during scoring, and may lead to low comprehension scores that fall below some acceptability criterion. It is possible that eliciting more detailed responses would increase comprehension scores, and thus, an increased number of acceptable symbols. Underestimating comprehension levels could add time and cost to producing acceptable symbols (Wolff & Wogalter, 1998).

The limitations of traditional, open-ended written tests might be overcome with oral testing and probing. Oral tests are generally more costly to conduct than written tests, as participants must be tested individually. However, if it facilitates finding understandable symbols and reduces the amount of iterative development and testing, then oral testing could be cost-effective. The probing technique has been used in "cognitive interviews" to elicit information from eyewitnesses beyond their initial answers to police questions. Cognitive interviews involve follow-up questioning, or probing, after initial responses to questions and interviewers are

trained to avoid leading questions. For example, after initial responses are provided, the interviewer asks a general question, such as, "What else can you tell me about this?" Eyewitness testimony research indicates that cognitive interviews elicit more details than traditional interviews, without biasing responses (Hernandez & Alonso, 1997; Mello & Fisher, 1996).

The present study compares oral and written tests and manipulates the presence of follow-up probe questions in both test formats. Two types of probe questions were examined in the oral test format: a *terminal* probe, which occurred after all responses had been reported, and a *continual* probe, which occurred after each partially correct or incorrect response.

METHOD

Design

In this between-subjects design, the five treatment conditions were: Written-Probe Absent, Written-Terminal Probe, Oral-Probe Absent, Oral-Terminal Probe, and Oral-Continual Probe. The first four conditions comprised a 2 (written test vs. oral test) x 2 (probe-present vs. probe-absent) factorial design. The dependent variable was the percentage of correct responses. While the continual probe is compatible with an oral or automated test format, it would be unrealistic to include it in a written test format; therefore, the present study did not examine the use of a continual probe in a written test.

Participants

One hundred individuals participated in this experiment, with 20 participants assigned to each of the five conditions. Fifty percent were undergraduate psychology students, 25% were older adults, and the remaining 25% consisted of community volunteers recruited from such locales as flea markets and community organizations.

Materials

Thirty-one pictorial symbols were selected from existing symbol collections developed by U.S. Pharmacopoeia (USP), Electromark, and Canadian Electric Association. The USP symbols convey pharmaceutical warnings and instructions and the Canadian Electric Association symbols convey warnings about the physical hazards at a hydroelectric power station dam. The Electromark symbols convey warnings and instructions relevant in an industrial setting. The symbol sizes were approximately 4.5 x 4.5 cm and they were presented individually on 11 x 18 cm cards. All symbols were black and white. Additionally, a portable tape recorder was used to record the oral test sessions.

Procedure

Participants were randomly assigned to one of the five treatment conditions. All participants were shown the same 31 symbols and the order of symbol presentation was randomized for each participant. The Written-Probe Absent condition is the traditional symbol comprehension test method, in which participants were asked to write the meaning of each symbol. In the Written-Terminal Probe condition, participants wrote the meanings of all the symbols, then the experimenter circled partially correct or incorrect interpretations and asked the participant to clarify and expand on those answers.

In the Oral-Probe Absent condition, participants orally reported the meanings of the symbols. In the Oral-Terminal Probe condition, participants orally reported the meanings of all the symbols, then, later, the experimenter asked participants to provide more details for partially correct or incorrect responses. In the Oral-Continual Probe condition, participants were asked to provide more details after each partially correct or incorrect response, rather than at the end. The probe questions provided an opportunity for participants to clarify and extend partially correct responses, as well as modify incorrect responses after reviewing the symbol more carefully.

Participants were tested individually in all conditions. Prior to each test session, they were told the contexts and environments in which the symbols might be viewed. In a preliminary practice trial, participants were shown an airline safety pictorial and they reported what it meant. As part of the practice trial, they were shown a complete and specific interpretation and were told that their responses should be complete and specific. Following the test session, participants were told the correct meanings of the symbols that they interpreted incorrectly.

Each response was scored three times – by the experimenter and two independent judges. All responses were scored strictly, meaning that correct responses reflected knowledge of the hazard or instruction, as well as the participant's understanding of an appropriate behavior in response to the symbol. When there was disagreement among the three judges, the score reported by two of the three judges was used in the analysis. There were no responses where all three judges reported different scores. Oral responses were reviewed by the judges for potential biases; biases were reported in less than 1% of the cases. In cases where the judges reported biases, scores for responses given prior to the probe questions were used in the analysis. This approach resulted in more conservative scoring to help eliminate potential bias effects.

RESULTS

Table 1 shows the percentage of correct responses in each condition.

Table 1. *Mean Percentage of Correct Responses as a Function of Test Method*

Condition	% Correct Responses
Written/Probe Absent	66.6
Written/Terminal Probe	81.3
Oral/Probe Absent	56.6
Oral/Terminal Probe	86.9
Oral/Continual Probe	81.8

A between-subjects analysis of variance (ANOVA), based on the percentage of participants who correctly interpreted each symbol for the five test methods, was conducted. The ANOVA indicated a significant effect of test method, $F(4, 150) = 10.8$, $MSe = .043$, $p < .0001$. Tukey's Honestly Significant Differences (HSD) test showed that the three probe-present conditions did not significantly differ and the two probe-absent conditions did not significantly differ. However, the probe-present conditions produced significantly higher comprehension scores than the probe-absent conditions.

A 2 (oral vs. written test) x 2 (present vs. absent probe) ANOVA showed a significant main effect of the probe, $F(1, 120) = 31.9$, $MSe = .046$, $p < .0001$. When probe questions were present, 84.1% of responses were correct, whereas only 61.6% of responses were correct when the probe questions were absent. There was no significant main effect of test format (oral vs. written test) and there was no significant interaction, $p < .05$.

DISCUSSION

Compared to the other test methods, the Written-Probe Absent and comparable Oral-Probe Absent tests produced lower symbol comprehension performance than the tests that included the follow-up probe questions.

There are at least two possible reasons why the use of probe questions was effective. First, the probing encouraged participants to look at the symbols more carefully and, in some cases, understand a few more symbols. Participants tended to complete both the Written-Probe Absent and Oral-Probe Absent test sessions in less time than the other test sessions. Particularly in the Oral-Probe Absent condition, they seemed to spend little time looking at the symbols and they appeared to report only their first impressions.

Second, the probing elicited more information from participants, thus providing a basis for higher comprehension scores. Although participants were instructed to state how they would respond to each symbol, they often neglected to address this in their initial responses. The probing elicited this information, and made the responses more complete.

The two conditions most similar to the test format recommended by ANSI were the two probe-absent conditions. The relatively low comprehension performance in these two conditions could, in real applications, result in a number of symbols failing to meet a specified comprehension criterion. Higher symbol rejection rates, due to erroneously low comprehension scores, may increase the cost of the iterative development process.

There are other aspects of the probing worth mentioning. Although it is not reflected in the average percentage of participants who provided correct responses in the Written-Terminal Probe condition, the follow-up questioning seemed cumbersome and awkward when used in a written test; however, the oral test sessions with the probes resembled the verbal protocol often used in usability tests. Participants "thought aloud" as they deciphered the symbols. This think-aloud protocol identified specific design deficiencies that may provide a basis for design changes. Finally, in a few cases, the probing actually elicited less accurate responses. In these cases, partially correct initial responses were followed by incorrect responses after the probe. This may be due to participants feeling compelled to provide some sort of response after the probe.

While effective symbols must be comprehended at a glance, the probe technique may be valuable to comprehension testing. This research shows that probing in both written and oral test formats elicits more information from participants and yields more correct responses. There was not, however, a significant difference in the number of correct responses between the oral test and written test conditions.

There is a tradeoff associated with the benefits of probing, as it is not conducive to testing groups of participants at once. ANSI (1998) recommends a

sample size of 50 participants; therefore, the cost of one-on-one comprehension testing relative to the amount of information gained must be considered. Further research on symbols and symbol development will likely contribute to effectively producing understandable symbols.

ACKNOWLEDGEMENTS

REFERENCES

American National Standard Institute. (1998). *ANSI 535.3 Criteria for safety symbols*. Washington, DC: National Electrical Manufacturers Association.

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

Dewar, R. (1994). Design and evaluation of graphic symbols. *Proceedings of Public Graphics*. Lunteren: 24.1-24.18.

Hernandez, F. E. & Alonso, Q. M. (1997). The cognitive interview and lie detection: A new magnifying glass for Sherlock Holmes? *Applied Cognitive Psychology,11*, 55-68.

Mello, E. W. & Fisher, R. P. (1996). Enhancing older adult eyewitness memory with the cognitive interview. *Applied Cognitive Psychology,10*, 403-417.

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

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

Wolff, J. S. & Wogalter, M. S. (1998). Comprehension of pictorial symbols: Effects of context and test method. *Human Factors, 40*, 173-186.

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 and Ergonomics Society.