Evaluating Risk Communications: Completing and Correcting Mental Models of Hazardous Processes, Part II

Ann Bostrom,[1] Cynthia J. Atman,[2] Baruch Fischhoff,[3] and M. Granger Morgan[3]

We propose a decision-analytic framework, called the *mental models approach*, for evaluating the impact of risk communications. It employs multiple evaluation methods, including think-aloud protocol analysis, problem solving, and a true–false test that allows respondents to express uncertainty about their answers. The approach is illustrated in empirical comparisons of three brochures about indoor radon.

1. INTRODUCTION

Part I[1] of this two-part article prescribes a process for designing risk communications based on (a) the decisions faced by readers and (b) their prior knowledge. Although the method is advanced on logical and theoretical grounds (based on research in other areas), direct empirical evaluation is needed to assess its products.[2,3] In the following, we advance a general approach to evaluating the impact of risk communications on readers' mental models, which we define as the set of concepts a person uses to understand and generate inferences about a hazardous process. The approach involves a set of reader-based evaluation methods, which supplement the text-based design and evaluation methods presented in Part I. These evaluation methods are demonstrated by application to communications about indoor radon, two of which are developed in Part I.

[1] School of Public Policy, Georgia Institute of Technology, Atlanta, Georgia 30332.

[2] Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

[3] Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

1.1. Dimensions of Text Evaluation in a Mental Models Approach

1.1.1. Defining Goals

A clear set of objectives is needed for any evaluation. Our goal is to help people make decisions about risk. As discussed in Part I,[1] risk communications should improve mental models by (a) adding missing knowledge, (b) restructuring a person's knowledge when it is too general or overly focused on peripheral information, and (c) dispelling misconceptions by deleting inaccurate pieces. Before including a question in an evaluation, the evaluator should establish explicitly its relevance to the goals of the communication (e.g., examining the persistence of misconceptions identified in the preceding exploratory analyses).

1.1.2. Structuring Data Collection

Reader-based evaluation methods vary along a continuum from completely open-ended, in which respondents formulate their own responses, to completely closed-ended, in which respondents select among inves-

Table I. Data Collection Options for Reader-Based Evaluations of Risk Communications

	Strengths	Weaknesses
Concurrent		
Think-aloud protocol	Protocols identify specific problems with text content and organization; can produce surprises	Costly, time-consuming difficult to analyze; samples usually small
Retrospective		
Open-ended	*Least reactive—avoids structuring answers for respondents*	*Coding scheme necessary—data potentially difficult to analyze*
Interview	Identifies how reader structures knowledge, is less reactive than most methods	Costly, time-consuming; samples usually small
Short question, recall	Measures what "sticks" in readers' minds; can measure how readers assign importance	May not elicit information used in actual decision-making; responses context driven; difficult to analyze
Problem solving (scenarios)	Elicits decision-making information and strategies	Frames problems for respondents—may be reactive
Closed-ended	*Data structured, easier, and less expensive to collect and analyze; large samples more feasible*	*Potentially reactive—may misrepresent respondents' knowledge and attitudes*
Knowledge tests (true–false, multiple-choice)	Can verify specific misconceptions and beliefs; data readily comparable	Costly and difficult to design valid questions and response scales

tigator-generated responses. Open-ended evaluation procedures reduce the risk of underestimating people's understanding in cases where their frame of reference differs from that of the evaluator. They reduce the risk of overestimating people's knowledge in cases where people harbor misconceptions that the evaluators do not suspect. They can measure what "sticks" in readers' minds, as well as what information people are able to apply in solving problems. On the other hand, open-ended procedures are expensive to administer and their scoring has an obviously subjective element (which is hidden in structured procedures, where one does not see the variety of interpretations given to particular questions and answers). Closed-ended knowledge tests (e.g., multiple-choice, true–false tests) are relatively cheap to administer, but they are also particularly vulnerable to several potential design flaws: (a) *reactivity*[4]—changing people's beliefs through the cues offered by questions or answer options; (b) *illusory expertise*—restricting the expression of nonexpert beliefs; and (c) *illusory discrimination*—suppressing (or at least not allowing) the expression of inconsistent beliefs.

1.1.3. Timing of Data Collection

In retrospective evaluations, subjects report what they remember from a text. However, retrospective measures do not provide specific information about where the text is confusing or incomplete or about how

readers make sense out of what they read or integrate it with their prior knowledge. Concurrent evaluation methods study text comprehension concurrently.[5–9] For example, *think-aloud protocols* ask subjects to verbalize any thoughts that come to their mind as they read text aloud.[10,11]

Table I summarizes these issues.

1.2. Overview of Studies

The most informative evaluations incorporate multiple methods with compensatory strengths and weaknesses. Here, we demonstrate a mixture of concurrent, retrospective, open-ended, and closed-ended evaluation methods.

Study 1 reports the results of (a) a concurrent evaluation using think-aloud protocols, (b) a multiple-choice test taken from a U.S. EPA study,[12] and (c) a true–false (TF) test derived from mental models interviews. Study 2 presents retrospective evaluations using open-ended recall questions, problem-solving questions, and the two closed-ended tests from Study 1. The studies evaluate three brochures designed to inform and motivate action regarding radon. Two brochures are based on a mental models methodology and employ a decision-analytic perspective to organize information relevant to the radon problem. CMUN[13] (for *n*etwork structure) used an influence diagram.[14] CMUD[15] used a *d*ecision tree.[16,17] The third brochure is *A Citizen's Guide to Radon*

(EPA),[18] widely disseminated by the U.S. EPA.[4] The communications are described in detail in Part I.[1]

In Study 1, subjects participated in a mental models interview about radon both before and after giving a think-aloud protocol as they read one of the three brochures.[19] The subjects were 15 undergraduates from a social science communications class at the University of Pittsburgh who volunteered for extra class credit. They included 14 females and a male. Their average age was 21.

Each of the three brochures was randomly assigned to five subjects. Each experimental session lasted approximately 90 min. The initial mental models interviews were conducted by a single experimenter. A second experimenter then administered the verbal protocol of the reading of the brochure. Following a 5-min break, the first experimenter returned to administer the second mental models interview (blind as to which brochure the subject had read). Finally, each subject completed a TF test and a demographic form.

Study 2 had four experimental groups, one receiving each of the three brochures (CMUN, CMUD, EPA) and a control group. Each task packet contained (in order) (a) a set of instructions (including the instruction not to go back to previous parts of the experiment), (b) a brochure or a filler task (for the control group), (c) three open-ended questions asking readers to describe the main point of their brochure, and (d) the two knowledge tests (EPA and TF).

Subjects were undergraduate students from a social science communications class at the University of Pittsburgh. Fifty-four women and 39 men volunteered for the 50-min experimental session in return for extra class credit. Eleven subjects (12%) reported that their home had been tested for radon. Subjects were almost all about 20 years of age and had a variety of majors. Most (75%) considered themselves "not technically or mechanically inclined."[5] Twenty-four subjects received the first CMU brochure (CMUN); the other two brochures and the filler task were each received by 23 subjects.

In each of two sessions, subjects were randomly assigned, in the order that they arrived, to one of the four experimental conditions. Subjects were instructed to "read the brochure just as you would if you had obtained it because you are interested in radon (not because

you are going to be tested on it). For example, if there is a glossary, just use it as a reference if you need it to understand terms." Subjects were also told that they had 15 min to read and study the brochure. In pretests, the slowest subject took 13 min to read the EPA brochure, which had the most words. After each 5 min, the time remaining was announced. Subjects were instructed to put the brochure away and proceed to the test materials if they finished early.

2. CONCURRENT EVALUATION (STUDY 1)

2.1. Coding

Subjects' comments from the think-aloud protocol were coded as *content* (referring to what was said) or *presentation* (referring to how something was said) and as *negative, positive,* or *nonevaluative.* Negative content comments included confusion about what was being said, or questions about missing information, such as "I think that's kind of, I don't know, fuzzy." Positive content comments included associations with prior knowledge and spontaneous (correct) inferences, for example, "I think they make that really clear." Better-written texts tend to generate fewer comments, because they are easily comprehended.[6]

Comments made by subjects as they read the brochures were coded by two individuals independently. Coders agreed 62% of the time at the detailed level and 72% at the general level. For example, "reader questions about information covered in the brochure" was treated as a general-level content category for which there were three detailed subcategories: the answer was (1) provided previously, (2) within a single page, and (3) later in the brochure. Coding differences between the coders were resolved.

2.2. Concurrent Evaluation Results

As summarized in Table II,[6] each brochure evoked an average of two positive comments per subject, equally divided among comments referring to its content and to its presentation. The EPA brochure produced more than twice as many negative comments (Mann–Whitney rank-order test, $P=0.01$). Three-quarters of these dealt with its contents, expressing confusion about

---

[4] Over a million had been distributed by April 1991 (personal communication with Mark Dickson, U.S. EPA, April 2, 1991).

[5] As chance would have it, a disproportionately large number (22/23) of subjects who received the EPA brochure declared themselves "not technically or mechanically minded." However, when this judgment is covaried with the brochure effect, it contributes little ($P>0.10$) to explaining subjects' knowledge.

[6] The total comments for each individual subject were as follows: CMUN—6, 10, 11, 14, and 23; CMUD—1, 7, 13, 22, and 45; and EPA—11, 17, 20, 35, and 59.

Table II. Total Number of Comments by Condition (*N*=5 per Condition)

Comments	Brochure		
	CMUN	CMUD	EPA
Positive			
On content	4	7	8
On presentation	6	5	4
Total	10	12	12
Mean per subject	**2.0**	**2.4**	**2.4**
Negative			
On content	10	24	48
On presentation	9	4	18
Total	19	28	66
Mean per subject	**3.8**	**5.6**	**13.2**
All			
On content	40	73	112
On presentation	23	17	30
Total	63	90	142
Mean per subject	**12.6**	**18.0**	**28.4**

specific wording or irritation about particular omissions. The details of these comments (and the comparable ones for the CMU brochures) are given in Ref. 19. Most of the negative comments were concentrated on the five and a half pages in the EPA brochure that describe how to use radon detectors and interpret test results. These pages also include risk comparison data. Thus, the EPA brochure confused subjects much more frequently than the CMU brochures.

As expected, the think-aloud protocols highlighted specific problems with the texts' structures and organizations. Despite the small sample size, they provided useful qualitative data for comparisons, indicating that the EPA brochure was more difficult to read and why this was so.

3. RETROSPECTIVE EVALUATION (STUDY 2)

Two structured tests are used here, one developed by the authors on the basis of mental models interviews and the second used in research sponsored by the U.S. EPA to evaluate its radon risk communications.[12] Although the rationale for the EPA test is not given, presumably it reflects EPA's conception of what people need to understand about radon. In this section, we explain our open-ended test (3.1), present and critique the EPA test (3.2), and describe the design of our own structured test, intended to improve on EPA's (3.3).

3.1. Open-Ended Test

All subjects receiving brochures were asked the following open-ended questions designed to assess the immediate impact of what they had read.

(1) What was the main message of the brochure you just read?

(2) What was its single most important point?

(3a) Assume that you own a house in a small town in Ohio. A family with young children has moved into the house next door. They tested their house for radon and found that their level of radon is 8 pCi/L (this is two times the EPA action level). They ask you for your advice. What advice would you give them?

(3b) Asssume that you have not tested your house yet. Would you test your house for radon after finding out about your neighbor's level?

The first two questions involve open-ended recall and judgment about emphasis. Questions 3a and 3b are problem-solving questions, designed to assess the effects of the communications on inferences about radon mitigation.

3.2. EPA Radon Test Design

The EPA test[12] had seven multiple-choice questions regarding what radon is, what health effects it can cause, how one can be exposed to it, how to detect it, and how to mitigate radon problems. Some of these questions exhibit the design problems discussed above (1.1). For example, *EPA test item 6*: What kind of problems are high levels of radon exposure likely to cause? (a) minor skin problems; (b) eye irritations; (c) or lung cancer? (Subjects were allowed to say that they did not know, although this option was not offered explicitly.) This item seems to have been intended to reveal whether people believe the specific misconceptions specified in (a) or (b). Whatever its intent, it exhibits all of the problems described in Section 1.1. The test may inflate the apparent level of knowledge of respondents who believe that radon causes cancer (but do not know what specific kind)—if they infer from response (c) that radon causes lung cancer or answer (c) because it is the only response that includes the term "cancer." Respondents who believe that radon is likely to cause more than one of these health problems may be tipped off by the adjectives "minor" and "irritation" that (a) and (b) are not right answers and choose (c) (illusory discrimination). Finally, this question does not allow the expression of any other

misconceptions, including many expressed in our mental models interviews (e.g., that radon causes breast cancer or noncancer lung problems).[20-22] One person in our open-ended interviews mentioned skin lesions as a possible health effect [and hence might have chosen response (a), if not dissuaded by the adjective "minor"]. None mentioned (b). Thus, both alternatives appear to be weak distracters, thereby encouraging respondents to choose (c).

These problems are illustrated further by EPA test item 3: When radon is measured in a home, which of the following will affect the level most? (a) the time of year it's measured; (b) the amount of industrial pollution around the home; (c) the number of appliances in the home? Theoretically, the amount of radon in a home could be influenced by (a), (b), or (c), depending on respondents' default assumptions about measurement conditions. A home surrounded by uranium mill tailings could give relatively high radon measurements (implicating industrial pollution). Major appliances that move air can affect the air pressure in a home, thereby influencing radon concentrations. Appliances that burn gas with significant amounts of radon can also affect indoor radon concentrations, although this is rare. Homes may be more carefully sealed for heating or air-conditioning purposes, hence retain more radon in winter or summer; radon flux from soil gas can also vary considerably with weather conditions.[23] If all respondents were perfectly informed about radon, the distribution of their answers would reflect the distribution of their assumptions about what the question meant.[24,25]

Thus, both the content and the format of questions can affect their validity and usefulness as measures by lay understanding.[26,27]

3.3. True-False (TF) test design

A 58-item TF test was designed to meet the criteria advanced in Section 1.1. Several examples of items are given in the Appendix. About half of its items address the most common misconceptions observed in the mental models interviews. The remainder cover the basic concepts in an expert model (see Part I, Fig. 1).[1] For each statement, five possible responses were offered: true, maybe true, don't know, maybe false, and false. This response scale allows a finer analysis of responses than a conventional test. Scoring reflected the distance of the respondent's answers from the expert answers: Zero indicated that the respondent agreed with the expert answer; 1 indicated that the respondent answered "maybe true" if the expert answer was "true" or

"maybe false" if the expert answer was "false"; 2 indicated a "don't know" response, 3 meant that the respondent was wrong, but uncertain about the answer; and 4 indicated a wrong answer, for example, "false" when the expert answer was "true."[7]

4. RETROSPECTIVE EVALUATION RESULTS

This section begins with the results of the individual tests, then considers consistency across tests.

4.1. Open-Ended Results

Most subjects receiving each brochure listed lung cancer and test as its main message (Q1). Lung cancer and radon is dangerous were two of three most frequently mentioned "single most important points" (Q2). Test was the third member of this set for CMUD and EPA subjects, while problem is fixable was for CMUN subjects. When asked how they would respond to a test result of 8 pCi/L (Q3a), the most common response for CMUN and CMUD subjects was to hire a contractor; most EPA and control-group subjects gave a general admonition to fix the problem or get more information. All subjects said that they would test their house if their neighbor had found that high a level of radon (Q3b). Thus, while differences among the three brochure groups are not dramatic, they suggest that those who read the CMU brochures may have received more effective messages about contractors and mitigation than those who received the EPA brochure.

4.2. EPA Test Results

Overall, the three brochure groups performed equally well on the seven-item multiple-choice quiz developed for EPA (and presumably addressing the information that EPA hoped to convey).[21] All did much better than the control group [$F(3,88)=21.03$,

$P<0.001$].[8] Differences among the brochures emerged on just two of the EPA items. On item 3 (on radon measurement, see above), 50% of the EPA group chose the first response, (a) *The time of year it's measured*, compared to 17% of CMUN, 30% of CMUD, and 17% of the control group. The modal responses for CMUN (61%) and CTRL (57%) were *Don't know*, whereas for CMUD the modal response (39%) was (b) *The amount of industrial pollution around the home*. As explained above (3.2), all of these responses could be correct. The results may indicate, however, that CMUD did not successfully dispel the misconception that industrial waste is a primary source of indoor radon.

The last question on the EPA radon test asks *What can homeowners do to reduce high radon levels in their home?* Virtually all CMUN and CMUD respondents (100 and 96%, respectively) chose the correct response, (b) *Hire a contractor to fix the problem*, whereas 43% of the EPA group chose *Don't know*, and 9% chose *There is no way to fix the problem*. Most of the control group (78%) responded *Don't know*, with the rest of their responses divided between (a) *Remove the appliances causing the problem* and (b).[30] Thus, in this summary sense, the EPA brochure was less effective than the CMU brochures in motivating reasoned action.

4.3. TF test results

Consistent with the EPA radon test results, a one-way ANOVA on distance scores (3.3) shows that all three brochures improved subjects' performance relative to the control group $[F(3,89)=64.64, P<0.001]$.[21] The 95% confidence intervals for group means showed no performance differences between the respondents who read one of the two CMU brochures, but significantly lower performance for those who read the EPA brochure.[9]

Bonferroni t tests found that the proportion of correct answers was the same in the CMUN and CMUD groups, significantly less in the EPA group, and significantly smaller still in the control group $(P<0.01)$.[10] On

items testing for misconceptions, only CMUN produced a mean proportion of wrong responses lower than that for the control group. When wrong and maybe-wrong responses are combined, the mean proportions were highest for the control group, next highest for the EPA group, and lowest for the two CMU groups. Misconception test items evoked a higher rate of "maybe" responses that did questions about concepts from the expert model. Thus, it appears that stating the facts alone is unlikely to correct misconceptions that belong to readers' initial mental models.

We compared the two CMU brochures with the EPA brochure for differences on individual items, using Bonferroni adjusted t tests to avoid overstating significance levels.[11] By this criterion, CMU respondents outperformed EPA respondents on five questions, those regarding contamination $(P<0.0001)$, decay $(P<0.0001)$, pet death from radon $(P=0.0001)$, radon from mines $(P<0.0001)$, and the health effects of waiting a few weeks before mitigating a radon problem $(P=0.0002)$. Coding the *Don't know* responses as missing values and repeating the analysis produced the same outcome. On no item did EPA subjects perform better.

4.4. Consistency Across Tests

We defined a subject as having consistent and accurate knowlege if that subject's responses to every question on a particular topic agreed with the expert model (i.e., responded *true/maybe true* when the expert answer was *true, false/maybe false* when the expert answer was *false*). There were opportunities to test for such beliefs with five topics that were addressed by the EPA radon test and by multiple items in the TF test: radon detection, mitigation, source, odor, and health effects. Decay and contamination are addressed by several items in the TF test but omitted in the EPA radon test.

Table III shows the number of respondents who agreed consistently with the expert model on each topic. Fewer EPA than CMU brochure subjects performed consistently well; almost no control-group subjects gave complete and consistent responses on any topic. Compared to CMU subjects, EPA subjects had more trouble with the decay and contamination questions, as well as with detection, mitigation, and health effects. Each brochure improved correctness, completeness, and individ-

---

[8] Mean proportion correct: (SD) CMUN, 0.89 (0.05); CMUD, 0.91 (0.06), EPA, 0.90 (0.08), and control, 0.55 (0.34); pooled SD = 0.18. These proportions were not compared directly but were first transformed using an arcsine square-root transformation (Ref. 29, pp. 368–369). The analysis of variance was performed on the transformation of the proportion of correct responses for each subject.

[9] Mean distance scores (SD): CMUN, 0.65 (.34); CMUD, 0.64 (.33); EPA, 1.07 (.28); and control, 1.70 (.23).

[10] For the Bonferroni adjustment, the desired α level is divided by the number of comparisons being made to give the P value required for a "significant" difference.

[11] For this number of comparisons, mean differences of about 1 are significant at a 0.05 level (an unadjusted P value of 0.0009). The unadjusted P values are given here.

Table III. Frequencies of Consistently Expert ("Maybes" Included) Sets of Beliefs by Condition (Study 2)

Topic	Decay	Mitigation	Detection	Odor	Source	Effects
CMUN (N=24)	10	7	7	22	15	16
CMUD (N=23)	11	8	17	23	13	14
EPA (N=23)	0**	2*	2**	21	9	9*
Control (N=23)	0	1	0	10	1	0

* Significant difference between the CMU (combined) and the EPA brochures (χ^2 test), $P<0.05$.
** Significant difference between the CMU (combined) and the EPA brochures (χ^2 test), $P<0.001$.

ual consistency, both across individual test items and across tests; but the CMU brochures did this better.

The TF test shows clear differences between CMU and EPA brochure readers on health effects. As discussed above, the EPA radon test item asked subjects to choose a single health effect from three relatively specific options, possibly producing reactivity, illusory discrimination or illusory expertise. In contrast, the TF test included separate questions on each of several health effects, all based on the mental model interviews. These included questions on cancer (unspecified type), lung cancer, and breathing difficulties. While almost all respondents who read a brochure correctly answered the EPA radon test item on health effects and the TF test items on cancer and lung cancer (93% of CMU, 96% of EPA subjects), a larger proportion of EPA subjects (57%) than CMU subjects (22%) said that radon causes breathing difficulties. In the control group, the TF cancer question produced 78% correct answers and the EPA lung cancer question 43%. Most respondents in the control group (87%) missed the TF breathing difficulties question. Hence responses to items asking for equivalent information are consistent, but the TF test picks up on differences in knowledge that are not captured by the EPA radon test. These differences appear to favor the CMU brochures (and tests). In this case, as anticipated, the EPA radon test appears to underestimate respondents' general knowledge and overestimate their specific knowledge.[21]

5. DISCUSSION

5.1. Conclusions from Studies 1 and 2

In both concurrent and retrospective evaluations, both open-ended and structured procedures were applied to three competing radon brochures. One was EPA's widely distributed Citizen's Guide to Radon[18]; the others, CMUN[13] and CMUD,[15] embodied decision-ana-

lytic structures intended to complete the lay mental models revealed in earlier studies.[19–22]

The CMU brochures appear to outperform the EPA brochure in filling knowledge gaps, contradicting misconceptions, and enabling readers to solve problems about radon. Although these differences are consistent with the mental models approach, they may also reflect other aspects of our procedures, such as the use of comprehension aids (e.g., advanced organizers) or having a single lead author have final editorial authority (rather than allowing the members of a committee to alter the product at all stages).

Of the three brochures, the EPA text appeared to confuse readers most, specifically in passages containing detailed testing and risk comparison information. Readers of the CMU brochures outperformed readers of the EPA brochure on the 58-item structured TF test. Control subjects had even fewer correct answers, primarily as a result of not knowing the answers, rather than having misinformation. CMU brochure subjects' more comprehensive knowledge was seen both in individual items and in the consistency of their answers regarding related questions.

Although the use of small convenience samples may limit the generalizability of the findings, a pilot study study produced comparable results. It compared a preliminary version of the CMUN brochure with A Citizen's Guide. Subjects were 37 students in a Pennsylvania summer high-school science program for those who excel in science. On the EPA quiz, the EPA and CMU groups performed similarly to one another (81 and 88% correct, respectively) and better than the control group (66%; $P<0.01$).[22]12 On a 57-question pilot version of the TF test, the EPA and CMU groups were

---

12 Group and question effects were tested using a two-way analysis of variance on the proportion of correct responses for each group. These proportions were not compared directly but, as in the earlier analyses, were first transformed using an arcsine square-root transformation. The test indicates that group [$F(2,12)=7.30, P<0.01$] and question [$F(6,12)=4.68, P=0.012$] effects are statistically significant.

again similar to one another (66% correct) and better than the control group (35% correct). As in Study 2, the CMU brochure produced greater and more consistent knowledge about the source and decay of radon.

5.2. Choosing Evaluation Methods

Poor risk communications may cause more damage than the risks they are intended to control.[31] They can lead to wrong decisions by omitting key information or failing to contradict misconceptions: They can create confusion by prompting inappropriate assumptions or emphasizing irrelevant information and produce conflict by eroding the audience's faith in the communicator. They can cause recipients to be unduly alarmed or complacent or to undertake ineffective actions. Because communicators' intuitions about recipients' perceptions cannot be trusted, there is no substitute for empirical validation.[2,31-35]

The most demanding kind of evaluation is to see whether recipients undertake actions recommended in the communication.[36,37] However, that standard requires recipients not only to understand the message, but also to see it as relevant to their circumstances. For example, a radon communication intended to motivate testing might not "work" in this sense with recipients who lack the resources to remediate any problems that they find.[17] It might actually be considered a failure, or even unethical, if it motivated action that was personally ill advised.[38]

We addressed the more limited, but still essential, question of whether recipients understand the message, remember it whey they have finished reading, hearing, or seeing the communication, and make appropriate inferences from it. That is, do they have a coherent mental model of the topic? The present study demonstrates a systematic approach to evaluating this kind of success. It is derived from a larger project designed to describe peoples' mental models and create communications aimed at complementing them.[30] It also incorporates principles designed to avoid introducing bias through the evaluation procedure (thereby understating or overstating respondents' knowledge).

The most revealing kind of evaluation is an open-ended interview, allowing respondents to express their beliefs in their own terms (albeit directed at topics of the investigators' choosing). Collected either concurrently or retrospectively, open-ended interviews can be used to evaluate many attributes of text, including both content and organization. They are, however, very resource-intensive and ill suited to studying large numbers of individuals. Open-ended interviews are best used in developmental work, such as improving the design of communications and structured evaluation procedures. Knowing how people think about a topic can increase the chances of formulating test questions that will be understood as intended, tapping rather than shaping beliefs.[3,22,33]

Concurrent evaluations can also include physical measurements such as detailed eye movement protocols or response latencies for problem solving and information search in a communication. Behavior protocols such as eye movement protocols, performance tests, and cloze tests can provide information such as what text or graphics readers are actually processing, how fast they can find information in a brochure, and how easily they can interpret a narrative.[39] Like open-ended interviews, these measures are typically expensive to collect and analyze. However, that may be justified by the increased information yield of knowing how people are processing information.

One popular form of retrospective analysis that was not applied here is the focus group, in which groups discuss a focal topic (e.g., a brochure that each has read). We believe that open-ended interviews come much closer than focus groups to simulating the actual conditions in which brochures are read.[40] The subjects in Study 1 participated in mental models interviews after they gave the think-aloud protocols on the radon brochures.[19] Although it is difficult to generalize from these interviews because of the small sample size, all subjects did offer more expert concepts after they read the brochure.

5.3. Conclusion

Poorly structured or superfluous risk information may bore recipients or frustrate their attempts to understand what is really important. Unless the decision relevance of information is made explicit, people may fail to extract it or not trust their own inferences. Misconceptions that are left unchallenged may coexist with more accurate beliefs. We believe that the mental models approach provides a systematic way to identify and avoid these pitfalls. We combined this approach with findings and methods derived from the survey, reading, and psychological research literatures. The resulting communications and evaluations seem more effective than those produced by EPA's thoughtful, but less theoretically integrated efforts. It is encouraging that the U.S. EPA has revised their 1986 *A Citizen's Guide* in response to these research findings, as well as other pub-

lic commentary.[41] It is less encouraging that it has done so without systematic empirical evaluation of the new text in its final form. That response is akin to failing to remeasure radon levels after remediation designed to correct a problem.

APPENDIX

Excerpt from the CMU Mental Models-Based True–False Test

For each statement listed, please circle the spot on the scale that reflects your opinion about that statement. An answer in the middle means that the proposition is in your opinion neither true nor false (i.e., to the best of your knowledge, the statement could equally well be true or false, you don't know). The scale should be interpreted as follows:

True:	To the best of my knowledge, this is true.
Maybe true:	I think this might be true.
Don't know:	I don't know if this is true or false.
Maybe false:	I think this might be false.
False:	To the best of my knowledge, this is false.

1. Under normal conditions radon is a gas.
   True—Maybe true—Don't know—Maybe false—False
2. Radon can be found outdoors.
   True—Maybe true—Don't know—Maybe false—False
3. Radon-contaminated surfaces stay contaminated unless they are cleaned or renovated.
   True—Maybe true—Don't know—Maybe false—False
4. Over a few days, radon decays (transforms itself) into other substances.
   True—Maybe true—Don't know—Maybe false—False
5. Some radon to which people are exposed comes from rotting garbage.
   True—Maybe true—Don't know—Maybe false—False

ACKNOWLEDGMENTS

REFERENCES

1. C. J. Atman, A. Bostrom, B. Fischhoff, and M. Granger Morgan, "Designing Risk Communications: Completing and Correcting Mental Models, Part I" *Risk Analysis* 14, 779–788 (1994).
2. B. Rohrmann, "Analyzing and Evaluating the Effectiveness of Risk Communication Programs," *Studies on Risk Communication* 17 (University of Mannheim, Mannheim, Germany, Dec. 1990).
3. B. Rohrmann, "The Evaluation of Risk Communication Effectiveness," *Acta Psychologica* 81, 169–192 (1992).
4. Compare with the definition given by D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Houghton Mifflin, Boston, 1966).
5. C. Bereiter and M. Bird, "Use of Thinking Aloud in Identification and Teaching of Reading Comprehension Strategies," *Cognition and Instruction* 2(2), 131–156 (1985).
6. K. A. Ericsson, "Concurrent Verbal Reports on Text Comprehension: A Review," *Text* 8(4), 295–325 (1988).
7. J. Laszlo, D. Meutsch, and R. Viehoff, "Verbal Reports as Data in Text Comprehension Research: An Introduction," *Text* 8(4), 283–294 (1988).
8. Y. Waern, "Thoughts on Text in Context: Applying the Think-Aloud Method to Text Processing," *Text* 8(4), 327–350 (1988).
9. K. A. Schriver, *Plain Language for Expert or Lay Audiences: Designing Text Using Protocol-Aided Revision* (Communications Design Center, Carnegie Mellon University, Pittsburgh, PA, 1989).
10. K.A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data* (MIT Press, Cambridge, MA, 1984).
11. A. Newell and H. A. Simon, *Human Problem Solving* (Prentice-Hall, Englewood Cliffs, NJ, 1972).
12. W. H. Desvousges, V. K. Smith, and H. H. Rink III, *Communicating Radon Risk Effectively: Radon Testing in Maryland*, EPA 230-03-89-408 (U.S. Environmental Protection Agency, Office of Policy, Planning and Evaluation, Washington, DC, 1989).
13. C. J. Atman, *A Citizen's Guide to Radon: What It Is and What to Do About It* (Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, 1989).
14. R. Howard, "Knowledge Maps," *Management Science* 35, 903–922 (1989).
15. M. G. Morgan, *A Citizen's Guide to Radon: What It Is and What to Do About It* (Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, 1989).
16. H. Raiffa, *Decision Analysis: Introductory Lectures on Choices Under Uncertainty* (Addison-Wesley, Reading, MA, 1968).
17. O. Svenson and B. Fischhoff, "Levels of Environmental Decisions," *Journal of Environmental Psychology* 5, 55–67 (1985).
18. U.S. Environmental Agency and U.S. Department of Health and Human Services, *A Citizen's Guide to Radon*, OPA-86-004 (U.S. Government Printing Office, Washington DC, 1986).
19. C. J. Atman, *Network Structures as a Foundation for Risk Communication: An Investigation of Structure and Format Differences*, unpublished doctoral thesis (Carnegie Mellon University, Pittsburgh, PA, 1990).
20. A. Bostrom, B. Fischhoff, and M. Granger Morgan, "Characterizing Mental Models of Hazardous Processes: A Methodology and an Application to Radon," *Journal of Social Issues* 48(4), 85–110 (1992).

21. A. Bostrom, C. J. Atman, B. Fischhoff, and M. Granger Morgan, "Public Knowledge about Indoor Radon: The Effects of Risk Communication," in J. Geweke (ed.), *Decision Making Under Risk and Uncertainty: New Models and Empirical Findings* (Kluwer Academic, Dordrecht, The Netherlands, 1992).

22. A. Bostrom, *A Mental Models Approach to Exploring Perceptions of Hazardous Processes*, unpublished dissertation (Carnegie Mellon University, Pittsburgh, PA, 1990).

23. W. W. Nazaroff and A. V. Nero (eds.), *Radon and Its Decay Products in Indoor Air* (Wiley, New York, 1988).

24. B. Fischhoff, "Value Elicitation: Is There Anything There?" *American Psychologist* **46**(8), 835–847 (1991).

25. B. Fischhoff and M. J. Quadrel, "Adolescent Alcohol Decisions," *Alcohol Health and Research World* **15**(1), 43 (1991).

26. R. Driver, "Pupils' Alternative Frameworks in Science," *European Journal of Science Education* **3**(1), 93–101, (1981).

27. F. Haslam and D. R. Treagust, "Diagnosing Secondary Students' Misconceptions of Photosynthesis and Respiration in Plants Using a Two-Tier Multiple Choice Instrument," *Journal of Biological Education* **21**(3), 203–211 (1987).

28. S. Sudman and N. M. Bradburn, *Asking Questions: A Practical Guide to Questionnaire Design* (Jossey–Bass, San Francisco, CA, 1985).

29. Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice* (MIT Press, Cambridge, MA, 1975).

30. M. G. Morgan, B. Fischhoff, A. Bostrom, L. Lave, and C. J. Atman, "Communicating Risk to the Public," *Environmental Science & Technology* **26**(11), 2048–2056 (1992).

31. B. Fischhoff, "Treating the Public with Risk Communications: A Public Health Perspective," *Science, Technology and Human Values* **12**, 13–19 (1987).

32. C. F. Turner and E. Martin (eds.), *Surveying Subjective Phenomena* (Russell Sage Foundation, New York, 1984).

33. M. Jacobs-Quadrel, *Adolescent Risk Perceptions: Quantitative and Qualitative Dimensions*, unpublished dissertation (Carnegie Mellon University, Pittsburgh, PA, 1990).

34. National Research Council, *Improving Risk Communcation* (NRC, Washington, DC, 1989).

35. E. Roth, M. G. Morgan, B. Fischhoff, L. Lave, and A. Bostrom, "What Do We Know About Making Risk Comparisons?" *Risk Analysis* **10**(3), 375–387 (1990).

36. R. Lau, R. Kaine, S. Berry, J. Ware, and D. Roy, "Channeling Health: A Review of the Evaluation of Televised Health Campaigns," *Health Education Quarterly* **7**(1), 56–89 (1980).

37. N. E. Weinstein, (ed.), *Taking Care: Understanding and Encouraging Self-Protective Behavior* (Cambridge University Press, Cambridge, 1987).

38. B. Fischhoff, "Giving Advice: Decision Theory Perspectives on Sexual Assault," *American Psychologist* **47**(4), 577–588 (1992).

39. K. Schriver, "Evaluating Text Quality: The Continuum from Text-Focused to Reader-Focused Methods," *IEEE Transactions on Professional Communication* **32**(4) (1989).

40. R. K. Merton, "The Focussed Interview and Focus Groups," *Public Opinion Quarterly* **51**, 550–566 (1987).

41. U.S. Environmental Protection Agency, Radon Division, Office of Radiation Programs, *Response to Public Comments on EPA's Draft "A Citizen's Guide to Radon"* (U.S. Environmental Protection Agency, Washington, DC, Sept. 1992).