COGNITIVE ASPECTS OF REPORTING CANCER PREVENTION EXAMINATIONS AND TESTS

Seymour Sudman, Richard Warnecke, Timothy Johnson, Diane O'Rourke, University of Illinois
Andrew M. Davis, RUSH-Anchor HMO
Jared B. Jobe, National Center for Health Statistics
Seymour Sudman, Survey Research Laboratory, University of Illinois,
909 West Oregon, Suite 300, Urbana, IL 61801

KEY WORDS: Regularity, Cancer Screening, Reporting Accuracy, Schema

1. Introduction

Papanicolaou (Pap) smears, mammograms, and breast physical examinations are widely accepted screening procedures commonly used for detection of cancer at an early stage. Knowledge of cancer screening practices in the general population, of course, is based in large measure on self-reports collected in survey interviews, such as the National Health Interview Survey (NHIS).

Several investigators have validated self-reports of these early detection procedures using information from patient medical records as the standard. The general finding is that such procedures are over-reported by respondents.

One approach to improving the accuracy of self-reports in surveys that has received considerable attention in recent years is through the theories and methods of cognitive psychology (Jobe and Mingay, 1991). The cognitive approach seeks to identify respondents' errors in understanding and answering questions and to design questions that minimize such errors.

The remainder of this paper presents findings from an exploration of the cognitive processes that older (50+) female respondents employ in answering questions related to their experiences with several cancer screening procedures.

Hypotheses

Based on the results of focus groups and thinkaloud interviews and the literature review, we formulated the following three hypotheses:

1. More accurate reporting will be found among respondents using a questionnaire form that activates schema relating to health events, physical checkups and gynecological examinations compared to a questionnaire that asks about each of the screening tests separately.

2. For questionnaires that ask about each screening test separately, more accurate reporting will be found for respondents for whom questions are asked about the details of the procedure (who performed it, where it was performed, what happened) before questions about when it happened.

3. For questionnaires that evoke schema by asking about health events, physical checkups and gynecological examinations, more accurate reporting will be found if respondents are required to count individual episodes rather than giving a rate.

Hypothesis 1, which is derived from the ideas of schema formation, asserts that the schema will be easier to retrieve and to use in estimation if they have already been previously activated by questions that parallel the schema that most respondents use.

Hypothesis 2 is based on work of Brewer (forthcoming), Means et al. (1991), Hermann (forthcoming) and Wagenaar (1986) that strongly suggest that time is the least well remembered aspect of an event. Asking about the better remembered aspects first may provide additional cues as to when the event occurred.

Hypothesis 3 was developed by observing that respondents sometimes remember events as being more regular than they really are and forget the exceptions. For such cases, getting respondents to count rather than using a rate should increase accuracy.

2. Methodology

Target Population

The population was women aged 50 and older who were members of the RUSH-Anchor HMO. Women over age 50 are the primary target population for mammograms and breast physical examinations. Below this age, these procedures are not recommended for routine case-finding in the absence of symptoms.

Questionnaires

In order to obtain an approximately equal number of interviews with each of the four questionnaire versions, each interviewer was given a stack of blank questionnaires in rotating version order (different for each interviewer) and told to use the top questionnaire for each subse-

quent interview. Half of the sample was interviewed with a questionnaire based on schema related to health events, physical checkups and gynecological examinations. The other half used a questionnaire asking about each screening test separately. Within the schema questionnaire, half of the respondents were asked questions that encouraged the use of a rate, with possible adjustment (Version 3). The other half were encouraged to count individual events (Version 4).

Within the sample where screening tests were asked about separately, half of the respondents were asked first about details of the diagnostic procedure before being asked about when it occurred (Version 2); the other half were asked first for when it occurred, then about the details (Version 1).

At the end of the interview, we asked respondents for permission to examine their medical records at RUSH-Anchor; 178 of 211 granted permission. These respondents' interviews were then compared with the abstracted records. The analyses are based on this sample of 178 respondents for whom validation information is available.

3. Results

Outcome Measures

To test the hypotheses, it was necessary to specify what is meant by accuracy. There is no single measure since accuracy is really a multidimensional concept.

We report the following four response outcome measures:

- Matched Data: the percentage of reports in which the respondent indicated receiving a screening test and the test was verified in the medical records (verified reports/total sample);
- False Reports: the percentage of reports in which the respondent indicated receiving a screening test for which no matching test was found in the records (unverified reports/total sample);
- Omissions: the percentage of reports in which the respondent indicated no test, but a test was found in the records (unreported tests/total sample);
- No Test: the percentage of reports in which the respondent indicated no test and no evidence of a test was found in the records (verified absences of test/total sample);

Using these response outcome measures, we also estimated the following response quality measures to summarize respondent reports.

- Gross Accuracy: (Matched Data + No Test); this measure is also commonly referred to as an indicator of "concordance;" or "raw agreement rate" (Brown and Adams, 1992);
- Percentage Reporting: (Matched Data + False Positives);
- Percentage Records: (Matched Data + False Negatives); and
- Report to Record Ratio: (Percentage Reporting/Percentage Records); we use this as a measure of net bias in test reporting. A similar measure has been previously used by Brown and Adams (1992).
- Sensitivity: Matched Data /(Matched Data + Omissions). In previous research, measures that are operationally-identical to sensitivity have been referred to as "accuracy" measures (Jobe et al. 1990; Loftus et al., 1992) and "medical record confirmation rates (Brown and Adams, 1992).
- Specificity: No test/(no test + false reports). Sensitivity has also been referred to as a measure of "completeness" (Jobe et al. 1990; Loftus et al. 1992).
- False Negatives: 1 - Sensitivity.
- False Positives: 1 - Specificity.

The Over-reporting of Screening Tests

The basic finding of this study is that respondents generally over-report having received screening tests. As may be seen in Table 1, which summarizes the data over the entire six-year time period, the relative over-reporting averages 29 percent for all three procedures and ranges from 16 percent for mammograms to 51 percent for Pap smears. To put this another way, the percentage of false reports averages 16 percent and the percentage of omissions averages only about five percent.

The null hypothesis that memory errors are unbiased would lead to a test of whether the percentages of False positives do not differ significantly from the percentages of False negatives. To test this hypothesis, the McNemar test for related samples was calculated for each type of examination for each reporting year. Overreporting (False Reports) was found to be significantly greater than underreporting (Omissions) for 6 of the 6 reporting years for Pap smears and for 3 of 6 reporting years both for clinical breast exams and mammograms.

Tests of Hypotheses

We had suggested three hypotheses. In testing these hypotheses, we rely both upon intuitive interpretation of the measures presented in Table 2 and on random effects logistic regression models (Hedeker and Gibbons, forthcoming). This statistical procedure is employed to adjust for the fact that all observations in these analyses are not independent of one another. In fact, most of our 178 respondents contribute 18 self-reports (3 screening procedures for each of 6 years) to the analysis. The clustered nature of these data therefore necessitate an analysis approach that is capable of testing our hypotheses while simultaneously modeling this dependency.

Hypothesis 1 - Hypothesis 1 proposed that more accurate reporting would be found for respondents using a questionnaire form that activates schema related to health events, physical checkups and gynecological examinations compared to a questionnaire that asks about each of the screening tests separately. The results, shown in Table 2, do not support this hypothesis.

One problem in testing this hypothesis is that the sample sizes for each of the four treatments are small, ranging from 37 to 51 cases. Thus, it is not surprising that there is substantial variability by year and between the three screening tests. The most stable measure is simply the total across all years and across all tests.

It may be seen in Table 2 that the ratios of reporting to records by form range from 1.26 to 1.33. Differences in these measures are not statistically significant, nor do they have any practical importance. To confirm these results, there is no consistent pattern of superiority of any form over all three screening tests or over years.

In retrospect, it appears that the treatment was ineffective because most respondents used schema in answering the questions regardless of the form. To put it another way, schema did not need to be activated by the form of the question; they were already activated simply by the topic.

Hypothesis 2 - Hypothesis 2 was also not supported. Hypothesis 2 stated that more accurate reporting would be found on Form 2, which asked details about the procedure before asking about the date, than on Form 1, where the date was asked first. As can be seen in Table 2, there are no differences. The retrospective explanation is that since respondents are using schema to report events, the details of specific events are not used in retrieval as they would be if the respondents were attempting episodic recall.

Hypothesis 3 - Hypothesis 3 predicted that Form 3, which (a) asked respondents to count individual physical examinations first before giving a total and (b) specifically asked respondents if they had missed having a physical exam in any of the past five years, would result in more accurate reporting (i.e., a reduction in the positive bias) as compared to Form 4, which simply evoked a regularity schema.

Again, it may be seen that there are no significant differences between Forms 3 and 4. It is evident that neither of the forms changed the process respondents used in retrieving their schema. In retrospect, this result is not too surprising since others (Blair and Burton, 1987; Means and Loftus, 1991) have also found it difficult to change the retrieval procedures that respondents use.

Menon (forthcoming) found that it was possible to reduce over-reporting by specifically asking respondents about exceptions to regularity. Her questions, however, asked about a very short time period. In this study, it is much more likely that respondents forgot about the non-occurrence of a perceived regular event several years earlier.

Hypotheses 1-3 were tested using logistic regression analysis with random effects. This analysis confirms that questionnaire form had no significant effect on the Gross Accuracy of reports. Additional analyses (not shown) also failed to identify any effects of questionnaire form on Sensitivity and Specificity.

4. Discussion

We had noted based on the focus groups that most women appeared to use schema, such as "I get a mammogram every year" or "I get a Pap smear along with my annual physical." This same heavy use of schema was noted in the Phase II interviews. We thought that we could affect this use of schema by revising the forms used, but we were unsuccessful.

Schema do not necessarily provide poor estimates. For very regular behavior, schema may provide better estimates than efforts to remember individual episodes. Schema can result in overstatements of behavior when respondents forget occasions when the regular behavior was interrupted. That is the case for all

of the health care behaviors studied. As was seen in Table 1, respondents, on average, overstate receiving health care procedures by 29 percent as compared to records.

Pap smears were the most over-reported procedure, possibly because some women may have them regularly every two years. It might be possible to test a question on Pap smears that made this explicit in the introduction. We had speculated that women who had hysterectomies would report Pap smears more accurately, but that did not occur. Although, of course, they reported fewer tests, the percentage of false positives and negatives did not differ between women who did and did not have hysterectomies.

We have assumed that the major cause of over-reporting was the way information was retrieved. It should be pointed out, however, that nothing in our study could rule out the possibility that some of the over-statement is caused by the perceived social desirability of preventive care behavior. This does not mean that many respondents deliberately falsified their answers. Rather, respondents who may have been uncertain about whether they had a procedure every year, said that they did because they knew that this was what they should have done.

This sample was a desirable one, because of our ability to validate information from records. It is possible, however, that membership in an HMO results in more regular behavior and greater use of schema than is found in the general population of women over age 50. It would be useful to replicate this study with a general population sample.

Broader Implications for Cognitive Research on Survey Response

Aside from a better understanding of how women report on these three cancer screening tests, the results also have implications for future research on cognitive aspects of survey response. First, these results strongly suggest that respondents are likely to use schema in reporting about behavior even when the total number of events is small if they perceive the events as regular. The use of schema is even more likely as respondents are asked about less recent events.

The use of schema can lead to highly accurate reporting if indeed the events are very regular. Schema, however, will lead to over-reporting of behavior if respondents forget to exclude exceptions. One might expect that the likelihood of forgetting exceptions would increase with longer time periods, but we saw no evidence of this in this study. If schema are used, then our results would suggest that the order in which questions are asked about details of an event would have no effect on the accuracy of reporting that the event occurred.

Obviously, it would be desirable to tell respondents what retrieval method they should use for greatest accuracy, but this research is in agreement with past efforts that indicate that it is enormously difficult to get respondents to change the way they find easiest to retrieve information. We do not say that it is impossible to do so, but we were unable to do it, even though our focus groups and thinkaloud interviews had given us a good understanding of what methods respondents were actually using.

REFERENCES

Blair, E., & Burton, S. (1987). Cognitive processes used by survey respondents in answering behavioral frequency questions. Journal of Consumer Research, 14, 280-288.

Bowman, J.A., Redman, S., Dickinson, J.A., Gibberd, R., & Sanson-Fisher R.W. (1991). The accuracy of Pap smear utilization self-report: A methodological consideration in cervical screening research. Health Services Research, 26, 97-107.

Brewer, W. (forthcoming). Distortions in autobiographical memory. In N. Schwarz & S. Sudman (eds.), Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer-Verlag.

Brown, J.B., E. Adams, M.E. (1992). Patients as reliable reporters of medical care process. Medical Care, 30, 400-411.

Degnan, D., Harris R., Ranney, J., Quade, D., Earp, J.A., & Gonzales, J. (1992). Measuring the Use of Mammography: Two Methods Compared. American Journal of Public Health, 82, 1386-1388.

Gordon, N.P., Hiatt, R.A., & Lampert, D.I. (1993). Concordance of self-reported data and medical record audit for six cancer screening procedures. Journal of the National Cancer Institute, 85, 566-570.

Hedeker, D. and Gibbons, R.D. (1993). A random-effect ordinal regression model for multilevel analysis. Biometrics, in press.

Hermann, Douglas (forthcoming). The validity of retrospective reports as a function of the directness of the retrieval processes. In N. Schwarz and S. Sudman (eds.), Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer-Verlag.

Jobe, J.B., & Mingay, D.J. (1991). Cognition and survey measurement: History and overview. Applied Cognitive Psychology, 5, 175-192.

Jobe, J.B., White, A.A., Kelley, C.L., Mingay, D.J., Sanchez, M.J. and Loftus, E.F. (1990). Recall strategies and memory for health-care visits. The Milbank Quarterly, 68: 171-189.

King, E.S., Rimer, B.K., Trock, B., Balshem, A., & Engstrom, P. (1990). How valid are mammography self-reports? American Journal of Public Health, 80, 1386-1388.

Loftus, E.F., Klinger, M.R., Smith, K.D., & Fiedler, J. (1990) A tale of two questions: benefits of asking more than one question. Public Opinion Quarterly, 54, 330-345.

Loftus, E.F., Smith, K.D., Klinger, M.R. and Fielder, J. (1992). Memory and mismemory for health events. Pp. 102-137 in J.M. Tanur (ed). Questions about Questions. NY: Russell Sage.

McKenna, M.M.T., Speers, M., Mallin, K. & Warnecke, R. (1992). Agreement between patient self-reports and medical records for Pap smear histories. American Journal of Preventive Medicine, 8, 287-291.

Means, B., & Loftus, E.F. (1991). When personal history repeats itself: Decomposing memory for recurring events. Applied Cognitive Psychology, 5, 297-318.

Menon, G. (forthcoming). Judgments of behavioral frequencies: Memory search and retrieval strategies. In N. Schwarz & S. Sudman (eds.), Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer-Verlag.

Michielutte, R., Dignan, M.B., Wells, H.B., Bahnson, J., Smith, M.,American Journal of Public Health, 79, 1036-1037.

Sawyer, J.A., Earp, J.A., Fletcher, R.H., Daye, F.F., & Wynn, T.M. (1989). Accuracy of women's self-report of their last Pap smear.

Wagenaar, W. A. (1986). My memory: a study of autobiographical memory over six years. Cognitive Psychology, 18, 225-252.

Walter, S.D., Clarke, E.A., Hatcher, J., & Stitt, L.W. (1988). A comparison of physician and patient reports of Pap smear histories. Journal of Clinical Epidemiology, 41, 401-410.

Warnecke, R.B., & Graham, S. (1976). Characteristics of blacks obtaining Papanicolaou smears. Cancer, 37, 2015-2025.

Wooten, R., & Hale, L.N. (1991). Errors in reporting cervical screening among public health clinic patients. Journal of Clinical Epidemiology, 44, 403-408.

TABLE 1
RESPONSE QUALITY MEASURES BY PROCEDURES
(All years combined)

	Percent			
	All Tests	Pap Smears	Breast Examinations	Mammograms
A. Matched data	31.8	27.9	29.8	37.9
False reports	16.2	18.9	18.5	11.1
Omissions	5.3	3.0	8.4	4.5
No test	46.7	50.2	53.3	46.5
B. Percent reporting	48.0	46.8	48.3	49.0
Percent records	37.1	30.9	38.2	42.4
Ratio	1.29	1.51	1.26	1.16
C. Gross accuracy	78.5	78.1	73.1	84.4
Sensitivity	85.7	90.2	78.0	89.5
Specificity	74.2	72.7	70.0	80.7
False negatives	14.3	9.8	22.0	10.5
False positives	25.8	27.3	30.0	19.3
D. N respondents	(178)	(178)	(178)	(178)
N reports	93168)	(1059)	(1058)	(1051)

TABLE 2
RESPONSE QUALITY MEASURES BY FORM
(All tests and years combined)

	Percent			
	Form 1	Form 2	Form 3	Form 4
A. Matched	31.3	32.7	34.0	30.2
False reports	14.4	16.8	18.0	16.2
Omissions	5.1	4.5	6.5	5.3
No test	49.2	46.0	41.5	48.3
B. Percentage reporting	45.7	49.5	52.0	46.4
Percentage records	36.4	37.2	40.5	35.5
Ratio	1.26	1.33	1.28	1.31
C. Gross accuracy	.81	.79	.76	.79
Sensitivity	.86	.88	.84	.85
Specificity	.77	.73	.10	.75
False negatives	.14	.12	.16	.15
False positives	.23	.27	.30	.25
D. N respondents	(51)	(41)	(37)	(48)
N reports	(306)	(246)	(221)	(286)

** $p < .01$.
*** $p < .001$.