Understanding the Cognitive Processes Used by Women Reporting Cancer Prevention Examinations and Tests

Seymour Sudman,[1] Richard Warnecke,[1] Timothy Johnson,[1] Diane O'Rourke,[1] Andrew M. Davis[2], and Jared B. Jobe[3]

This article reports the results of a study that determined the cognitive processes used by women to answer questions about the number of cancer screening tests they received during the previous five years. Most respondents estimated based on a schema that they received such tests on a regular annual basis. These estimates overstated the receipt of health care when compared to record data because the actual gaps between tests were longer than a year and because missed tests were forgotten. An attempt to improve respondent memory by revising the questionnaires was not successful.

Key words: Regular behavior; reporting accuracy; schema.

1. Introduction

There are two main reasons why understanding how women report about cancer prevention examinations and tests is important. At a specific level, cancer prevention tests have significant implications for health policy decisions. At a more general level, such questions are typical of a broad range of questions that ask respondents to report about their behaviors over time periods of many years. It is not surprising that memory may become distorted. What may be surprising is that respondents are often able to report with reasonable accuracy for many kinds of autobiographical behavior for long time periods. Recent work in this area has been summarized in Sudman, Bradburn, and Schwarz (1995), Schwarz and Sudman (1994), Jobe et al. (1990), Jobe and Mingay (1991), Tanur (1992), Tanur and Fienberg (1992), and Bradburn, Rips, and Shevell (1987), but there is still much about reporting of long-term behavior that is not well understood.

In this article we report the results of a study that first attempted to determine the cognitive processes respondents use to answer questions about cancer screening tests over the previous five years, and then used this information to design alternative data gathering procedures that were tested. It was possible to validate reported information from medical records because all the women in this study were members of the same Health Maintenance Organization (HMO).

[1] Survey Research Laboratory, University of Illinois.
[2] RUSH-Anchor HMO.
[3] U.S. National Center for Health Statistics.

1.1.   Focus groups and think-aloud interviews

Two focus groups were conducted, with women over the age of 50, the first with ten African-Americans and the second with eight whites. After the focus groups, a draft questionnaire was developed and tested on 16 respondents using *think-aloud* methods. That is, respondents not only answered the questions, but also indicated the processes they used to answer each question. The major cognitive findings were:

1.1.1.   Comprehension

Pap smears and mammograms are very well understood by women in our focus groups. We found no cases where these terms were misinterpreted or not known. There was some confusion about the term "breast examination." Many women in our focus group and think-aloud samples thought first of mammograms when the term "breast examination" was used, rather than about a physical examination of the breast by a physician.

1.1.2.   Retrieval

From the focus groups and from the think-aloud responses to both versions of the questionnaire, it was clear that only a few respondents were retrieving individual screening examinations. Instead, almost all respondents were relying on schemas. The reasons for this are fairly clear. First, a majority of respondents reported the behavior being studied as very regular, with tests being done on an annual basis. (An examination of the records indicated that actual behavior was less regular than what was remembered.) Second, five years is a long time period, which makes it difficult to retrieve individual events. Both regularity (Menon 1994) and long time periods (Blair and Burton 1987) have been shown to lead to the retrieval of schemas rather than of individual episodes.

   By far the most common schema used was to associate screening tests with annual complete physical checkups. Another major schema was associated with serious illnesses or conditions. Several respondents reported getting regular annual checkups, but in later discussion indicated that the time period between checkups varied from 12 to 15 and sometimes even 18 months. A few reported that they knew they were supposed to get the test annually, but skipped one or more years out of five for a variety of reasons. This could, of course, have happened to others who no longer remembered the exceptions. Even in such cases, however, the schemas provided the basic structure for estimating frequencies.

1.1.3.   Judgment formation

Obviously, respondents use the schemas to make their judgments. Thus, nearly all judgments reported in the think-aloud interviews were based on estimation and not on counting. The easiest and most common estimate was made by women who reported an annual physical every year without exception and who reported having a pap smear or mammogram each time. They simply reported getting five tests in the five years. A few women used this rate estimation procedure, but then adjusted for exceptions.

1.1.4.   Editing

Respondents edit responses if they are concerned about the social desirability of their answers. There was no evidence that respondents found screening tests difficult to discuss

and report. It must be recognized, however, that screening tests for these women are seen as something that is socially desirable, so respondents may be reluctant to report *not* getting these tests. Thus, overreporting of tests could occur.

2.   Methodology

Based on the pilot test results, we conducted face to face interviews with over 200 women in the City of Chicago. The purpose of these interviews was:

1. To determine if the results obtained in the focus groups and cognitive interviews that indicated the widespread use of schemas were also found in a more general population.
2. To determine if respondents who receive screening procedures on a regular basis report more accurately than respondents who receive screening tests irregularly.
3. To test the accuracy of new questionnaires that encouraged the use of schema.

Target Population. The population was women aged 50 and older who were selected at random from the file of members of the RUSH-Anchor HMO in Chicago. Women over age 50 are the primary target population for mammograms and breast physical examinations.

Questionnaires. Half of the sample was interviewed with a questionnaire based on schemas related to health events, physical checkups, and gynecological examinations. That is, respondents were first asked about checkups and examinations, and then asked whether they had received screening examinations as part of the process. The other half used a questionnaire asking about each screening test separately. Additionally, each of these half samples were split in half to study order effects. Complete versions of the questionnaire may be found in Sudman et al. (1994). Appendix A shows abbreviated versions of the questionnaires giving only the major questions.

At the end of the interview, we asked respondents to sign permission forms to allow us 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 was available. We did not observe any differences in demographic characteristics or behavior between those who did and did not grant permission. We recognized, of course, that records are not error-free, but used the records as the standard of comparison.

To measure the effect of regularity of physical examinations, respondents were classified as regular test takers if there was evidence in the medical record that they were tested every year, every other year, or if they received examinations during four of the six years included in the study. Other respondents were classified as irregular, except those with no medical-record evidence of cancer screening examinations during the study period, who were excluded from the comparison of regular and irregular respondents.

3.   Results

3.1.   Outcome measures

To test the hypotheses, it was necessary to specify what is meant by accuracy. There is no single measure because 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). Events that occurred within three months of a self-report were considered a match;
- *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 reports);
- *Percentage records:* (matched data + omissions);
- *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). (Jobe et al. 1990; Loftus et al. 1992);
- *False negatives:* 1 − Sensitivity;
- *False positives:* 1 − Specificity.

3.2. The overreporting of screening tests

The basic finding of this study is that respondents generally overreport having received screening tests. As may be seen in Table 1, which summarizes the data over the entire time period, the relative overreporting averages 29 per cent for all three procedures and ranges from 16 per cent for mammograms to 51 per cent for pap smears. The percentage of false reports averages 16 per cent and the percentage of omissions averages only about five per cent.

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 (Fleiss 1981). Overreporting (false reports) was found to be significantly greater than underreporting (omissions) for each

Table 1.   *Response quality measures by procedures (All years combined)*

		Per cent		
		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	43.3	46.5
B.	Per cent reporting	48.0	46.8	48.3	49.0
	Per cent 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	(3,168)	(1,059)	(1,058)	(1,051)

of the reporting years for pap smears and for half the reporting years both for clinical breast examinations and mammograms.

3.3.   The effects of time

Table 2 shows the response quality measures separately by year for all three tests combined. The results differ between the false reports and the omissions. There appears to be no effect of time on the false reports. We believe that the false reports occur because most respondents are using a schema (that they get the tests on a regular, yearly basis) and forgetting possible exceptions. If such schemas are used, and exceptions are not related to time, then one would expect a consistent level of false reports, and this is what is observed.

Respondents who do not get screening tests regularly are forced to attempt to remember individual tests. It has long been known that the rate of forgetting increases with increased time. This is exactly what is seen with the omissions. There is a steady increase of omissions from 2.1 per cent in 1991 to 11.1 per cent in 1987.

3.4.   The effects of alternative forms

We had hoped that more accurate reporting would be found for respondents using a questionnaire form that activates schemas related to health events, physical checkups, and gynecological examinations compared to the current questionnaire used by the National Center for Health Statistics (NCHS) that asks about each of the screening tests separately. The results, shown in Table 3, do not support this hypothesis. There are no form differences on any of the accuracy measures. There were also no order effects.

Table 2.   Response quality measures by year (All procedures combined)

		Per cent					
		1987	1988	1989	1990	1991	1992*
A.	Matched data	16.9	34.9	40.5	43.2	40.8	15.2
	False reports	14.1	17.4	19.7	17.2	21.3	7.5
	Omissions	11.1	8.4	5.0	3.4	2.1	1.9
	No test	57.9	39.3	34.8	36.1	35.8	75.4
B.	Per cent reporting	31.0	52.3	60.2	60.5	62.1	22.7
	Per cent records	28.0	43.3	45.5	46.6	42.9	17.1
	Ratio	1.11	1.21	1.32	1.30	1.45	1.33
C.	Gross accuracy	74.8	74.2	75.3	77.3	76.6	90.6
	Sensitivity	60.4	80.6	89.0	92.6	95.2	89.0
	Specificity	80.4	69.4	63.9	67.7	62.6	91.0
	False negatives	39.6	19.4	11.0	7.4	4.8	11.0
	False positives	19.6	30.6	36.1	32.3	37.4	9.0
D.	*N* respondents	(178)	(178)	(178)	(177)	(178)	(178)
	N reports	(532)	(524)	(522)	(523)	(534)	(533)

*Partial year.

Table 3.   Accuracy by current and new forms

		NCHS current form	New schema activation form
A.	Matched	31.9	31.9
	False reports	15.5	17.0
	Omissions	4.8	5.8
	No test	47.8	47.3
B.	Percentage reporting	47.4	48.8
	Percentage records	36.8	37.7
	Ratio	1.29	1.29
C.	Gross accuracy	.80	.78
	Sensitivity	.87	.85
	Specificity	.75	.73
	False negatives	.13	.15
	False positives	.25	.27
D.	*N* respondents	(92)	(85)
	N reports	(552)	(507)

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

Menon (1994) found that it was possible to reduce overreporting 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 nonoccurrence of a perceived regular event several years earlier.

3.5. Regularity

We had predicted that respondents receiving regular screening examinations would report better than those receiving irregular examinations, and this prediction was confirmed, as may be seen in Table 4. Section A of the table reveals that the major difference between regular and irregular test recipients is that the percentage of false reports is significantly higher among irregular test recipients. The cognitive explanation of this finding makes it almost tautological. Many respondents use regularity schemas in reporting about these tests. Those respondents whose records indicate that indeed the tests were received regularly would certainly have a lower level of false reports than respondents who did not receive the tests regularly but who thought they did. Net reporting bias is shown in Section B of the table. Overall, the net biases are far smaller for respondents who received tests regularly for each procedure. They range from +8 per cent for pap smears to −18 per cent for breast examinations, with mammograms reported with a net bias of only one per cent for those getting tests regularly. On the other hand, the net biases for those getting the tests irregularly range from a low of +31 per cent for mammograms to +90 per cent for pap smears.

Table 4. *Response quality measures by test regularity (Percentage)*

	Pap smears		Breast examinations		Mammograms	
	Regular	Irregular	Regular	Irregular	Regular	Irregular
Section A						
Gross accuracy	(83.8)	(74.1)	(72.9)	(71.4)	(85.2)	(82.5)
Matched data	67.1	20.1	52.8	24.7	63.9	30.3
No test	16.7	54.0	20.1	46.7	21.3	52.2
False reports	11.0	23.2	6.9	23.2	7.1	14.0
Omissions	5.2	2.7	20.2	5.4	7.7	3.5
Section B						
Percentage reporting	78.1	43.3	59.7	47.9	71.0	44.3
Percentage records	72.3	22.8	73.0	30.1	71.6	33.8
Ratio	1.08	1.90	.82	1.59	.99	1.31
Section C						
N respondents	(35)	(129)	(39)	(131)	(50)	(117)
Reporting years	(6)	(6)	(6)	(6)	(6)	(6)
N reports	(210)	(774)	(234)	(786)	(300)	(302)

Recognizing that the observations examined in this article were not independent of one another we also analyzed these data using random-effects logistic regression (Hedecker 1993). This procedure is able to adjust for the fact that we are assessing multiple reports per respondent by including random person-specific effects and then estimating the variability associated with these random effects. This approach is similar to the way repeated-measures ANOVA accounts for the dependency in data resulting from repeated observations of the same individuals across time; the logistic model was used here because the dependent variable was dichotomous. A random-effects model confirmed our basic findings that questionnaire form had no significant effect on the accuracy of reports, but that test regularity did.

4.   Discussion

4.1.   Survey results

We had noted based on the focus groups and think-alouds that most women appeared to use schemas, such as "I get a mammogram every year" or "I get a pap smear along with my annual physical." This same heavy use of schemas was noted in the interviews. We thought that we could affect this use of schemas by revising the forms used, but we were unsuccessful, primarily because most respondents were already using schemas.

Schemas do not necessarily provide poor estimates. For very regular behavior, schemas may provide better estimates than efforts to remember individual episodes. Schemas 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 per cent as compared to records. Researchers typically observe sharp increases in underreporting caused by forgetting for longer time periods. Such differences in false negatives are observed in Table 2. Because the false positives are far larger, they dominate the net results.

We have assumed that the major cause of overreporting 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 overstatement 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 schemas 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.

4.2.   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 schemas in

reporting about behavior even when the total number of events is small if they perceive the events as regular. The use of schemas is even more likely as respondents are asked about less recent events, as had earlier been observed by Blair and Burton (1987).

The use of schemas can lead to highly accurate reporting if indeed the events are very regular, as was demonstrated in this study. Schemas, however, can lead to overreporting of behavior if respondents forget to exclude exceptions (see also Lessler, Tourangeau, and Salter 1989). 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 schemas 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.

The major source of error in reporting of screening tests thus appears to be from those women who report getting a test on a regular basis, but who, according to the records, do not. 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, and that respondent selected strategies may already be optimal. Instead of trying to change retrieval strategies, it may be better to determine the strategy used, and to then use appropriate memory cues. If respondents use schemas, they can be asked about exceptions for a relatively short time period. If they counted, they could be given additional cues to stimulate memory of events.

Appendix A

Standard version of questionnaire

4a. A pap smear is a routine test, often part of a pelvic examination, where the doctor uses a swab to take a sample from the mouth of the womb or cervix. Have you ever *had* a pap smear test?

4b. How many pap smear tests have you had in the *past five years*, since (MONTH) 1987?

8a. In what month and year did you have your *last* pap smear test?

8c. What about the one before *that*? (KEEP ASKING FOR ALL IN PAST 5 YEARS)

11a. A mammogram is an x-ray taken only of the breasts by a machine that presses the breast against a plate. Have you ever *had* a mammogram?

11b. How many mammograms have you had in the *past five years*, since (MONTH) 1987?

Then, as above for pap smear tests.
Same set for breast examinations.

Schema version

4. Have you had a complete physical or gynecological (female) checkup in the *past five years*, since (MONTH) 1987?

5a. Do you get checkups on a regular basis, or only when you have a health problem?

5b. How often do you get a complete physical or gynecological checkup?

5c. Does the time between checkups ever stretch out to *more* than a year, because of any reason, including problems with making an appointment?

6. Have you ever *skipped* getting your regular checkup during the past five years?

9a. As part of your complete physical or gynecological checkups in the past five years, have you had:
1) a pap smear test
2) a mammogram
3) a breast examination

9b. How many times since 1987 have you had a (TEST) as part of your checkups?
11a. In addition to the pap smear tests you already mentioned, have you had any (*other*) pap smear tests in the *past five years* since (MONTH) 1987?

Same set for mammograms and breast examinations.

5.   References

Blair, E. and Burton, S. (1987). Cognitive Processes Used by Survey Respondents in Answering Behavioral Frequency Questions. Journal of Consumer Research, 14, 280–288.

Bradburn, N.M., Rips, L.J., and Shevell, S.K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. Science, 236, 157–161.

Brown, J.B. and Adams, M.E. (1992). Patients as Reliable Reporters of Medical Care Process. Medical Care, 30, 400–411.

Hedeker, D. and Gibbons, R.D. (1994). A Random-effect Ordinal Regression Model for Multilevel Analysis. Biometrics, 50, 933–944.

Jobe, J.B. and 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.

Lessler, J.T., Tourangeau, R., and Salter, W. (1989). Questionnaire Design in the Cognitive Research Laboratory. National Center for Health Statistics. Vital and Health Statistics, 6(1).

Loftus, E.F., Smith, K.D., Klinger, M.R., and Fielder, J. (1992). Memory and Mismemory for Health Events. In Questions About Questions, J.M. Tanur (ed.). New York: Russell Sage.

Menon, G. (1994). Judgments of Behavioral Frequencies: Memory Search and Retrieval Strategies. In Autobiographical Memory and the Validity of Retrospective Reports, N. Schwarz and S. Sudman (eds.). New York: Springer Verlag.

Schwarz, N. and Sudman, S. (eds.) (1994). Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer Verlag.

Sudman, S., Bradburn, N.M., and Schwarz, N. (1995). Thinking About Answers. San Francisco: Jossey Bass.

Sudman, S., Warnecke, R., Johnson, T., O'Rourke, D., and Davis, A.M. (1994). Cognitive Aspects of Reporting Cancer Prevention Examinations and Tests. National Center for Health Statistics. Vital and Health Statistics, 6(7).

Tanur, J.M. (ed.) (1992). Questions About Questions: Inquiries into the Cognitive Bases of Surveys. New York: Russell Sage Foundation.

Tanur, J.M. and Fienberg, S.E. (1992). Cognitive Aspects of Surveys: Yesterday, Today, and Tomorrow. Journal of Official Statistics, 8, 5–18.