MODE EFFECTS AND CONSUMER ASSESSMENTS OF HEALTH PLANS

Floyd Jackson Fowler, Jr., Patricia M. Gallagher, University of Massachusetts-Boston
Floyd J. Fowler, Jr., Center for Survey Research, 100 Morrissey Blvd., Boston, MA 02125-3393

Key Words: Mode effects, health plan surveys

Background

In the fall of 1995, the Agency for Health Care Policy and Research let three cooperative agreements with the Research Triangle Institute, RAND, and the Harvard Medical School, to work together to develop an instrument to measure consumer assessments of their health care plans. The goal was to have an instrument that would work across various kinds of plans, more and less managed, to provide a basis for comparing consumer experiences. Among the more challenging standards for the instrument were to produce comparable data by mail and by telephone, to be usable in Spanish or in English, and, most of all, to provide data that would be helpful to consumers in making choices among plans.

During the past year and a half, the three organizations, and their sub-contractors, have been working together to develop this instrument. The first public version was released this April. During the past year and a half, candidate questions and survey instruments have been subjected to extensive cognitive testing and field testing, using different modes, with different populations, and with different kinds of health care plans. This paper addresses one particularly pervasive substantive challenge for those developing such an instrument, the way that challenge interacts with effort to design comparable instruments for mail and telephone administration, and the results to date of our tests of efforts to solve these problems.

The Inapplicable Problem

When we first started testing questions, it immediately became apparent that a major challenge was that some questions do not apply to all respondents. The most obvious, and possibly simplest, problem is that asking people to rate medical care within a specific reference period (for instance, we chose six months) does not work for people who have not received any medical care during that reference period. However, the problems are much more pervasive than that, and sometimes much more difficult.

For example, if we want to ask people about whether or not they participate in medical decision making, we have to identify people who have actually had a medical decision to make. If we want to ask about emergency medical care, we have to identify people who have experienced an emergency. If we want to find out about whether health plans approve needed tests and treatments or seeing specialists, we have to identify people who think they have needed tests and treatments or tried to get specialist care.

There are basically four ways that researchers who have tried to assess health care experiences have dealt with the problem of potentially inapplicable questions (Figures 1).

1. They have ignored it, and had everyone answer all the questions.

2. They have offered a "does not apply" option, without exactly specifying what the criteria for applicability were.

3. They offered an inapplicable alternative to questions, which explicitly describes what inapplicable means.

4. Prior to the focus question, they have asked respondents explicitly whether or not they have had the kind of experience that the follow-up question is designed to measure.

Issues Related to Mode of Data Collection

Being able to collect data both by mail and by phone is very important to having a universally useful instrument. Depending on the available information about sampled individuals, and the characteristics of samples, one approach or the other may be best in order to carry out a survey with an adequate response rate. Indeed, the potential for using combinations of modes, in order to maximize the rate of response, is a particularly desirable feature. In order to have the option of collecting data by either mode, however, it is important that the results be comparable.

There is extensive literature comparing data collected by various modes. When Hochstim (1967) did one of the earliest such studies, he did over 1000 comparisons of between-mode results and found only 51 differences in aggregate answers. Many subsequent researchers have found that comparable data emerge from different modes of data collection. When

Figure 1

Four Ways to Deal with Inapplicable Questions

In the past 6 months, how often were you involved as much as you wanted in decisions about your health care?

A) ☐ Always
☐ Usually
☐ Sometimes
☐ Never

B) ☐ Always
☐ Usually
☐ Sometimes
☐ Never
☐ Does Not Apply

C) ☐ Always
☐ Usually
☐ Sometimes
☐ Never
☐ There have been no decisions about my health care in the last 6 months

D.) (SCREENER) For the last 6 months, did you and a doctor or other health profession have any decisions to make about your health care?

☐ Yes
☐ No (Skip to...)

---

questions pertain to issues that have a high component of social desirability, self-administered forms tend to elicit more responses that might be judged socially undesirable. There also is some evidence that when people are asked for self-reports about their current status, self-administration produces more negative or critical self-descriptions. Finally, the mechanics of self-administered forms pose some problems for making them comparable to interviewer administered surveys, particularly when there are significant skips. This latter issue, the fact that skipping questions in a consistent way without the assistance of an interviewer may be problematic, is the area in which the issue of mode comparability intersects with the problem of the inapplicable questions discussed above.

The Experiment

A sample of Washington state employees enrolled in a single health plan were randomized to one of two data collection protocols. Half were sent an advance letter; then interviewers called to attempt to conduct interviews with the designated individuals. For the other half, a fairly standard mail protocol was used: initial mailing, postcard, second mailing of questionnaire to non-respondents, followed finally by a telephone reminder call. Response rates were similar for the two protocols, around 79 percent; there were about 100 respondents in each sample.

The questionnaires were designed to be as comparable as possible. The wording of questions themselves was virtually identical across forms. However, there was a fundamental difference in the way that the applicability of questions was handled. In the mail survey, questions that did not apply to all respondents had a response alternative that explicitly described the class of people to whom the question did not apply (Option C in Figure 1). Using such a strategy proved to be impossible by telephone, because the definition of the inapplicable conditions was so complicated. Therefore, we had to go to an approach that asked a prior screening question to identify those people to whom questions applied (e.g., Option D in Figure 1). An effort was made to make the screening questions mirror the inapplicable alternative in the mail questionnaire. However, as will be seen shortly, that effort was not consistently successful.

Results

The first step was to compare aggregate distributions by mode to find out if they were the same or different (Table 1). It turns out it makes a big difference how the data are analyzed. Overall, when all respondents are included in a table, including those to whom the questions did not apply, 10 of 22 questions had a different distribution if they applied to fewer than 90% of respondents.. In contrast, if the "INAPS" are left out of the tables, and only the distributions of substantive answers are compared, there are only 5 significant differences. The way the "INAPS" were handled was responsible for 9 out of 14 significant differences. Table 1 shows the breakdown by the percentage of respondents to whom questions applied.

The four questions that applied to all respondents and that produced significant differences were an odd group: 3 questions asked for factual information in

categories (number of visits to doctors and specialists and dollars spent out of pocket on medical care) and one of 4 0-to-10 ratings differed by mode. In addition, the answers to one of 7 distributions using a 5-category hard to easy rating differed by mode. Also, it should be noted that 3 of the 4 0-to-10 ratings showed a small but statistically significant difference in the means, and the easy to hard responses consistently bordered on being significantly different by mode. We are doing more work on those items. However, because it accounts for 9 significant differences, the balance of this paper focuses on the problem of applicability.

The dominant feature of the differences in distributions was the way that the inapplicable category was handled. On the telephone, when the skip pattern was under interviewer control, and the respondent was asked a screening question without knowing what the followup question would be, respondents were much more likely to give the answer that indicated that the followup question did not apply to them. In contrast, when the inapplicable alternative was provided as one more check box in the mail version, many more respondents answered the substance of the question rather than checking the inapplicable box.

We examined the inapplicable check box to see if there were ambiguities in the definition that did not appear in the screening questions used in the telephone interview. Our analysis suggested that some of the differences in the way the boxes were used could be explained by different wording. However, for the most part, there seems to be a real tendency for respondents to want to answer questions rather than skip them.

The key question is whether or not these mode differences make a difference in the conclusions you would reach about the plans. As noted previously, when the inapplicable responses are omitted, the distributions of substantive answers usually were similar for both groups, even though some mail respondents answered the question who probably would have been declared inapplicable if they had been interviewed by phone.

Tables 2 through 4 provide three different examples of how distributions were affected by mode of data collection and the different approaches to identifying people to whom questions did not apply. In each table, the mail and phone responses are compared with and without the inapplicable responses being included.

In Table 2, the focus is on peoples' answers about whether or not they were included in decisions as much as they wanted. When the inapplicable answers are not included, the distributions are virtually identical. However, almost a third of the telephone respondents said that there were no decisions made, so the question did not apply, while only 5% of the mail sample checked the box indicating the question did not apply.

In Table 3, we look at the answers to a question about whether respondents said they had to wait more than 15 minutes past an appointment time in the past 6 months. In this case, there is close to a significant difference in the answers even when the inapplicable responses are not included. The difference would no doubt be significant if there were more cases. As the bottom table shows, most of the people in the mail sample to whom the question probably did not apply responded that they had to wait more than 15 minutes in a doctor's office. Apparently, they could not resist the opportunity to communicate their experience.

In Table 4, we compare answers to a question about whether or not the health insurance plan refused to approve or pay for a test or treatments that the respondent thought were needed and should be covered. Again, as in Table 2, there was no difference in the distributions of answers by mode among those who answered the question. Again, the bottom of the table shows that many more people answered the question in the mail version than in the phone version, creating significantly different overall patterns of response.

Conclusions

Our first conclusion from this field test was that we needed to standardize the way that the inapplicable answers were handled across the modes of data collection. Even though it means more skip instructions, and hence a more complex questionnaire in some ways, our revised instruments have explicit screening questions in the mail and telephone versions. We are in the process of testing them to see if the results produced are more consistent across modes.

Second, it is encouraging that there are not more differences between modes. When we think that the questions are consistent and the same populations of respondents are answering them, most of the distributions are similar. Despite the fact that there is a body of data that suggests that self-descriptions differ by mode of data collection, these descriptions of peoples' health care experiences do not seem to differ very much. A careful examination of the results suggests that there may be a tendency to elicit more criticism or negative reports by mail than by phone. As noted, there also was some evidence that the 0 to 10 scale produced more negative average ratings in the mail version than by phone. We are doing further tests of these issues, with larger samples for more power. We also have worked on the presentation of the 0 to 10 ratings, in the hope of making the experience more comparable across the two modes. In general, while social desirability may be the driving force in the results having to do with self-descriptions,

we think that difference in the way the questions are structured is the key to most of the differences we observe here. The fact that there was very little evidence for primary mode effects encouraged us that it may be possible to collect comparable data by mail and by telephone.

Finally, for anyone collecting data about peoples' health care experiences, we certainly have ample evidence that how the problem of inapplicable questions is handled makes a big difference in the resulting data. Moreover, this particular sample moderated the problem, because they used more health care than average--and hence contingent questions applied to more respondents.

Even when there are fairly explicit instructions about what to do if the question does not apply, in a self-administered questionnaire people have difficulty following the instructions. Surely it is unreasonable to think that people will consistently figure out whether or not questions apply, if they are given no guidance (and just told to check a box if the question "does not apply"). Depending on mode and how the data are analyzed, there are big differences in what one would say about a plan. In Table 2, one could say that 48% or 68% of enrollees say they were "always" involved in decision making. In Table 3, 38% or 65% said they waited 15 minutes past an appointment. In Table 4, 9% or 15% had payment for needed tests or treatments refused. Those are some pretty big differences.

Anyone who wants to gather data about patient experience has to deal with the inapplicable problem. Our research to date strongly suggests that doing so explicitly, with well-defined screening questions, is probably the best approach, regardless of mode of data collection.

Table 1

Whether Responses by Telephone and Mail are the Same or Different ($p<.05$) by Percent to Whom Questions Applied, With Respondents Who Said Question Did Not Apply Included and Excluded from Distribution

How Distributions Compare	% To Whom Question Applied			
Including INAPS	*≥90%	50-90%	≤50%	Total Questions
Same	NA	5[1]	7[1]	12
Different	NA	7	3	10
Excluding INAPS				
Same	35[2]	12[1]	9[2]	56
Different	4	0	1	5

1 Includes one comparison for which $p<.10$
2 Includes 2 comparisons for which $p<.10$

* Most of the these questions applied to all respondents

Table 2

Distribution of Answers by Telephone and Mail, "In the past 6 months, how often were you involved as much as you wanted in decisions?" With Respondents Who Said Questions Did Not Apply Excluded and Included In Distributions

EXCLUDING INAPS		MAIL	PHONE (WITH SCREENER Q)
Always		68%	70%
Usually		24	24
Sometimes		6	5
Never		2	1
		100%	100%
	N=	104	67
	p=.97		
Including INAPS			
Always		65%	48%
Usually		23	16
Sometimes		5	3
Never		2	1
INAP-No		5	32
Decisions		100%	100
	N=	106	98
	p <.0001		

Table 3

Distribution of Answers by Telephone and Mail, "In the past 6 months, have you ever had to wait more than 15 minutes past your appointment time?" With Respondents Who Said Questions Did Not Apply Excluded and Included in Distributions

EXCLUDING INAPS		MAIL	PHONE (WITH SCREENER Q)
Yes		65%	51%
No		35	49
		100%	100%
	N=	105	74
	p=.10		
Including INAPS			
Yes		63%	38%
No		34	37
INAP-No		3	25
Appointments		100%	100%
	N=	106	98
	p <.0001		

Table 4

Distribution of Answers by Telephone and Mail, "In the past 6 months, has your health insurance plan refused to approve or pay for any tests or treatment? With Respondents Who Said Question Did Not Apply Excluded and Included in Distributions

EXCLUDING INAPS	MAIL	PHONE (WITH SCREENER Q)
Yes	15%	14%
No	85	86
	100%	100%
N=	94	65
p=.85		
Including INAPS		
Yes	13%	9%
No	76	59
INAP-No Tests or Treatments Needed	11	32
	100%	100%
N=	106	98
p <.001		