CHOOSING QUESTIONS TO MEASURE THE QUALITY OF EXPERIENCE WITH MEDICAL CARE PROVIDERS AND HEALTH CARE PLANS

Floyd Jackson Fowler, Jr., University of Massachusetts, Boston
Center for Survey Research, 100 Morrissey Boulevard, Boston, MA 02125-3393

Key Words: Consumer surveys, satisfaction

Introduction

The Consumer Assessment of Health Plans (CAHPS) project was designed to develop a survey protocol to measure consumer experiences with their health plans. The purpose of the survey instrument was to gather consistent data across health plans so they could be compared and so people making choices could benefit from the experiences of those who were enrolled in the plans. In order to meet this goal, the instrument had to meet a number of complex standards. Among the criteria for an acceptable instrument were that it provide meaningful data for all kinds of health insurance plans and programs, that it contain a core of data that would be useful to virtually all consumers of health care, and that it meet high standards for survey research methodology.

In designing the survey instrument, the first steps built on consumer focus groups, conducted by the Research Triangle Institute, and a collation of survey items collected from many instruments that had been used to assess responses to health plans. As one can imagine, the number of possible things to ask about was large, and the way the questions were designed varied greatly.

The first phase of the CAHPS project involved extensive testing of alternative questions and ways of asking questions. A key part of the initial question evaluation involved cognitive interviews. Some of the cognitive interviews used think-aloud interview techniques, while others used extensive debriefing questions after individual questions or series of questions. The basic goal was to learn how respondents understood proposed questions, how they constructed their answers, and what the answers actually meant. The goal of the work was to find a set of questions that would best measure what consumers had to tell us about their experiences with their health care plans.

While many issues were addressed during this phase of cognitive testing, in this paper, I will address four specific methodological issues:

1. Sampling events - Which health care experiences should we ask respondents to describe or report.

2. The problem of dealing with questions that do not apply.

3. What to ask about.

4. Ratings versus reports.

   4a. Numerical versus adjectival rating tasks.

   4b. How to ask report questions.

Sampling Time and Events

Because people have belonged to their health plans for varying lengths of time, it is important to standardize the period about which they are reporting to make results comparable. When thinking about how to ask about interactions with medical care providers, three candidate approaches were: a) the last visit to a doctor, b) all the encounters with medical providers in the past six months, c) all the interactions with providers in the last year.

From a question design point of view, asking about a single interaction (as the last encounter would do) is quite appealing. It avoids the complexity of asking people to summarize across experiences that may vary. However, when we tested questions about the last encounter, it became apparent that it was a very unsatisfying way of capturing patient experiences. Many times, respondents found that the last visit was inconsequential or atypical. On the one hand, respondents found it frustrating to answer questions about an interaction that they did not consider to be representative. In turn, as researchers, it became apparent that by restricting reporting in that way, we were not efficiently using survey time to capture what respondents had to tell us.

Based on these experiences, we turned to asking people to describe their experiences over a specific time period. There were many appeals to asking about one year: there is a certain roundness to one-year data, and more people see doctors in a year than in 6 months, so one-year reporting yields more data. However, when we thought about surveying plan members, a six-month reporting period meant that it was not possible to get meaningful responses from people who had switched plans during the most recent enrollment period. Going to a one-year reference period meant that it would extend over the change in enrollment point for switchers, thereby compromising the comparability of their answers. There also was some advantage to a six-month as a recall

period because it is shorter. A downside of this decision was that respondents found it hard to restrict their answers only to a six-month period, when they had significant interactions and experiences with plans that were somewhat or slightly before the reference period. Nonetheless, if one is willing to make the assumption that such pressures are consistent across plans, a six-month time frame seemed the best way to get standardized reporting and include new members.

When Questions Do Not Apply

One of the important realizations that came from cognitive testing was that many questions one might want to ask patients about their experiences with health plans and medical providers do not apply to everyone, particularly when questions focus on a time period such as six months. Obviously, if respondents had not seen a medical provider in the last six months, questions about how they were treated do not apply. In addition, even for those who have seen a medical provider, there often is a challenge in finding those patients to whom particular questions really apply. For example, we wanted to ask about participation in medical decisions. Whether or not there have actually been any medical decisions can be ambiguous, yet it is necessary to find people for whom it is appropriate to ask how decisions have been made. In the same way, if one wants to know about how payments have been handled, access to specialists, or problems with getting needed tests or treatments, one has to give thought to how to define the group to whom these questions truly apply.

One sometimes sees survey instruments about health care plans that ask respondents to rate how decisions are made, access to specialists, and the like. If no effort is made to identify those people to whom the questions do not apply, the statistics can vary markedly solely because of differences in the numbers of patients who have actually had needs or experiences relevant to the question. Moreover, the answers are uninterpretable if those to whom the questions do not apply have not been identified.

We found people will answer questions who have not had relevant experience, based on inference from their other experiences with providers. Our cognitive testing made it quite clear that an explicit strategy for asking people whether or not they have had a relevant experience, prior to asking questions that may not apply, is essential to producing interpretable data.

What to Ask About

As noted previously, extensive focus group work was done to find out what people wanted to know from respondents. Among the topics that clearly made the list were the quality of interactions that people had with providers, with office staff, and with health care plans. Also experiences with access, and problems getting the care they thought they needed, fit the bill. However, two issues that made the top of most peoples' lists, technical quality and physician choice, were more problematic and became the focus of a distinctive amount of testing and evaluation.

With respect to technical competence, we did numerous cognitive interviews, whereby we asked respondents to rate the technical competence of their physicians, then asked them to explain the basis on which they made those ratings. Time after time, it was apparent that respondent ratings of the technical quality of their physicians had little to do with technical quality. The single most common justification for a positive rating of competence was that the physician spent enough time with patients to examine them thoroughly. In essence, length of time with a patient was a surrogate for technical competence. While spending enough time with a patient may be a good thing, it has nothing to do with technical competence. We concluded that very few patients have experiences with their physicians in a six-month period that would enable them to evaluate their technical competence. Moreover, there are real difficulties in believing that patients can make meaningful evaluations. In the end, we decided that an overall rating of the physician essentially captured what most respondents had to say about the competence of their physicians.

With respect to choice, while it might seem to be a good idea to ask patients about their sense of choice, in general the number of physicians to which patients had access, or their credentials, was unknown and irrelevant to respondents. The real measure of whether or not people had the choice they wanted, for most respondents, was whether or not they found a personal physician they liked, and whether or not they were able to find and go to specialists they wanted to go to. Thus, when we had answers to those questions (do you have a good doctor; can you get to specialists who are good), even though the word "choice" was not in the questions, we felt we had what respondents could tell us about whether or not the choices available in a plan met their needs.

Ratings Versus Reports

We also spent a good deal of time addressing the problem of how to ask people questions that would best capture what they had to say about their experiences. There were two essential candidates, which might be described as ratings and reports. For example:

1. How would you rate how promptly you were seen

when you went to doctors' offices -excellent, very good, good, fair, or poor?

2. In the last six months, have you had to wait in the waiting room more than fifteen minutes past the time of your appointment with a doctor or nurse?

I would consider the first question to be a rating; the second is more of a report.

In our testing, we found three different kinds of difficulties with ratings. First, when respondents had more than one encounter with providers, and the experiences were different, they had real trouble using ratings. Suppose one experience was excellent and the other was poor. Respondents had three choices: they could ignore the good experience and report the poor one; they could do the opposite; or they could report some average, such as "good," which described none of the experiences that they had. We found respondents dealt with this ambiguity differently, and none of the answers captured what the respondents had to say very well.

A second problem was that ratings often seemed to be an inappropriate response task. Waiting in a waiting room is a good example. Seeing a doctor on time did not seem to many respondents to be a "excellent" experience, even though it couldn't get any better. There were numerous questions about which we wanted information, where the rating task simply didn't fit what respondents had to say and how they felt about it.

A third downside of the rating approach was that some users of the statistics felt that ratings *per se* were not very informative. They worried about the standards that respondents were using. What if the respondents were the kind of people who didn't mind waiting in waiting rooms? Their "good" rating might be intolerable to others. A question which came closer to having people report their experiences, rather than evaluate them, seemed to provide information that would be more useful to those to whom it would be presented.

In fact, we ended up using a combination of reports and ratings. Overall ratings of providers and health plans enabled respondents to give us their own weighted evaluation of how things worked for them. At the same time, when it came to the details of their experience, sticking closer to reports seemed the most meaningful way for respondents to share their experiences with others.

For both types of question, ratings and reports, we also had to make choices about how to pose the questions. With respect to the ratings, the biggest issue we faced was whether to ask people to use a common adjectival rating scale, such as excellent to poor, or to have them use a numerical scale, such as 0 to 10. We did extensive testing with both approaches with a variety of populations.

We had been concerned that the numerical task might be less acceptable to people with low incomes and educations. However, in our testing with a Medicaid population, there was no evidence of aversion to the 0-10 task. Among most groups, when respondents were debriefed, they expressed a preference for the numerical rating scale.

The numerical rating scale also had some desirable psychometric properties. By giving respondents more options, it spread out the answers. Because people's ratings of their plans and providers tend to cluster at the positive end of the rating scales, having more categories increased the distribution of answers, thereby improving ability to distinguish between providers and plans.

The CAHPS project also was committed to creating an instrument that could be used in English and in other languages. Translating adjectives into other languages in an exact way is not possible. Using the numerical scale helped solve that problem.

Thus, after numerous tests with various populations, it was concluded that the rating task of choice was asking people to use a scale from 0 to 10.

When we were asking for reports, we ran into a different kind of problem. One of the easiest ways to design questions to measure people's experiences with their plans and providers was in the following form:

> In the last six months, was there ever a time when your doctor failed to explain things in a way you could understand.

A question in that form easily adapted itself to patients who had multiple experiences or only one. It was also a clear question. However, in our testing, we found two kinds of problems with these questions. First, from the respondent's point of view, it was very hard for many respondents to answer "yes" when their main sense was that the provider usually explained things pretty well. When they had only one bad experience, and several good ones, to have to be confronted with only two choices, and to have to use the negative category, was more than many respondents could do. We found there was unreliability between respondents based on their willingness to use the negative category. The other problem was that the survey instrument that evolved with this strategy had a very negative tone to it. In essence, it became a list of possible negative things that could happen to people with their plans and providers over a six month period. Many of the potential users objected to the notion that nothing good was ever being reported - only negative things.

In response to these concerns, a different kind of question emerged:

> In the last six months, how often did doctors and other medical providers explain things to you in a way you could understand - always, usually, sometimes, or never?

This question design solved the two problems that were raised above. First, it enabled people to say "usually" when things weren't perfect, and that felt comfortable and accurate to respondents. Second, it enabled the questions to be phrased in a positive way, which made respondents and other reviewers of the instrument feel as if we were giving respondents a chance to say positive things, as well as negative. However, most solutions come with a problem, and this is no exception. The question form essentially assumed multiple events. If there had been only one encounter, technically one could argue that only "always" or "never" are possible answers. In our testing, however, we found that all the answers for one visit respondents were not in those two categories; respondents would use the middle categories as well, even when they only saw a physician once.

How can we explain that? There are at least two explanations. First, conceptually the denominator for many questions is not the visit but rather the interactions with a provider. Hence, when asked how often things were explained clearly, one could say that each verbal exchange with a physician potentially could be in the denominator, even if all the exchanges occurred on a single visit to an office. In that context, "usually" or "sometimes" answers are not meaningless; they are quite appropriate.

Second, we found that for some characteristics of providers, such as explaining things or spending time with the patient, respondents whose history with the provider extended beyond six months drew on their general experience when they answered questions. They did not carefully limit themselves to events in the last six months. This is a very understandable tendency, and one which affected the answers in all question forms. We do not think it affected these questions more than others; it was just more obvious. Moreover, since most of the questions are going to be combined with others to create multi-item indices, the anomalies of having an occasional "usually" or "sometimes," when the denominator is only one, become blurred.

Conclusion

The above is only a small sampling of the most important issues that have been addressed in the CAHPS project. Probably no instrument has been subjected to more testing and evaluation then this one. Moreover, this is still a work in progress. Extensive testing is going on now, and revisions will no doubt emerge from that process.

Some of the outstanding health researchers in the country have been involved in the process, and those on the project have been humbled by the complexity of the task of designing a good instrument to capture what people have to say about their experiences with their providers and the health care plans. However, we believe that the experience derived from this work will not only produce an excellent CAHPS instrument but also should be helpful to others who want to achieve similar goals.