DISCUSSION: ISSUES IN THE REDESIGN OF THE NHIS QUESTIONNAIRE

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

The basic design for the National Health Interview Survey (NHIS) was established in the late 1950's. At that time, the world was a simpler place. In particular, most treatments for serious medical conditions took place in hospitals. Virtually all medical care was delivered by, or under the supervision of, a medical doctor.

In 1994, much more medical care is being delivered outside of hospitals. A good deal of hospital care, including a significant amount of surgery, does not involve an overnight stay. Increasingly nonphysicians, including physical therapists, visiting nurses, and psychologists, are providing important medical care. Moreover, it is now recognized that mental health problems, and their treatment, rival physical problems in their importance. Thus, one essential reason for redesign of the NHIS is to collect data that more accurately and completely captures important elements of medical care in the 1990's.

At the same time, this is an occasion for rethinking the way things are measured. Probably no research organization in the country has given more attention to survey methods and the quality of survey measurement than the National Center for Health Statistics (NCHS). There is a long history within NCHS of sponsoring methodological studies. Through those studies, ways in which some of the important questions in the National Health Interview Survey could be improved have been identified; significant problems with achieving some of the measurement goals of the NHIS have been found. Although there have been occasional changes in the design of the NHIS, they have been modest in nature, in part to maintain the continuity of the program. The fact that the entire NHIS is now being redesigned provides an opportunity to rethink a number of issues about what questions are asked, and how they are asked, that holds the potential to improve the overall quality of measurement in the redesigned National Health Interview Survey.

The papers by Pennell and Jay are in a long tradition of record-check studies. When such studies were first done in the 1960's, the general assumption was that data from medical records constituted a gold standard; to the extent that the survey reports differed from medical records, reporting error could be inferred.

We now know that there is a good deal of error in most medical records systems. In particular, when researchers attempt to derive information from records in a form that was not built in to the design of the record-keeping system, they are likely to run into problems. One common feature of the Pennell and Jay papers is that they recognize that survey reports and medical records both may be imperfect ways of measuring what researchers want to measure. Rather than assuming that discrepancies between data from the two sources reflect survey error, the papers tabulate the rate of agreement between the data from the two sources, then explore possible reasons in one or both sources for the discrepancies. The notion that we have to define what we want to measure, then find the best source for measuring accurately what we need to measure, is a central theme and part of the task of the redesign of the NHIS. Surveys are the very best way to measure certain things; there are other things that survey respondents are unable or unwilling to tell us, and alternatives must be found. This sorting process, thinking through the strengths and weaknesses of survey respondents to provide information, is an essential part of the task of the NHIS redesign.

The paper by Makuc et al illustrates a different approach to the search for a gold standard against which to assess error. The reporting of visits to physicians within a two-week period is used as the standard; against that, the reports of characteristics of the "last visit" within the past year are compared. Discrepancies are inferred to reflect error in reporting the details of the "last visit".

This again is a very useful paper, because there is some real utility in being able to learn useful things by asking people about their most recent visit to a physician. A minority of people, about 10 percent, account for about half of all visits to doctors. These people who visit doctors very frequently contribute disproportionately to those visits occurring within two weeks of an NHIS interview. Since only 15 percent of the population has a visit within the two weeks preceding an interview, it means that those people who see physicians infrequently have their medical care experiences rather poorly represented in the population of two-week visits; their experiences could be described better if meaningful information could be provided about the "last visit".

The analyses presented suggest that the two-week results cannot be replicated by aggregating the "last visit" data; the population of reported "last visits" seems to have some systematic omissions. I hope, because of the potential utility of last visits, that the issue will not be completely dropped yet. The results seem to be quite sensitive to weighting. The weight used was the number of doctor visits in a year, which we know has some error. It may be that a corrected weight would make the last visit data look better. There also may be certain kinds of errors, particularly getting the wrong "last visit" because emergency room visits or inoculations are not reported as the "last visit", that could be fixed. For example, there may be ways to design the questions to do a better job of avoiding such misses. Finally, the standard error of estimates in the paper are for the sample as a whole. The authors should explore the effect on standard errors for those subgroups of the population that see doctors less frequently; the gains for last visit questions, or some variation, may be greater for such subgroups. At least that possibility needs to be explored before these questions are dropped.

The fact that record data and survey reports do not correspond, and even the fact that two-week reporting of visits to doctors is more accurate than reporting over an entire year, does not constitute real news. Other studies have been done that produced similar results. The glue that holds the session together, and provides the key to the importance of the studies being done here, is provided in the paper by Blixt et al. Blixt reports results when interviews are tape-recorded and the behavior of interviewers and respondents are coded to identify indications of question problems. Such coding is an imperfect way to capture problems of comprehension or tasks that people cannot perform. However, it does identify many questions that pose problems for interviewers and respondents. Moreover, once researchers begin to look hard at questions that pose problems for respondents, and try to figure out what the source of the problem is, it leads researchers back to the problem of question objectives. Often, it is poorly defined concepts, poorly defined terms, and poorly thought-through question objectives that lead to question problems. When we look at the kinds of results that Blixt and his associates produce, it highlights the centrality of thinking through question objectives in any study of survey error.

The problem of the heterogeneity of the reality that was to be measured was quite apparent in the Pennell paper on the measurement of chronic conditions. Some conditions, such as diabetes, are identified primarily through clinical tests. In contrast, arthritis is primarily evident through symptoms that patients experience. Moreover, arthritis is poorly defined from a clinical perspective. For those conditions that are well defined and usually require a physician for diagnosis, and particularly for those that require ongoing treatment, it is reasonable to think that medical records may provide good measurement, and the correspondence between patient reports and medical records may be an indication of the quality of survey reporting. In contrast, for conditions such as arthritis, lower back pain, and benign prostate disease, to name a few, the presence and severity of the condition is measured primarily through patient report. Reliable clinical indicators do not exist. As a result, survey instruments that reliably and validly measure the symptoms that patients experience should be able to provide the best measurement of these conditions. Moreover, discrepancies are likely to be the result of faulty records.

The issue of how to measure the presence of conditions also raises the distinction between having patients name the condition accurately, which may be affected by the kind and quality of interactions patients have had with medical care providers, versus having patients report on the symptoms that they experience, that they know about, and the reporting of which does not depend on interactions with physicians.

Conceptual issues in what is to be measured are also apparent in both of the papers devoted to measurement of physician visits. It may be difficult to replicate some ideal conception of a visit to a physician either from patients' reports or from records. Ambiguities abound. If a patient visits a medical-care facility, sees a nurse practitioner, a physician, has an X-ray, and has laboratory tests, does that constitute a single episode of medical care, or is it four episodes of medical care? If everything happens on the same day, during the same visit to one building, the patient is likely to perceive the events as one visit. In contrast, it may yield four different billing events in a record system. Moreover, does it make sense if the patient comes back the next day for the X-ray to count that event as a second visit, when it would be one visit if all the services are received on the same day? However issues like this are handled, they become essentially arbitrary definitional decisions that are unlikely to be shared either by those designing record systems or by respondents reporting on their medical care experience. Thinking through carefully what it is we want to measure, and what people can tell us, is essential to solving this problem, as it is to all survey reporting

problems.

LeClere and Parsons' paper emphasizes the dominance of the question, what is trying to be measured and how the question is worded, over other sources of error in surveys. Repeatedly, as we study the effect of interviewers on data and the correlates of respondent characteristics with error in data, we find that these effects, while sometimes statistically significant, are dwarfed in importance by the variation between questions in the quality of data that are produced. Question selection, design, and evaluation is the key to reducing error in any survey, including the NHIS.

In conclusion, there are many choices to be made. These studies constitute only a small portion of the many methodological analyses being carried out at the moment that bear on the redesign of the National Health Interview Survey. Moreover, it is clear from listening to these papers that methodological studies, by themselves, do not make survey design decisions. Nonetheless, such studies form an essential information base on which to make design decisions. The National Center for Health Statistics has a long history of sponsoring methodological studies to evaluate its procedures and to inform users of the data it produces. These papers, and the related research, constitute an important further step in that tradition. Methodological research will not free the NCHS staff from having to make value judgements about what is important to measure, and how to make trade-offs among alternative values. However, these studies surely will make a major contribution to the quality of the ultimately redesigned National Health Interview Survey.