

# Survey Design and Methodology in the Health and Retirement Study and the Wisconsin Longitudinal Study

ROBERT M. HAUSER

ROBERT J. WILLIS

Large-scale data collection has become the kernel of the growth of knowledge in the social sciences. Nowhere is this more evident than in research in the demography of aging and the life course, where scientific progress has been stimulated and sustained by complementary longitudinal studies of aging populations. In this chapter, we review the history, organization, and design of two of these public resources, the Wisconsin Longitudinal Study (WLS) and the Health and Retirement Study (HRS). Recent innovations in each of these studies hold promise for major advances in knowledge about the demography, economics, sociology, and epidemiology of aging. The stories of the WLS and HRS highlight important clues about the creation of systems of continuing surveys to inform science and public policy.

## Sample design

There are many ways to collect information about populations, ranging from journalistic observation to full enumerations and from sparse data, like those in a census, to detailed histories of individuals, families, or communities. While demographers take probability sampling for granted, there are alternatives, some of which are chosen regularly and on a scale as large as or larger than the best-designed probability surveys. Depending on the availability of appropriate lists or sample frames, it may be more or less costly to establish a probability sample for longitudinal study.

Almost five decades ago, when a professor of education at the University of Wisconsin-Madison carried out a survey of high school seniors to assess the demand for postsecondary schooling, no one would have

imagined that it would create the foundation for a major study of the life course and aging. The WLS began in large part because that survey provided a complete, identifiable list of the population. In the early 1960s, sociologist William H. Sewell learned that the questionnaires from a 1957 survey of all Wisconsin high school seniors were stored in the basement of the University of Wisconsin administration building (Sewell et al. 2003). When he examined the instruments, he learned that each questionnaire contained the name of the student and the name and address of the parents of the student. From a list of those schedules, Sewell chose a simple one-third random sample of more than 10,000 persons for further study.<sup>1</sup> That sample was followed up successfully in 1964 and 1975—and expanded to include randomly selected siblings in that wave. Only in 1992 to 1993, when the graduates were 53 and 54 years old, was the content of the surveys expanded to shift the focus of the study from education, careers, and family to health and aging. At present, the fourth wave of the WLS is in the field, as the graduates reach retirement age, and the sample has again been expanded to include the spouses or widows of graduates and their siblings.

In contrast, the Health and Retirement Study was created in a deliberate effort to establish a longitudinal observatory for multi- and interdisciplinary studies of aging populations (Juster and Suzman 1995), but its design and coverage, also, have evolved across time. The HRS design began with a decision to represent the US population in 1992 at ages 51 to 61 (and their spouses of whatever age). Because only 16.6 percent of households contained people in this age range, a large screening survey was undertaken by the Survey Research Center at the University of Michigan. As opportunities arose, coverage was first increased at older ages by incorporation of participants in a companion study, Assets and Health Dynamics of the Oldest Old (AHEAD), which covered cohorts born in 1890 to 1923, who were 70 and older in 1993 (Soldo et al. 1997). A decision was made to integrate the HRS and AHEAD questionnaires in 1995 and to combine the field periods in 1998. At that time, the HRS committed to a steady state design by adding two new cohorts, the “Children of the Depression” (born in 1924 to 1930) and the “War Babies” (born in 1942 to 1947). By 1998 the HRS became cross-sectionally representative of the US population older than 50, and a decision was made to maintain the steady state by adding a new six-year cohort every six years (Willis 1999).

Thus, in 2004, a fresh screening sample is being drawn to add a cohort of “Early Boomers” born in 1948 to 1953, who will then be 51 to 56 years old;<sup>2</sup> here the household eligibility rate is expected to be even lower, about 12.4 percent. This effort will generate a random sample of younger cohorts that will be used in the 2010 wave of the HRS, and those in older cohorts will be selected for another NIA-supported study, “Social Life, Health, and Illness at Older Ages,” conducted by Linda Waite and colleagues. Other older

participants in the screening sample may be asked to join a proposed project on measurement of cognition that will be used to “reengineer” HRS cognitive measures using state-of-the-art ability tests and psychometric methods. Finally, a small number of cases will be retained for use by the HRS as a “test bed” sample.

Whatever else may be problematic in the repertoire of aging surveys that are sponsored in whole or part by NIA, they all start with probability samples. This has great advantages. One can generalize to each of the populations that has been sampled, whether it is national or regional, or representative of the whole adult population or of specific age cohorts.<sup>3</sup> Moreover, if they are large enough, these population samples may be used as frames for the study of special populations. For example, ADAMS (Aging, Demographics, and Memory Study) has identified for intensive study a stratified random sample of about 850 persons from the Health and Retirement Study, based on survey measures of cognitive impairment. These individuals have had in-home visits by a trained nurse and psychometric technician to obtain clinically valid measures of dementia.

Similarly, Marsha Seltzer and Jan Greenberg’s component of the Wisconsin Longitudinal Study is identifying several hundred participants (high school graduates or their siblings) for intensive study if one or more of their children have a developmental disability or severe mental illness. In each of these cases, the base samples are large enough to generate more cases for intensive study than one often finds in clinical research, but the subpopulations are also representative of all instances meeting the definition of the special subpopulation. Another example of this is simply that the baseline samples, both in the HRS and the WLS, are of populations large enough to yield a substantial sample, many years later, of participants at advanced ages.

Longitudinal observation of the same individuals is necessary but not sufficient to provide a useful base for analyses of change in the process of aging. Thus, while studies of unique cohorts, like the WLS graduates, are of great scientific value, they do not provide a sound basis for studies of inter-period change, nor do they permit accumulation of age-specific data for scientifically significant subpopulations in different cohorts.<sup>4</sup> Moreover, unique cohort studies are likely to become obsolete as the world changes and as policy issues and scientific methods evolve.

An important advantage of the repeated longitudinal cohort design is the ability to cumulate rare cases across cohorts. For example, in order to obtain an adequate sample size, the most impaired stratum of ADAMS participants has been recruited from both the 2000 and 2002 rounds of the HRS.

To be sure, repeated longitudinal studies are complemented by other research designs. Repeated cross-section surveys, including the Census, Current Population Survey, Survey of Consumer Finances, National Health Interview Survey, Medical Expenditure Survey, and the National Health and Nutrition Examination Survey, provide important trend assessments with

large numbers of cases and without the large added costs of longitudinal coverage. Also, despite and perhaps because of their distinctly nonrepresentative design, Snowdon's studies of aging nuns (Snowdon et al. 1996; Snowdon 2001) have made uniquely important scientific contributions to our understanding of the precursors of longevity and of Alzheimer's disease. Likewise, Robert Fogel's studies of Union Army veterans are providing unique insights into the epidemiological transition. In an earlier period, the NBER-Thorndike data—from a study of men who volunteered to become pilots, navigators, and bombardiers in the Army Air Force in World War II—provided the first major longitudinal data on the relationship between early-ability test scores and later educational and economic success (Thorndike and Hagen 1959; Taubman and Wales 1973) and enabled researchers to address issues such as lifetime inequality in human wealth (Lillard 1977) and the relationship between ability and self-selection in educational choice (Willis and Rosen 1979).

### **Creating surveys: Serendipity and institutional support**

What has led to continuing, large-scale data creation? In our opinion, the successes are marked by the combination of policy relevance and scientific design and control. This combination is not easily achieved. Aside from the constitutionally mandated US Census, there is no institutional structure to create and maintain a complementary system of demographic, social, economic, or health surveys. Indeed, there is no centralized statistical system in the United States, such as exists in Canada and some other countries. Rather, America's national statistical "system" is an agglomeration of activities that are housed in line agencies with responsibilities to execute policies, such as the Department of Labor, combined with the ongoing activities of scientific agencies, interest groups, and individual investigators. The decentralization of statistical activities has both disadvantages and advantages.

Many key statistical series exist through historical accident. The Current Population Survey, which now serves many purposes and audiences, was created for the sole purpose of measuring unemployment. Sharing of samples and data across statistical agencies has been severely restricted by laws that have only recently been modernized. For many decades, agency interests and mandates have reduced or precluded the use of exceptionally valuable samples. Surveys that fall under the purview of the Office of Management and Budget (OMB), because they are conducted by federal agencies or under federal contract, are more restricted in content, design, and procedure than surveys conducted with grant support by nonfederal agencies.

Despite problems like these, the decentralized US research and statistical systems have yielded excellent and innovative data for research on

aging. First, the diversity of academic institutions in the United States has allowed for innovation and for open competition among research agendas and designs. Second, federal support for investigator-initiated research, notably at the National Institutes of Health and the National Science Foundation, but also at the Social Security Administration, the Department of Education, the Department of Labor, and the Department of Health and Human Services, has permitted innovative and sustained large-scale data collection and analysis. Third, unlike countries in which a central statistical agency has a monopoly on data collection, the United States has developed numerous independent survey organizations, including several that are capable of large-scale operations.

Chance, if not sheer luck, has played an important part in the development of current data resources on aging. As we mentioned earlier, in the 1950s and 1960s no one anticipated that the Wisconsin Longitudinal Study would evolve into a major longitudinal survey of health and aging. In retrospect, some of its design elements seem more valuable now than at the time they were introduced. These include the baseline measures of cognition and academic performance (from 1957) and the introduction of parallel surveys of sisters and brothers of the graduates (in 1977). However, sponsorship of the WLS shifted across time as the sample aged, and there were low points when the project might easily have been abandoned. From the inception of the 1964 follow-up through the mid-1970s, most support came from the National Institute of Mental Health, but a grant from the Spencer Foundation facilitated creation of the sibling sample. The 1990s brought initial sponsorship from the National Institute on Aging—again supplemented by the Spencer Foundation—which facilitated the shift to aging as the focus of the study and supported the latest two waves of data collection.

In contrast, the institutional framework for the HRS represents an innovative approach for the design and sustained support of a scientific and policy-relevant data collection effort (Juster and Suzman 1995). The HRS is supported through a cooperative agreement between NIA and the University of Michigan that provides extraordinary resources to allow the involvement of a broad range of scientists in the design and conduct of the study. For example, NIA provided over \$1 million and 18 months of time to prepare the baseline questionnaire and sample design for the original HRS baseline survey fielded in 1992. An important part of the motivation for the HRS arose from concern about the implications of the aging of the US population for the health and economic well-being of Americans during the latter part of life and the costs that will be borne by younger portions of the population to support the elderly through public programs such as Social Security, Medicare, and Medicaid. Planning for the HRS took place with a set of working groups with expertise in economics, demography, sociology, psychology, medicine, epidemiology, health services, and survey methodology. The experts

developed measures, usually by truncating or streamlining state-of-the-art measures from their fields. These working groups were overseen by an interdisciplinary group of co-investigators and advisory committees who made decisions on priorities across topical areas. These decisions were based on the potential of a given measure to contribute to the goals and theoretical framework of the HRS rather than their contribution to particular disciplinary concerns. The HRS has continued to obtain expert advice from diverse specialties as cohorts have been added and as new scientific and policy issues have arisen. For example, during 2002 some one dozen review papers were commissioned by the HRS Data Monitoring Board to assess the strengths and weaknesses of the HRS in a variety of research areas and to suggest ways of improving the quality and coverage of the survey.

Sustained support for a single data collection activity, however well designed, is not sufficient to inform public policy and the scientific enterprise. Thus, we need multiple data collection activities: to cross-validate findings, to exploit unique analytic opportunities, to support competing explanatory frameworks, and to expand content without overburdening research participants. We should prefer to see substantial overlap, comparability, and complementarities in content, along with collaboration among the users of multiple survey vehicles.

For example, in the 1992 round of the WLS, many of the economic measures were drawn from those used in the first round of the HRS. The same measures of psychological well-being have deliberately been used in the WLS, the National Survey of Families and Households (NSFH), and the survey of Midlife in the United States (MIDUS). The HRS has been employed as a model for the development of many other surveys, including the English Longitudinal Study of Ageing (ELSA), the Mexican Health and Aging Study (MHAS), and the Survey of Health, Ageing and Retirement (SHARE) in nine European countries. Similar efforts are now under discussion in Canada, Australia, and Israel. The scientific and policy rationale for developing comparable, multinational longitudinal studies of aging has been well argued in the US National Research Council's (2001) report, *Preparing for an Aging World*.

## Public responsibility

Along with innovation and entrepreneurship, and beyond the relevance of research to policy, public responsibility also plays a growing part in large-scale survey research. Surveys are successful only when members of the population are willing to join with funding organizations and researchers in adding to the society's knowledge. The WLS and HRS, along with other long-term studies, bear a special burden in this regard, not only because of the large public subsidies required to sustain them, but because the scien-

tific uses of the studies should warrant the periodic time and effort of research participants and the risks, however small, of harm to them.

Why do people become research participants? Individual motives are surely heterogeneous, and research participants have no collective voice except through the political process (Groves and Couper 1998; Hill and Willis 2001). Researchers seek to behave as if participants were behaving altruistically, by informing them of the purposes and importance of each study, even though pre-survey monetary incentives are increasingly common and are sometimes effective in securing participation (Rodgers 2002).

In the case of the WLS, the recruitment appeal to research participants goes beyond altruism and seeks to take advantage of their identification with the state, its university, and their high school class. In the past, participants were never compensated. In the 2004 round of the survey, following the completion of a 75-minute telephone interview, each respondent receives a self-administered questionnaire. Two crisp five dollar bills are attached to the front cover of the instrument with a red paper clip, but many participants have returned the money. At present, among 85 percent of participants who complete the telephone interview, 89 percent also complete the 50-page self-administered questionnaire.

Even in good times, research support is a scarce and largely public good, and respondents' time, effort, and willingness to risk disclosure of private information are of critical importance. Thus, the WLS and the HRS attempt to minimize respondent burden and risk in several ways.

First, despite some risk of disclosure, data-sharing is becoming the rule. If research participation is a public contribution, then data so obtained should be a public good. Given the growing emphasis on protecting data in secure enclaves, through licensing agreements with researchers, and with other technologies for guaranteeing privacy, we think it is fair to say that the protection of research participants provides no rationale for the unwillingness of investigators to share primary data. We believe that researchers have a responsibility, not merely to share data, but to encourage their use by the research community. Thus, rather than relying on traditional data archives for dissemination, the WLS and HRS projects have created websites designed to inform potential data uses, to register active users, and to permit downloading of data and documentation: «<http://dpls.dacc.wisc.edu/wls/> (WLS)» and «<http://hrsonline.isr.umich.edu/> (HRS)».

Second, while linking data from multiple sources potentially increases disclosure risks, it also increases the value of existing observations without increasing respondent burden. For example, the HRS has linked its survey data files to administrative records from the Social Security Administration, from Medicare, from employer pension plans, and to the National Death Index, while the WLS data include some of the same administrative record data. These make both sets of data much more valuable to a community of

users that extends well beyond the primary investigators to include other researchers and the participating agencies. However, from the perspective of the agencies as well as the research participants, possible violations of privacy and confidentiality loom large as threats to their partnership in the research enterprise. It is thus highly significant that the Social Security Administration is contributing additional resources to the HRS in its 2004 wave for personal household interviews that will improve the chances of obtaining informed consent for links from SSA records to the HRS.

Third, one common type of data link, namely geocoding, deserves special mention because it presents serious risks of re-identification of research participants, but it often does not require partnership with a third party. For example, detailed addresses are easily obtained and are sufficient to create links with many types of public data. Indeed, it was the potential identifiability of deaths of individuals for whom only the date of death and city of residence were known that led to the controversial Shelby Amendment, mandating disclosure of some research data through the Freedom of Information Act (National Academy of Sciences (US), Science, Technology, and Law Panel, National Academy of Sciences (US), Policy and Global Affairs, and National Research Council (US) 2002). Because of the threat of disclosure, we believe that researchers should provide access to geocoded data under the same restrictions as apply to linked administrative data for individuals.

With the risks of identifiability in mind, the HRS and the WLS have each created special licensing arrangements for exceptionally detailed data extracts and linked administrative data within secure data enclaves. These make it possible to analyze sensitive data statistically without direct access to individual and actually or potentially identifying information.

Fourth, although violations of privacy and confidentiality are the main source of risk and harm to research participants in the social, economic, and behavioral sciences (National Research Council, Panel on Institutional Review Boards, Surveys, and Social Science Research 2003), researchers also have a continuing responsibility to avoid other potential risks. This is an issue of growing importance as social surveys are increasingly combined with other forms of potentially sensitive data collection, for example of biomarkers (National Research Council, Committee on Population 2001).

### **Interdisciplinary innovations in survey methodology**

Aging is a process whose study is not the property of any given discipline. Gerontology must inherently be interdisciplinary. Aging also, by definition, proceeds through biological, demographic, social, and economic stages of the life course. As described earlier, the planning and oversight of the HRS is one

positive example of multidisciplinary collaboration in the study of aging, and the new international studies are following the same model. The WLS is also a case in point. That study was dominated by sociological interests in its first 35 years, but since the early 1990s its content and direction have shifted to incorporate psychological, economic, and biomedical concerns.

The importance of cross-disciplinary partnership is well illustrated by growing interest in the ubiquitous correlation between socioeconomic status and health. Traditional epidemiology sees this correlation as dominated by the effects of economic and social resources and stresses on health, while economists have tended to view health as a form of human capital (Grossman 1972) that in part reflects deliberate choice and in part is influenced by third factors such as time discounting (Farrell and Fuchs 1982). In the past, surveys with good health measures have had poor economic measures, and vice versa, and surveys with both measures have been cross-sectional. This has hampered serious investigation. The availability of longitudinal observations with high-quality data on both health and socioeconomic status has stimulated analysis and discussion by leaders in several disciplines of the interactions between health and social and economic standing across the life course (Adams et al. 2003a, 2003b; Adda, Chandola, and Marmot 2003; Florens 2003; Geweke 2003; Granger 2003; Hausman 2003; Heckman 2003; Hoover 2003; Mealli and Rubin 2003; Poterba 2003; Robins 2003). This collaboration is discussed in more detail by James Smith in this volume.

Some of the best examples of interdisciplinary progress have come in the area of measurement. In the following paragraphs, we discuss several instances in which methods, variables, or motivating ideas cross the boundaries of economics, sociology, psychology, and medicine. This is not merely a matter of bringing "the state of the art" to bear on specific topics. Rather, in the context of large-scale data collection, the problem is often to modify specialized disciplinary measures in order to obtain valid measures within the knowledge base of respondents, without excessive intrusion, and within the time constraints of a multipurpose survey.

### Sources of nonresponse

The knowledge and information-processing capacity of respondents plays a dual role in studies of aging. On the one hand, it is an object of study because it is a major determinant of health and decisionmaking behavior. On the other hand, it influences the quality and type of information that may be obtained in a survey. The latter effects range from overall survey response to specific patterns of response error, such as item nonresponse, age-heaping, and systematic mismarking of questionnaires.

For example, in the Wisconsin Longitudinal Study, there were essentially no systematic ability gradients in refusals from 1957 through 1975.

Beginning with the 1992 telephone survey, there was a modest gradient in refusal, such that 93.2 percent of persons in the top 10th of adolescent IQ responded, while only 77.7 percent of persons in the bottom 10th of IQ responded. This gradient was equally steep in the 1992 self-administered mail questionnaire, which ranged from 20 to 24 pages in length. Among graduates who completed the telephone interview, 89.8 percent of persons in the top 10th of IQ responded, compared with 68.0 percent of persons in the bottom 10th. We speculate that less able respondents may have failed to understand the purposes of the study and found it difficult and tiresome to read and respond to a long series of questions.<sup>5</sup> Adolescent ability also affected item response patterns. For example, in the case of Carol Ryff's measures of psychological well-being (Ryff 1989; Ryff and Keyes 1995), 11.8 percent of individuals in the bottom 10th of IQ used the same response category more than half of the time in a series of 42 items, as compared to 3.2 percent of individuals in the top 10th of IQ.<sup>6</sup>

The role of cognitive functioning in survey measurement is nicely illustrated by progress that has been made in the measurement of income and wealth. Traditionally, economic surveys have asked about the ownership and value of assets in a separate part of the questionnaire from questions about the flow of income from assets. Typically, these methods have yielded dramatic underreporting of income from assets, relative to estimates derived from administrative data sources. Beginning with the AHEAD cohort in 1995, this discrepancy was eliminated in the HRS in a very simple way: For each type of asset, the questionnaire asked about ownership, value, and income in that sequence (Juster and Suzman 1995).

### Measuring economic quantities

A more general issue in the measurement of income and wealth arises from the problem of item nonresponse in conjunction with the fact that the most widely used measures of income and wealth are equal to sums of individual measures. For example, wealth is the sum of the value of checking accounts, house value, stock ownership, 401(k) balances, and so on. In most surveys, for example, many individuals answer "don't know" or "refuse" in response to at least part of the sequence of questions about the value of individual wealth components. In order to compute the desired aggregate measure of wealth, the analyst must impute some value to the missing items.

Traditionally, surveys have used hot-deck (borrowing values from similar cases) or regression methods for imputation, using observed individual and household characteristics to capture variation in the value of each component. However, wealth is distributed quite unequally across households, with a long right tail, and observed characteristics account for minuscule proportions of the variance of wealth. Moreover, the missing data are not

missing at random, but in rather complex ways. For instance, missingness may be negatively correlated with some components of wealth—since persons who fail to report the value of wealth may be those with poor cognition—while it is positively correlated with other components—because high wealth persons are particularly sensitive to privacy concerns.

The HRS has made a major advance, both in reducing the proportion of item nonresponse and in improving the quality of the imputations for non-responses by employing bracketing techniques (Juster and Suzman 1995). The technique itself is very simple. After getting a response of “don’t know” or “refuse” in response to a question about, for example, the amount of money in a checking, savings, or money market account, most surveys simply record a missing value. Instead, in the 1992 wave, the HRS interviewer immediately asked whether the value of the account is more than a given bracket amount, say \$5000. If the person answers “yes,” he or she is asked whether it would be \$10,000 or more; if “no,” whether it would be \$1,000 or more. If the answer to the \$10,000 or more bracket is “yes,” a final question is asked about whether the account has \$50,000 or more; no further question is asked if there is a “no” answer to \$1,000 or more.

The effects of using the bracketing technique are striking. Juster and Suzman (1995) report that about 90 percent of people who say “don’t know” are willing to respond to the bracket questions, while about 50 percent who “refuse” to give a continuous amount such as \$2,034 are willing to give a bracket amount. This results in a reduction of the item missing response rate from double-digit to low single-digit rates. Moreover, the brackets are extremely informative. For example, the checking and savings account value for a participant who completes the bracketing sequence described above is determined to be in one of the following intervals: 0–\$999, \$1,000–\$4,999, \$5,000–\$9,999, \$10,000–\$49,999, or \$50,000 or more. Within each interval, the HRS imputes a continuous value, using a hot-deck technique based on the distribution of continuous responses within the interval. Imputed values based on brackets typically explain about 80 percent of the variance (in logs) of continuous reports, as compared to only about 20 percent of the variance for an imputation based on observed characteristics of the respondent (Hurd 1999). In addition, Juster and Suzman (1995) argue that an unbiased estimate of asset values cannot be obtained from respondent characteristics because those who respond “don’t know” or “refuse” are often a highly select subsample of participants with given observed characteristics. For example, Juster and Smith (1997) find that item nonresponse for asset questions is reduced by 75 percent and the estimate of housing wealth is increased by 18 percent through the use of unfolding brackets in the HRS. The success of the bracketing technique in improving asset information in 1992 led the HRS to apply brackets to almost all income, wealth, and other quantitative questions in subsequent waves, with bracket values optimized

to maximize their explanatory power (Heeringa, Hill, and Howell 1995). In its current round of surveys, the WLS has followed the lead of the HRS in bracketing economic amounts, with similar success in reducing nonresponse. Also, the WLS has extended the HRS procedure by conditioning bracket points on certain characteristics of participants, for example, by sex in the case of current earnings from employment.

Why does the bracket technique reduce nonresponse to income and asset questions? There appear to be two main answers to this question, both psychological in nature. First, people who respond “don’t know” may be uncertain about the value of an asset at the level of accuracy they presume the interviewer desires. The use of brackets signals that an approximate answer is all that is needed. Second, people who refuse to answer may do so because they regard the details of their economic situation to be personal information that they do not wish to disclose to a stranger. In addition, despite assurances of confidentiality, they may worry that this information might fall into the hands of the government or others who could use it to their disadvantage. By allowing them to give rough approximations, the bracketing technique appears to reassure a significant number of participants who have these concerns.

Psychological and other determinants of survey response increasingly command the attention of economists, sociologists, psychologists, and health researchers. Cognitive capacity, personality, physical abilities, motivation, and other factors influence the willingness of individuals to participate in surveys, their willingness to answer any given question, and the quality of the information they provide in their answers. Survey designers attempt to minimize survey and item nonresponse and to ask questions in a way that will elicit “true” answers. One should also recognize that many of the factors that influence the quality of an individual’s survey responses may also influence his or her behavior in general. Because of this, researchers need to develop theories of survey response that provide a link between observed responses and the underlying true values. In addition, it may be useful to model an individual’s behavior in the real world and his or her behavior as a survey respondent simultaneously.

We illustrate these points with several examples, beginning with further analysis of the quality of data in the HRS generated by bracketing. Soon after introducing brackets and despite their apparent success in reducing nonresponse and overcoming bias, the HRS investigators became concerned about the potential influence of bracketing due to anchoring and acquiescence bias (Hurd 1999). As discussed above, the bracketing technique is successful in obtaining information from individuals who are uncertain about the true value of a given asset or income amount. However, beginning with Tversky and Kahneman (1974), psychologists have established a robust finding of anchoring effects in questions of the form “Would it be greater or less than X?”

where “it” might be the length of the Amazon River and  $X$  is the anchor. Moreover, these effects have been shown to be largest for items about which a subject is most uncertain of the true value. In the context of the HRS, the issue is whether, for example, the entry point of \$5,000 in the bracketing sequence on the value of checking and savings accounts has an influence on the final category in which a respondent ends up as compared to a bracketing sequence that began with an entry point of \$1,000 or \$10,000. Similarly, psychologists have found that a one-sided bracketing question—“Would it be more than \$5,000?”—may lead a respondent to agree, causing an upward acquiescence bias in the HRS question format.

The HRS investigators began a set of experiments over several waves of HRS and AHEAD that were designed to determine the degree to which anchoring and acquiescence are of empirical importance. The history of this effort and findings are described in Hurd (1999). An informal Bayesian theory of anchoring bias might involve interaction between the precision of an individual’s knowledge about the quantity in question and the amount of information about this quantity that the person believes to be contained in the anchor,  $X$ . The individual’s knowledge is given by a subjective distribution that is concentrated about a single value if knowledge is precise but spread over a considerable interval if it is imprecise. In responding to a bracketing sequence, the individual chooses, perhaps unconsciously, a weighted average of the central tendency of his prior subjective distribution—say, the mean, the median, or the mode—and the anchor, where the weight is inversely related to the precision of his prior distribution and positively related to the perceived relevance of the anchor. For example, people usually have precise knowledge of their height and thus their answers would be unlikely to be influenced by whether a bracketing sequence began by asking whether they were “five feet or more” or “six feet or more.” However, a person might be uncertain about the market value of his house and might believe that the anchor conveyed some information about the value of homes in his neighborhood. In this case, the final bracket might be sensitive to the choice of entry point unless the entry point was very far from market values in the neighborhood and hence was viewed as not relevant. The propensity to select into a bracketing sequence by answering “don’t know” or “refuse” to a request for an exact answer is likely to be negatively correlated with precision of beliefs, and, conditional on selection, the anchoring bias is also likely to be negatively correlated with precision.

The importance of anchoring bias for income and assets in a survey like the HRS is an empirical question. It is likely to vary across particular items, to vary across people with different degrees of knowledge and cognitive ability, and to depend on the distance between the anchor and the true value of the item in question. Hurd (1999) describes a number of experiments, both deliberate and inadvertent, in the HRS that enable measure-

ment of the empirical magnitude of anchoring and acquiescence effects. Qualitatively, his findings are consistent with the theoretical expectations described above. Quantitatively, the effects were sufficiently important that the HRS survey designers decided, after consultation with an expert working group, to randomize the entry points for all bracketing sequences and to replace the unbalanced format (“Would it be more than X?”) with the more verbose balanced format (“Would it amount to less than \$X, more than \$X, or what?”).

### **Expectations: Measuring subjective probabilities**

Another example of interrelationships between the behavior of individuals as survey participants and as actors in the real world is afforded by the innovative attempt in the HRS to measure subjective probability beliefs on a wide variety of topics. These questions depart from the conventional approach to expectations in economics. Specifically, as Dominitz and Manski (1999: 16) argue, “Economists have typically assumed that expectation formation is homogeneous, all persons condition their beliefs on the same variables and process their information in the same way.” Within the conventional approach, probability beliefs are treated as unobservable, and assumptions about beliefs, such as rational expectations, along with assumptions about unobservable preference parameters, such as risk aversion or time preference, are embedded in optimizing models from which testable relations between observable variables may be derived. An important motivation for asking directly about subjective probabilities is to relax the assumption of homogeneous expectations by converting probabilities from an unobservable to an observable quantity that may vary across respondents and thus capture individual heterogeneity in expectations.

The probability questions in the HRS follow an approach pioneered by Juster (1966) and Manski (1990) by asking a participant directly about the likelihood of various events by giving the interviewer a number from 0 to 100, where “0” means no chance and “100” means that the event is sure to happen. The questions cover a variety of topics that can be usefully classified into three types: (1) general events such as the chance of a severe economic depression in the next ten years, the chance that social security will become more generous, or the chance that the value of a mutual fund held in stocks will be higher at this time next year; (2) events with personal information such as the chance of surviving to age 75 or the chance that one’s income will increase faster than inflation over the next ten years; and (3) events under personal control such as the chance of working past age 62 or 65 or the chance of leaving an inheritance of a given size. As this classification implies, subjective probabilities may vary across people at a moment in time or over time for a given person because of external exog-

enous factors such as prices or aggregate income; because of personal knowledge of idiosyncratic factors such as health or promotion possibilities at work; because of internal psychological factors such as optimism or pessimism; or because of choices about controlled stochastic processes such as work or saving that reflect the interaction of preferences and constraints that may depend on the realization of random variables such as a spouse's survival, an employer's offer of an early retirement window, and so on.

The first analyses of the HRS probability questions focused on whether the answers made sense. Striking evidence for the value of these questions came in early analysis of the survival probabilities by Hurd and McGarry (1995), who showed that, on average, there is surprising agreement between these subjective reports and life table estimates of survival, including covariation with health status and health behaviors such as smoking.

Although the subjective probability responses in HRS seem to "work well" when averaged across respondents, individual responses appear to contain considerable noise and are often heaped on "focal values" of "0," "50," and "100." The existence of focal answers raises questions about what aspects of the probability beliefs of respondents are revealed by their answers to survey questions. Some psychologists, especially Fischhoff, de Bruin, and their colleagues, have argued that focal answers at "50" often reflect "epistemic uncertainty"; that is, a failure to have any probability belief at all about the event in question or, at least, to have no clear idea of what the probability could be (Fischhoff and de Bruin 1999; de Bruin et al. 2000). Alternatively, of course, an answer of "50" might reflect a very precise belief about the probability that a fair coin will come up heads or perhaps a less precise belief that a given event is about equally likely to occur or not occur.

There has been much less emphasis in the psychological literature on focal answers at "0" or "100." When a probability question concerns a general event such as the HRS question "What is the percent chance that mutual fund shares invested in blue chip stocks like those in the Dow Jones Industrial Average will be worth more this time next year than they are today?," it does not seem credible to assume that a respondent who gives a particular answer, such as 65 percent, is completely certain that such an event will or will not take place. Such an interpretation is more plausible for questions about events with personal control such as, "What do you think the chances are that you will be working full-time after you reach age 62?" Answers at "0" or "100" may simply indicate that the respondent feels that he has already made a decision. While this may be a natural interpretation, it is inconsistent with dynamic economic models in which labor supply decisions are assumed to be conditioned on variables—such as income, wealth, pension incentives, the health status of oneself or one's spouse, and the employer's demand for labor—whose future values cannot be known with certainty unless, perhaps, the target date in the question is near.

Lillard and Willis (2001) propose a model of survey responses to subjective probability questions that is broadly consistent both with the existence of focal answers and with their close correspondence to objective outcomes when averaged across respondents. They assume that probability beliefs consist of a subjective prior distribution which is highly concentrated for questions such as the probability that a fair coin will come up heads and highly dispersed for questions with great deal of uncertainty. HRS participants take about 15 seconds to answer each probability question. Lillard and Willis hypothesize that respondents report the modal value of the subjective prior distribution; that is, they report that value which they believe to be the most likely among all possible probability values. They argue that the modal response is cognitively less burdensome than alternatives such as the mean or the median of the subjective prior distribution. When the degree of uncertainty is relatively low, this answer mode provides a very good estimate of the mean. Thus, reports of the mode might be construed as a “fast and frugal” algorithm of the sort that psychologists such as Gigerenzer et al. (1999) suggest often help people make complex judgments in situations in which a rapid response is called for. As uncertainty increases, Lillard and Willis show that this algorithm generates responses at “0,” “50,” and “100” as the median and mean of the subjective prior distribution increase from low to high values. Because of this, the average value of probability reports across people tends to track objective outcomes if people’s subjective beliefs are not biased in an optimistic or pessimistic direction.

Researchers have found the HRS probability measures useful in understanding behavior. For example, Lillard and Willis (2001) argued that the propensity to heap on focal values in answering probability questions provides an indicator of the degree of imprecision of probabilistic beliefs and, further, that economic theory implies persons with more imprecise beliefs will be more risk-averse. They found evidence for this in an analysis of savings and portfolio choice. Still another use of the probability data is a book-length analysis, *The Smoking Puzzle* (Sloan, Smith, and Taylor 2003), which examines how survival expectations are affected by smoking, how smokers and nonsmokers revise their expectations as they experience health shocks, and what role expectations play in decisions to quit smoking. As a final example, Chan and Stevens (2001) study incentive effects of pensions on retirement probabilities and on actual retirement behavior.

### Comparing subjective assessments

Recent methodological research in connection with the World Health Survey (Salomon, Tandon, and Murray 2001; Sadana et al. 2001; King et al. 2004) addresses a fundamental problem in survey measurement: whether all respondents use the same numeric scale in the same way when re-

sponding to questions that call for a graded response. For example, to measure negative affect, people in highly developed and less developed countries are asked, "Overall, in the last 30 days, how much difficulty did you have with feeling sad, low, or depressed?" and are given the response alternatives, "none, mild, moderate, severe, or extreme." How can we know, even after addressing language differences with careful back-translation, whether the several response categories are used similarly by individuals in such widely different circumstances? The phenomenon is essentially the same as that of differential item functioning (*dif*) in psychometric measurement, except there is no right answer to each item for all individuals.

In this context, the WHS group suggests that carefully drafted vignettes be used to calibrate self-assessments across populations. For example, each respondent is asked the self-assessment item in the domain of affect, followed by one or more vignettes of the form: "Imagine that the people described below are the same age that you are. Using the same scale that you used when talking about aspects of your own health, how would you rate the health of these people? ... Barbara feels depressed most of the time. She weeps frequently and feels hopeless about the future. She feels that she has become a burden on others and that she would be better off dead. How much of a problem does Barbara have with feeling sad, low, or depressed? (none, mild, moderate, severe, or extreme)." By making the crucial assumption that respondents use the categories in the same way in rating their own health as in rating the vignettes, it becomes possible to calibrate differences in the self-rating scale across populations.

Typically, calibration is improved by using more than one self-rating scale in each health domain and by administering several different vignettes for each health domain with corresponding rating scales. However, it is not essential that every vignette be administered to every respondent. Thus, although the vignette method requires more survey time or space than a simple rating scale, it can be made more efficient by random presentation of a subset of vignettes. In addition to other standard health measures, the WLS is administering vignette-based measures of affect and mobility in four alternate forms of a sex-specific self-administered mail instrument.

### Recording full-text interviews

In developing the 2003–05 round of the Wisconsin Longitudinal Study, project staff decided to record all of the telephone interviews. The original reason for our investigation of recording technology was that two of the collaborators, Nora Cate Schaeffer and Douglas Maynard, wanted to obtain high-quality recordings of about 1,000 randomly selected interviews that could be used for intensive analysis of respondent–interviewer interaction in an older population. A second reason, which applied to parts of all inter-

views, was that some of the more promising protocols for cognitive assessment could not be administered reliably unless the responses were recorded, and, furthermore, recordings could be used to validate appropriate administration of the assessments.

Staff learned that almost all respondents would agree to recordings and to their retention for research purposes. This was confirmed in pretests, and a consent protocol was developed. The WLS recording technology was developed to meet four main criteria:

(1) Recordings should be digital and stored as WAV or MP3 files, thereby permitting random access for research purposes. Thus, standard or digital audio taping was eliminated as a possibility.

(2) Control of the recording process should take place automatically through the Computer Assisted Telephone Interviewing software and not require separate activation or adjustment by interviewers.

(3) The recordings should be as high in audio quality as possible, given the limited bandwidth of telephone lines.

(4) Because audio files are potentially identifiable, as well as valuable for research purposes—including at some point in the future—file storage procedures should be both secure and redundant.

The cost of the recording software and equipment is modest, less than US\$250 per workstation.

Aside from the future value of the recordings for research, which will include an improved ability to edit the raw survey data, they have already proved useful in the process of instrument development in the WLS. For example, it has been efficient for researchers to listen to each instance of a pretest telephone module—for example, a family roster or employment history—in order to detect and solve problems in the logic and content of the instrument and to identify problems in interviewing that can be addressed in training sessions.

## Domains and units of observation

While survey respondents are individuals, the units of observation and analysis are by no means limited to individuals. For example, the Wisconsin Longitudinal Study actually samples persons within households and their siblings, while the Health and Retirement Study samples persons in specific cohorts and their spouses, regardless of age. In the WLS, observations are also linked to high schools and geographic location in each wave of the survey. In fact, almost all household-based survey observations may be linked to geography at some level of detail in private and sometimes publicly available files. In the HRS, individuals are linked to pension plans through their employers, and in the Medicare data it is possible to link individual records to

characteristics of hospitals or physicians providing treatment. Analyses based on the complex record structures resulting from such links have been facilitated by the development of new statistical software for hierarchical data.

The possibilities for useful data links continue to grow. We have already mentioned health and Medicare records, pensions, and Social Security earnings. The HRS is exploring the feasibility of links of individual survey data to characteristics of employers with the Census Research Data Center. This would map characteristics of the firm—for example, its wage policies—down to the level of the individual HRS participant, and it would also map other links, such as employer pension policies, at the level of firms. For younger cohorts, it would seem feasible to match individuals to birth records that include birth weight. For example, in Wisconsin, birth weight was not included in the birth certificates of 1939, when almost all members of the graduate cohort were born. However, for a few years following World War II, birth weight was stated on the public portion of the birth certificate. The WLS is experimenting with a self-reported birth weight item in its 2003 mail questionnaire, and for siblings of the graduates who were born during the right period, the self-reports will be validated against birth certificates.<sup>7</sup> In the WLS, it has already been possible to collect data on high school resources from archives in the Wisconsin State Historical Library. These are available for school districts, not individual schools, but in most of the state in the 1950s, each school district had only a single high school.

Another link in the WLS, whose construction is now in progress, is to data from high school yearbooks. The project has succeeded in borrowing and scanning the 1957 yearbooks from schools attended by about two-thirds of the graduates. Individual pictures of graduates have been extracted and used to identify African Americans in the cohort. Later, the full set of pictures will be graded for facial characteristics, such as attractiveness, obesity, and the Duchenne smile (symptomatic of muscular dystrophy), that may be associated with economic, health, and social outcomes (Hayes and Ross 1986; Hamermesh and Biddle 1994; Averett and Korenman 1996; Biddle and Hamermesh 1998; Smith 1999; Harker and Keltner 2001). In addition, although the baseline WLS survey of 1957 did not ascertain students' extracurricular activities, these data are now being coded from reports in the yearbooks.

### **Leadership, flexibility, and serendipity**

In our discussion of institutional support, we emphasized the important role that scientific agencies such as the National Institute on Aging, the National Institute of Mental Health, and the Social Security Administration have played in supporting the development and analysis of the HRS and the WLS. Serendipity shows its hand in this arena, too. A leading example is the pro-

found influence that Richard Suzman's arrival as a program officer at NIA in the early 1980s has had on the development of both of those surveys, among others. The idea for the HRS emerged out of a set of scientific meetings organized by Suzman early in his tenure at NIA, and the subsequent movement from idea to implementation required skill and persistence in getting both federal officials and a number of scientific constituencies to understand what was envisioned and to lend their resources to making the HRS a reality. This was not a one-off accomplishment. Suzman has been behind many other data collection initiatives, ranging from supporting international surveys such as the English Longitudinal Study of Ageing, Survey of Health, Ageing and Retirement in Europe, and Mexican Health and Aging Study, to supporting the addition of new kinds of measures, such as biomarkers, to social science surveys like the WLS (National Research Council, Committee on Population 2001). He has also been instrumental in enticing researchers of the highest caliber, from diverse disciplines and countries, many of whom were new to the field of aging, into research that exploits the data developed with the support of NIA. The effects of these investments on the growth of knowledge will be felt far into the future.

### Probability samples within probability samples

At first thought, one might expect that large-scale longitudinal surveys of probability samples of well-defined populations are useful only in providing superficial descriptions and models—providing “the big picture.” We think this is partly true, but large-scale observational studies also provide unusual opportunities for in-depth studies, and even experiments, in populations whose defining traits are rare. In our opinion, the major longitudinal studies should be designed to anticipate and encourage such uses of the samples, even if their exact content is unknowable at the outset. We have already mentioned the possibility of identifying, cumulating, and pooling rare observations across time, an idea that forms the basis for the ADAMS supplement to HRS, and the identification of atypical child outcomes in the WLS—developmental disability, severe mental illness, or early death.

Aside from providing a frame for intensive study of rare populations, large-scale longitudinal surveys may also provide opportunities for especially costly or intensive studies of the base population or, inversely, as ways of minimizing cost and respondent burden in segments of the base population that are not included. Current surveys provide many examples of this kind, all of which may be thought of as designs where data are deliberately missing at random, a notion that reminds us of recent advances in theory and methods for the analysis of data with missing observations (Arminger, Clogg, and Sobel 1995; Vermunt 1997; Little, Schnabel, and Baumert 2000; Little and Rubin 2002; Allison 2002).

One such example in the WLS is a plan to invite a random subsample of approximately 500 graduates to Madison for a two-day visit during which their health and psychological functioning will be assessed intensively. The protocol includes a physical examination, various psychological assessments, and functional magnetic resonance imaging (fMRI) of the brain. Richard Davidson has already established the feasibility of this project in examinations of a select subsample of about 100 WLS participants in the late 1990s (Jackson et al. 2003) and in repeated assessments of many of the same individuals in late 2002 and early 2003.

In the 2003–05 round of the WLS, there is simply too much content for all items to be administered to all eligible graduates or siblings in the telephone interviews. For that reason, the WLS has adopted a complex subsampling scheme, in which selected core items are administered to all participants, while other modules are administered to nested subsamples. This scheme is designed to provide larger numbers of observations for estimation of relationships of great interest, for example, health and access to health care, economic assets and economic transfers, and depression and alcohol use. The scheme will provide reasonable numbers of observations with complete data for analysts with typical interests and moderate statistical skill, along with richer, but incomplete data for more sophisticated analysts.

### Serendipity and social change

The contingencies of economic and social change, interacting with those of the life course, imply that no longitudinal survey design will ever anticipate all uses of a study—or guarantee that planned uses of data will pay off. Indeed, we are impressed that the best research based on a given survey often explores questions or tests theories that were not contemplated by the designers of the survey. What we can do is to give our studies enough strength to hold up across time and to keep our eyes open for new opportunities to create data and add to knowledge. Such opportunities may arise through secular or cyclical change, bringing external events to bear on the lives of study participants, or they may be generated within a study population by the dynamics of the life course.

The HRS and WLS provide examples of each of these opportunities. In the HRS, successive cohorts have experienced the rise and decline in equity markets since the early 1990s. These changes appear as exogenous shocks to the economic status and plans of HRS participants and, thus, provide useful information in estimating models about individual and familial provision for retirement, about economic transfers, and about labor market behavior in older populations. In its 2002 wave, the HRS began measuring subjective expectations about the value next year at this time of a hypo-

thetical mutual fund held in stocks. Preliminary analysis by Kézdi and Willis (2003) suggests that expectations vary greatly across respondents and are significantly related to actual investment behavior. Because the HRS was in the field from 1 April 2002 through 31 January 2003, actual variation in the recent history of stock prices at the time of interview will enable researchers to study how people alter their subjective beliefs about expected returns as market conditions change. Traditional finance theory with its emphasis on efficient markets and rational expectations has come under attack by researchers in the new field of behavioral finance, which emphasizes the heterogeneity of the beliefs of economic agents (Shleifer 2000). The HRS offers the first opportunity to address issues raised in this debate with data on a probability sample of the US population that combines information on expectations with actual behavior. Understanding of these issues is important for public policies regarding 401(k) plans and in assessing the potential benefits and pitfalls of introducing individual accounts into the Social Security system.

The WLS began, in the midst of the Cold War, as a study of the transition from high school to postsecondary schooling, military service, and the labor market. It has been sustained and increased in scientific value for two reasons: first, because the baseline measurements and survey data have survived across nearly half a century, and, second, because investigators have been willing to change direction as the sample aged. Thus, the study has now come to focus on family, health, and well-being, and the original focus on labor market outcomes is winding down. At the same time, by linking to external data the WLS has continued to expand its coverage of earlier phases of the life course, for example, using high school yearbooks, archival school district records, and, possibly, birth certificates.<sup>8</sup>

## Epilogue

We have examined six main desiderata in a system of survey-based research: (1) representation of real populations; (2) sustained institutional support; (3) responsibility to the public; (4) innovation based on multiple disciplinary perspectives; (5) coverage of multiple domains and units of observation; and (6) opportunities for flexibility, serendipity, and scientific opportunism. Each of these points is well illustrated in the development of the Wisconsin Longitudinal Study and the Health and Retirement Study.

Scientifically useful evolution of study design and content has not been automatic and inevitable. It is easy to think of contrary examples in the recent history of survey research. Obversely, we should not become so wedded to the rich, contemporary array of data resources that we fail to keep scanning the horizon for new and unexpected scientific opportunities. With

luck, the Wisconsin Longitudinal Study and the Health and Retirement Study will continue to extend their baseline measurements of aging processes and to generate novel scientific findings and innovations.

---

## Notes

This research was supported by grants to the University of Wisconsin–Madison and to the University of Michigan from the Behavioral and Social Research Program of the National Institute on Aging. We thank Willard Rogers, Jeremy Freese, and Wes Taylor for contributions to this text and James Smith and Linda Waite for helpful advice. The order of authorship of this paper is without social or academic significance. Correspondence may be sent to Robert M. Hauser (hauser@ssc.wisc.edu) or Robert J. Willis (rjwillis@umich.edu).

1 Sewell remained active in research using the WLS until his death in 2001 at the age of 91.

2 The same goal has been achieved in the Panel Study of Income Dynamics by following all individuals who enter sample households, rather than by refreshing the sample with new cohorts (Duncan, Hofferth, and Stafford 2004).

3 In Europe, surveys like the HRS are now underway, notably, ELSA (English Longitudinal Study of Ageing) and SHARE (Survey of Health, Ageing, and Retirement in Europe). Fortunately, each of these studies will use probability sampling, thus overcoming a tradition of quota sampling in several of the participating countries.

4 Because the WLS covers siblings as well as the 1957 graduates, it is possible to carry out some intercohort comparisons within the study population. However, such comparisons

necessarily pertain to differently selected populations. For example, graduates must have completed high school, but their siblings need not have done so. Graduates may be singletons, but siblings cannot be singletons.

5 For this reason, the WLS is planning special follow-up activities for low-ability graduates who do not complete the mail survey after the usual three mailings and reminders.

6 While these observations all refer to adolescent IQ test scores, differential non-response in the WLS is not solely a function of those scores. Rank in high school class has equally large effects on response. The key finding is that these cognitive/academic variables account for the association between socioeconomic variables and differential response. We thank Jeremy Freese for the tabulation of response heaping in Ryff's items.

7 Birth weights were also reported in some newspapers, and WLS staff members are exploring the feasibility of a search for such reports.

8 This parallels the evolution of the British birth cohort studies (Douglas and Blomfield 1958; Douglas 1964; Douglas, Ross, and Simpson 1968; Kerckhoff 1990; Wadsworth 1991; Kerckhoff 1993). As the cohorts have aged, scientific leadership of those studies has variously been exercised by pediatricians, developmental psychologists, education experts, demographers, sociologists, labor economists, and epidemiologists.

## References

- Adams, Peter, Michael D. Hurd, Daniel McFadden, Angela Merrill, and Tiago Ribeiro. 2003a. "Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status," *Journal of Econometrics* 112(1): 3–56.
- . 2003b. "Response," *Journal of Econometrics* 112(1): 129–133.
- Adda, Jerome, Tarani Chandola, and Michael Marmot. 2003. "Socio-economic status and health: Causality and pathways," *Journal of Econometrics* 112(1): 57–63.
- Allison, Paul D. 2002. *Missing Data*. Thousand Oaks, CA: Sage Publications.

- Arminger, Gerhard, Clifford C. Clogg, and Michael E. Sobel. 1995. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum Press.
- Averett, Susan and Sanders Korenman. 1996. "The economic reality of the beauty myth," *Journal of Human Resources* 31(2): 304–330.
- Biddle, Jeff E. and Daniel S. Hamermesh. 1998. "Beauty, productivity, and discrimination: Lawyers' looks and lucre," *Journal of Labor Economics* 16(1): 172–201.
- Chan, Sewin and Ann Huff Stevens. 2001. "Job loss and employment patterns of older workers," *Journal of Labor Economics* 19(2, April): 484–521.
- de Bruin, W. B., B. Fischhoff, S. G. Millstein, and B. L. Halpern-Felsher. 2000. "Verbal and numerical expressions of probability: 'It's a fifty-fifty chance,'" *Organizational Behavior and Human Decision Processes* 81(1): 115–131.
- Dominitz, Jeffrey and Charles F. Manski. 1999. "The several cultures of research on subjective expectations," in Robert J. Willis and James Smith (eds.), *Wealth, Work, and Health*. Ann Arbor: University of Michigan Press.
- Douglas, James W. B. 1964. *The Home and the School: A Study of Ability and Attainment in the Primary School*. London: Macgibbon & Kee.
- Douglas, James W. B. and J. M. Blomfield. 1958. *Children Under Five: The Results of a National Survey Made by a Joint Committee of the Institute of Child Health (University of London), the Society of Medical Officers of Health, and the Population Investigation Committee*. London: Allen & Unwin.
- Douglas, James W. B., J. M. Ross, and Howard R. Simpson. 1968. *All Our Future: A Longitudinal Study of Secondary Education*. London: P. Davies.
- Duncan, Greg J., Sandra L. Hofferth, and Frank Stafford. 2004. "Evolution and change in family income, wealth and health: The Panel Study of Income Dynamics, 1968–2000 and beyond," in James S. House, F. T. Juster, Robert L. Kahn, and Howard Shuman (eds.), *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*. Ann Arbor, MI: Institute for Social Research, Chapter 6.
- Farrell, P. and V. R. Fuchs. 1982. "Schooling and health: The cigarette connection," *Journal of Health Economics* 1(3): 217–130.
- Fischhoff, Baruch and Wändi B. de Bruin. 1999. "Fifty-fifty=50%?," *Journal of Behavioral Decision Making* 12(2): 149–163.
- Florens, Jean-Pierre. 2003. "Some technical issues in defining causality," *Journal of Econometrics* 112(1): 127–128.
- Geweke, John. 2003. "Econometric issues in using the AHEAD panel," *Journal of Econometrics* 112(1): 115–120.
- Gigerenzer, Gerd, Peter M. Todd, and ABC Research Group. 1999. *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.
- Granger, Clive W. J. 2003. "Some aspects of causal relationships," *Journal of Econometrics* 112(1): 69–71.
- Grossman, Michael. 1972. "On the concept of health capital and the demand for health," *The Journal of Political Economy* 80(2, March–April): 223–255.
- Groves, Robert M. and Mick Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.
- Hamermesh, Daniel S. and Jeff E. Biddle. 1994. "Beauty and the labor market," *The American Economic Review* 84(5): 1174–1194.
- Harker, L. and D. Keltner. 2001. "Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood," *Journal of Personality and Social Psychology* 80(1): 112–124.
- Hausman, Jerry A. 2003. "Triangular structural model specification and estimation with application to causality," *Journal of Econometrics* 112(1): 107–113.
- Hayes, Diane and Catherine E. Ross. 1986. "Body and mind: The effect of exercise, overweight, and physical health on psychological well-being," *Journal of Health and Social Behavior* 27(4): 387–400.

- Heckman, James. 2003. "Conditioning, causality and policy analysis," *Journal of Econometrics* 112(1): 73–78.
- Heeringa, S. G., D. H. Hill, and D. A. Howell. 1995. "Unfolding brackets for reducing item non-response in economic surveys," HRS Working Paper 94-029. Ann Arbor: Institute for Social Research, University of Michigan.
- Hill, Daniel and Robert J. Willis. 2001. "Reducing panel attrition: A search for effective policy instruments," *Journal of Human Resources* 36(3): 416–438.
- Hoover, Kevin D. 2003. "Some causal lessons from macroeconomics," *Journal of Econometrics* 112(1): 121–125.
- Hurd, Michael D. 1999. "Anchoring and acquiescence bias in measuring assets in household surveys," *Journal of Risk and Uncertainty* 19(1-3): 111–136.
- Hurd, Michael D. and Kathleen McGarry. 1995. "Evaluation of the subjective probabilities of survival in the health and retirement study," *Journal of Human Resources* 30(Supplement): S268–S292.
- Jackson, Daren C., Corrina J. Mueller, Isa Dolski, Kim M. Dalton, Jack B. Nitschke, Heather L. Urry, Melissa A. Rosenkranz, Carol D. Ryff, Burton H. Singer, and Richard J. Davidson. 2003. "Now you feel it, now you don't: Frontal EEG asymmetry and individual differences in emotion regulation," *Psychological Science* 14(6): 612–617.
- Juster, F. T. 1966. "Consumer buying intentions and purchase probability: An experiment in survey design," *Journal of the American Statistical Association* 61(315): 658–696.
- Juster, F. T. and James P. Smith. 1997. "Improving the quality of economic data: Lessons from the HRS and AHEAD," *Journal of the American Statistical Association* 92(440): 1268–1278.
- Juster, F. T. and Richard Suzman. 1995. "The Health and Retirement Study: Data quality and early results," *Journal of Human Resources* 30(Supplement): S7–S56.
- Kerckhoff, Alan C. 1990. *Getting Started: Transition to Adulthood in Great Britain*. Boulder, CO: Westview Press.
- . 1993. *Diverging Pathways: Social Structure and Career Deflections*. New York: Cambridge University Press.
- Kézdi, Gábor and Robert J. Willis. 2003. "Who becomes a stockholder? Expectations, subjective uncertainty, and asset allocation," presented at the Fifth Annual Joint Conference for the Retirement Research Consortium, 15–16 May 2003, Washington, DC.
- King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the validity and cross-cultural comparability of survey research," *American Political Science Review* 98(1): 191–207.
- Lillard, Lee A. 1977. "Inequality: Earnings vs. human wealth," *The American Economic Review* 67(2): 42–53.
- Lillard, Lee A. and Robert J. Willis. 2001. "Cognition and wealth: The importance of probabilistic thinking," Michigan Retirement Research Center Working Paper UM00-04, University of Michigan, Ann Arbor.
- Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical Analysis With Missing Data*. 2nd ed. Hoboken, NJ: Wiley-Interscience.
- Little, Todd D., Kai U. Schnabel, and Jürgen Baumert. 2000. *Modeling Longitudinal and Multilevel Data Practical Issues, Applied Approaches, and Specific Examples*. Mahwah, NJ; London: Lawrence Erlbaum.
- Manski, Charles F. 1990. "The use of intentions data to predict behavior: A best-case analysis," *Journal of the American Statistical Association* 85(412): 934–940.
- Mealli, Fabrizia and Donald B. Rubin. 2003. "Assumptions allowing the estimation of direct causal effects," *Journal of Econometrics* 112(1): 79–87.
- National Academy of Sciences (US), Science, Technology, and Law Panel, National Academy of Sciences (US), Policy and Global Affairs, and National Research Council (U.S.). 2002. *Access to Research Data in the 21st Century: An Ongoing Dialogue Among Interested Parties: Report of a Workshop*. Washington, DC: National Academy Press.

- National Research Council, Panel on Institutional Review Boards, Surveys, and Social Science Research. 2003. *Protecting Participants and Facilitating Social and Behavioral Sciences Research*, Constance F. Citro, Daniel R. Ilgen, and Cora B. Marret (eds.). Washington, DC: National Academies Press.
- National Research Council (US) and Panel on a Research Agenda and New Data for an Aging World. 2001. *Preparing for an Aging World the Case for Cross-National Research*. Washington, DC: National Academy Press.
- National Research Council (US), Committee on Population. 2001. *Cells and Surveys Should Biological Measures Be Included in Social Science Research?*, Caleb E. Finch, James W. Vaupel, and Kevin G. Kinsella (eds.). Washington, DC: National Academy Press.
- Poterba, James M. 2003. "Some observations on health status and economic status," *Journal of Econometrics* 112(1): 65–67.
- Robins, James M. 2003. "General methodological considerations," *Journal of Econometrics* 112(1): 89–106.
- Rodgers, Willard L. 2002. "Size of incentive effects in a longitudinal study," presented at the 57th Annual Meeting of the American Association for Public Opinion Research, St. Pete Beach, FL.
- Ryff, Carol D. 1989. "Happiness is everything, or is it? Explorations on the meaning of psychological well-being," *Journal of Personality and Social Psychology* 57(6): 1069–1081.
- Ryff, Carol D. and Corey L. Keyes. 1995. "The structure of psychological well-being revisited," *Journal of Personality and Social Psychology* 69(4): 719–727.
- Sadana, Ritu, Ajay Tandon, Christopher J. Murray, Irina Serdobova, Yang Cao, Wanjun Xie, Bedirhan Ustun, and Somnath Chatterji. 2001. "Describing population health in six domains: Comparable results from 66 household surveys," *The Global Burden of Disease 2000 in Aging Populations*. Research Paper No. 01.16. Cambridge, MA: Harvard Burden of Disease Unit, Center for Population and Development Studies.
- Salomon, Joshua A., Ajay Tandon, and Christopher J. Murray. 2001. "Using vignettes to improve cross-population comparability of health surveys: Concepts, design, and evaluation techniques," Discussion Paper No. 41. Geneva, Switzerland: World Health Organization.
- Sewell, William H., Robert M. Hauser, Kristen W. Springer, and Taissa S. Hauser. 2003. "As we age: The Wisconsin Longitudinal Study, 1957–2001," in Kevin Leicht (ed.), *Research in Social Stratification and Mobility*, vol. 20. London: Elsevier, pp. 3–111.
- Shleifer, Andrei. 2000. *Inefficient Markets an Introduction to Behavioral Finance*. Oxford; New York: Oxford University Press.
- Sloan, Frank A., V. K. Smith, and Donald H. Taylor, Jr. 2003. *The Smoking Puzzle: Information, Risk Perception, and Choice*. Cambridge, MA: Harvard University Press.
- Smith, James P. 1999. "Healthy bodies and thick wallets: The dual relation between health and economic status," *The Journal of Economic Perspectives* 13(2): 145–166.
- Smith, P. M. and B. B. Torrey. 1996. "The future of the behavioral and social sciences," *Science* 271(5249): 611–612.
- Snowdon, D. A., S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, and W. R. Markesbery. 1996. "Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study," *Journal of the American Medical Association* 275(7): 528–532.
- Snowdon, David. 2001. *Aging With Grace: What the Nun Study Teaches Us About Leading Longer, Healthier, and More Meaningful Lives*. New York: Bantam Books.
- Soldo, B. J., M. D. Hurd, W. L. Rodgers, and R. B. Wallace. 1997. "Asset and health dynamics among the oldest old: An overview of the AHEAD study," *Journals of Gerontology, Series B, Psychological Sciences and Social Sciences* 52 Spec No: 1–20.
- Taubman, Paul J. and Terence J. Wales. 1973. "Higher education, mental ability, and screening," *Journal of Political Economy* 81(1, January–February): 28–55.

- Thorndike, Robert L. and Elizabeth Hagen. 1959. *Ten Thousand Careers*. New York: Wiley.
- Tversky, Amos and Daniel Kahneman. 1974. "Judgment under uncertainty: Heuristics and biases," *Science* 185:1124–131.
- Vermunt, Jeroen K. 1997. *Log-Linear Models for Event Histories*. Thousand Oaks, CA: Sage Publications.
- Wadsworth, M. E. J. 1991. *The Imprint of Time: Childhood, History, and Adult Life*. Oxford: Clarendon Press.
- Willis, Robert J. 1999. "Theory confronts data: How the HRS is shaped by the economics of aging and how the economics of aging will be shaped by the HRS," *Labour Economics* 6(2, June): 119–145.
- Willis, Robert J. and Sherwin Rosen. 1979. "Education and self-selection," *The Journal of Political Economy* 87(5, Part 2: Education and Income Distribution): S7–S36.

