



The National Longitudinal Survey of Youth: 1979 cohort at 25

The 1979 cohort of the National Longitudinal Survey of Youth has been a font of information for researchers of all stripes; the Monthly Labor Review brings together the results of research on topics ranging from employment, to attrition in the survey, to data on education, to the children of survey respondents

Charles Pierret

This issue of the *Monthly Labor Review* celebrates the 25th anniversary of the National Longitudinal Survey of Youth, 1979 Cohort (NLSY79). The National Longitudinal Surveys (NLS) program, of which the NLSY79 is the flagship survey, is a bit of an anomaly among the Bureau of Labor Statistics many data collection efforts. None of the Bureau's key economic indicators relies on NLS data. Only a couple of the more than one hundred press releases the Bureau publishes each year involve data collected by the NLS program. It is doubtful that financial markets ever will react strongly to the release of NLS data. And unlike the current employment statistics, the inflation statistics, or the unemployment rate, measures from the NLSY79 are not likely to be discussed in everyday conversation or even in the business news.

Yet, the NLSY79 has been extremely influential. Over the last 25 years, it has provided the data for thousands of Ph.D. dissertations, working papers, journal articles, and books that have shaped theory and knowledge in disciplines such as economics, sociology, education, psychology, and health sciences. The survey's primary constituency includes hundreds of researchers within universities, think tanks, and government agencies both in the United States and abroad. Because of its quality, breadth, and thoroughness, the NLSY79 has become probably the most analyzed longitudinal data set in the social sciences. Almost every issue of leading labor

economics and demography journals contain at least one article that uses NLSY79 data.

The main product of the NLS program, unlike that of most other BLS programs, is the actual microdata generated by the surveys. For each respondent, researchers can access a record that details his or her responses to every question in the survey, along with summary and supporting information. Of course, this record is stripped of all information that could identify the individual respondent.¹ With 20 rounds of data currently available, the NLSY79 has become an enormous data set, comprising more than 75,000 variables for each respondent and requiring about 500 megabytes for its storage. The NLSY79 provides researchers with data from a nationally representative sample of 12,686 individuals who have participated in up to 21 hour-long interviews over the last 25 years.² These individuals were 14 to 22 years old when they were first surveyed in 1979; they are now in their forties. By observing their lives over the 25-year period, researchers can study the life course of a large sample of American men and women born at the end of the baby boom (1957 to 1964) as they navigate the years between adolescence or young adulthood and middle age.

Although the primary focus of the survey is labor force behavior, the content of the survey is considerably broader. The NLSY79 contains an expansive set of questions ranging from childcare costs to welfare receipt. For example, the survey

Charles Pierret is Director of National Longitudinal Surveys, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics.
E-mail: pierret.charles@bls.gov

includes detailed questions on educational attainment, investment in training, income and assets, health conditions, workplace injuries, insurance coverage, alcohol and substance abuse, sexual activity, and marital and fertility histories.

This wide array of subject matter is certainly one key to the survey's broad utilization for academic research. Recognizing that decisions made in one realm of life often affect and are affected by events in other realms, the NLSY79 questionnaire tries to get a comprehensive view of the lives of survey respondents. For example, health, childcare, and family constraints are important inputs to any labor market choices. By collecting information in many domains, the NLSY79 gives researchers the ability to simultaneously examine and control for multiple correlates of complex phenomena.

The NLSY79 also benefits from the inclusion of information collected outside of the main survey. For example, scores on the Armed Services Vocational Aptitude Battery, a series of 10 tests measuring knowledge and skill in areas from paragraph comprehension to electronics, are available for 94 percent of sample respondents. The Armed Forces Qualifying Test score, a composite of the scores on four of these tests, is one of the most popular variables in the survey, in that it can be used to control for differences in cognitive development prior to the start of the survey. Other ancillary data include a survey of the secondary schools attended by NLSY79 respondents and detailed information from the respondents' high school transcripts.

The breadth of the content in the survey is complemented by the longitudinal design, a second key to the success of the NLSY79. Not only can researchers correlate behavior in multiple domains, but they can do so over long periods; for example, they can investigate how family structure or educational experiences as a teenager affect employment decisions in a person's twenties and thirties. To facilitate this type of analysis, much of the information in the NLSY79 is gathered in event history format, in which dates are collected for the beginning and ending of important life events. Data on respondents' labor market status, marital status, fertility, and participation in government assistance programs such as unemployment insurance and Aid to Families with Dependent Children are all collected in this manner. The event history format allows researchers to sequence key events so that a first attempt at establishing causality can be made: only if *A* precedes *B* can we postulate that *A* caused *B*.

But the real advantage of longitudinal data in terms of studying causality is the existence of multiple observations of the same person. A classic econometric problem is the existence of unobserved personal characteristics that may be correlated with both the dependent variable of interest and an independent variable that is hypothesized to cause the dependent variable. For example, consider the wage premium associated with marriage among men.³

With cross-sectional data, one might compare the wages of married men with those of unmarried men to calculate the wage premium. Even after controlling for all observable characteristics, though, one might suspect that there are some differences that are not easily measurable (for example, interpersonal skills or ambitiousness), but that might make one both more likely to be married and more likely to earn higher-than-average wages. With longitudinal data, the standard technique is to analyze changes either directly, by regressing the change in wages on the change in marital status, or indirectly, by using a fixed-effects framework. In this way, longitudinal data give us the ability to control for individual effects by using multiple observations of the same individual.

A final key to the NLSY79's success is the high quality of its data. The fundamental measure of quality in longitudinal data is the *retention rate*—the percentage of initial respondents who respond in later rounds. Among social science surveys, the NLSY79 is the undisputed leader in this regard. In 1994, after 15 rounds of interviewing, more than 90 percent of the survey's eligible initial respondents were still being interviewed.⁴ While attrition has picked up since that time, more than 80 percent of living initial respondents were interviewed in the 2002 survey.

Other dimensions of data quality are also high in the NLSY79. Nonresponse to individual questions (either refusals or "don't know" responses) is quite low; only a handful of questions have nonresponse rates above 1 percent. Because of the longitudinal design of the survey, missing data can often be recaptured in subsequent interviews. As Randall J. Olsen explains, this approach can lead to effective sample sizes larger than the response rate would imply.⁵

Given the size of the NLSY79 data set, using the data is relatively easy.⁶ Researchers can download the entire data set, along with extraction software to pull off particular variables on their own computer, or they can use the extraction software at the Web site and download smaller data sets. Typically, researchers then use statistical software such as SAS, SPSS, or STATA to process the data and perform their analyses. This accessibility can be combined with the richness of the data to study many topics across myriad disciplines. The remaining articles in this issue of the *Review* highlight the contributions of the NLSY79 to research, with special emphasis on issues in the area of labor economics.

James A. Walker points out that the introduction of nationally representative longitudinal microdata sets, combined with ever-increasing computing power, has created a revolution in social science research over the last few decades.⁷ Previously, researchers had to rely on aggregate statistics to grasp the workings of large complex systems. Now they can study individual actors to build an understanding of social structure based upon those micro-

foundations. Before, analysts were forced to use small, potentially unrepresentative samples to infer “normal” behavior. Now, large surveys give them access to thousands of individuals to study at very low cost. The combination of massive amounts of data and growing computing power led to the development of new statistical methods to exploit this new source of information.

The first of these data sets, including the original NLS, failed to fully exploit their longitudinal design. Instead, they tended to ask questions as if the survey were cross sectional, concentrating on the situation at the time of the interview. The genius of the NLSY79 was that it attempted to capture information about what was happening *between* interviews. Now events of short duration—jobs, marriages, spells of unemployment or public assistance, and the like—were captured, and events across different domains could be sequenced. Analysts now knew, for example, whether the job change happened before or after the marriage, so issues of causality could be addressed.

A second major improvement in the NLSY79 over previous surveys using longitudinal data sets was the NLSY79’s strong emphasis on reducing attrition and the effects of attrition on the data. Even seemingly modest attrition can lead to sharp declines in sample size. If participants who do not respond in a particular round are never recontacted, and if 3 percent of the remaining sample declines to participate in every round, only slightly more than one-half of the initial sample will be left after 20 rounds. The NLSY79’s retention rate has required interviewing an average of 99 percent of the number of the previous round’s respondents every round. As Olsen’s article shows, this task involved increased efforts in terms of both survey design and field operations. All respondents were recontacted every round, regardless of how long it had been since they were last interviewed. A surprising number were reinterviewed after missing one or several rounds. In addition, questions were written to recapture data that had not been retrieved in interviews that had been missed. Respondents were asked about events that had occurred since the last interview, not those which were ongoing or had happened in the past year.

Of course, even the highest-quality data are worthless if they are unrelated to any valuable or interesting content. The other six articles in this edition focus on particular content areas. Kenneth I. Wolpin’s article on education data in the NLSY79 details a large number of correlates of educational attainment that can be found in the survey’s database.⁸ Not only do more educated individuals earn more money, on average, but they work more hours, are less likely to be unemployed or receive welfare payments, have better health and less obesity, drink less, wait longer to have children and have fewer of them, and, among women in the survey, bear children who weigh more at birth and score higher on cognitive tests. These findings hint at the richness of the

NLSY79 data, among which associations like these can be explored in great depth.

In her article on early career job churning, Julie A. Yates also touches on the richness of the education data in the NLSY79.⁹ She shows that the concept of educational attainment is not at all static: people leave school and return, sometimes getting new educational credentials, sometimes not. This ever-present activity adds significantly to the level of education in the United States. Indeed, only 11 percent of those born between 1961 and 1964 had a college degree when they first left school, but more than 25 percent had a degree by their 35th birthday.

The main focus of Yates’s article, though, is the length of time required after leaving school to acquire a job that will last 1, 3, or 5 years. She finds that, by age 35, about 60 percent of those who left school with a high school degree or greater have worked at a job for at least 5 years. For high school dropouts, however, the figure is only 36 percent. Interestingly, the route high school graduates take to acquire a job of this duration is different from that taken by college graduates: the median high school graduate takes 10.1 years and 5 shorter jobs before finding the job that will last 5 years, whereas a college graduate requires only 3.5 years and 2 jobs.

Audrey Light also uses the work history data of the NLSY79 to study employment dynamics—specifically job mobility and wage growth.¹⁰ She asks whether wage growth is greater for those who stay in one job or those who change employers frequently. The answer she arrives at depends at least partially on the reasons for mobility. Those who change jobs voluntarily have wage growth of the same magnitude as those who stay in jobs; those who change jobs involuntarily have lower wage growth. Light admits that her analysis is “cur-sory.” Given the complexities underlying her question—including the role of spells out of the labor force and their interactions with education and gender—it is not surprising that there is no simple answer to the question of how job mobility affects wage growth. But by providing longitudinal employment data along with comprehensive contextual data, the NLSY79 will continue to play a key role in our growing understanding of employment dynamics.

The NLSY79, states Robert W. Fairlie, has many features that make it attractive for the study of self-employment:¹¹ a large, nationally representative sample; detailed financial information; data on family and educational background; and measures of cognitive and psychological suitability. But the survey’s strongest asset is its comprehensive work history, which tracks transitions into and out of self-employment as well as wage and salary employment. A great deal of self-employment, at least among the young, is episodic, with workers moving often between employment states. But as workers age, their participation in self-employment grows and becomes more stable. Only 5 percent of 22-year-old men are self-employed; by age 42, 12 percent of men are self-employed.

The NLSY79, because of its longitudinal design, is ideal for investigating the dynamics of self-employment.

Since the 1950's, the reigning paradigm of labor economics has been human capital theory, in which workers invest in productive skills in order to earn higher wages. The study of human capital has divided skills acquisition into two components: education—in which general knowledge is acquired largely prior to the start of one's working career; and training—in which more specific, job-related skills are acquired, often during one's career. As Harley J. Frazis and James R. Spletzer assert, empirical labor economics research on training has lagged behind research on education because research on training placed much greater demands on the data.¹² Again, the NLSY79 has been a key data set in this literature, not only because it captures details about training, but because it contains a complete history of employment and thousands of other contextual variables.

The final article in this volume introduces an additional dimension of the NLSY79. In a unique collaboration between Federal Government agencies with complementary interests, the NLS program, with the financial support of the National Institute of Child Health and Human Development, has been surveying the children of all women respondents to the NLSY79 every 2 years since 1986. In their article, Lawrence L. Wu and Jui-Chung Allen Li describe this survey, called the Children of the NLSY79.¹³ Mothers are asked about pregnancy, prenatal care, and childcare, as well as the early development of their children. Young children are given cognitive and developmental assessments from the age of 4 to the age of 14. Children between ages 10 and 14 are interviewed briefly about school and family life. Those older than 15 years are given an interview that is similar to the NLSY79 interview. What emerges is a unique data set with a complete developmental history from birth and a comprehensive background of the mother, starting many years before birth. Over time, the sample size has grown to rival that of the NLSY79 itself. At its inception in 1986, the Children of the NLSY79 survey comprised 4,971 children. In 2002, 7,467 children were assessed or interviewed, ranging in age from 2 to 30 years.

These articles only scratch the surface of what is available in the NLSY79. Space limitations and the customary focus of the *Monthly Labor Review* dictate a concentration on employment, education, and training. But there is much more. The NLS program tracks research using NLS data in the NLS bibliography.¹⁴ Among other topics that have been the subject of at least 100 papers using NLSY79 data are income and earnings, marriage, maternal employment, adolescent fertility, alcohol and substance use, childcare, family structure, and government assistance programs. Smaller numbers of papers have used NLSY79 data to examine topics as diverse as job satisfaction, depression, migration, breast-

feeding, parental leave, savings, and the involvement of fathers in raising children. There truly are few topics in social science that have not seen a contribution from NLSY79 data.

The success of the NLSY79 also led the Bureau of Labor Statistics to introduce a new cohort to its own longitudinal survey program. The NLSY97 is a survey of 8,984 youths who were born between 1980 and 1984 and who were 12 to 17 years of age when first interviewed in 1997. While the basic design of this survey is quite similar to that of the NLSY79, several changes were introduced because of the NLS program's experience with earlier surveys. The most important change was that NLSY97 respondents were younger when first interviewed than were those in the NLSY79. It was felt that respondents should be observed while still in school and still in the parental home, so that all transitions to the world of work and to independent living could be recorded. Also, starting with a younger sample allowed the survey to collect more background on the respondents' introduction to the labor market. Information on informal jobs such as babysitting and yard work was collected for those as young as 12 years.¹⁵ In another attempt to understand initial conditions, the NLSY97 interviewed a parent for each respondent. These interviews focused on the child's family, health, and schooling background, as well as the resources available to the child in the parental home. The eighth annual interview of the NLSY97 is currently in the field, and the respondents are now in their twenties. Already, a number of articles using these data have been published on a wide array of topics. It appears that the NLSY97 is poised to emulate the success of the NLSY79.

Many people have contributed to the success of the NLSY79 program, which is managed and funded by the Bureau of Labor Statistics. Since the survey's inception, the Center for Human Resource Research at the Ohio State University has directed survey operations under contract to the Department of Labor. The Center is responsible for survey development, data processing, and user support. In turn, the actual interviewing has been performed under subcontract by the National Opinion Research Center at the University of Chicago. Designing, preparing, fielding, processing, and disseminating a survey of the complexity of the NLSY79 requires many dedicated professionals. The success of the survey is a tribute to their hard work and dedication.

The contributions of two groups are especially noteworthy. Over the last 25 years, the field-interviewing staff at the National Opinion Research Center has consistently exceeded response rate targets by locating elusive respondents, convincing reluctant respondents to participate, and maintaining high levels of data quality. Still, the most important group in the success of the NLSY79 is the respondents. For 25 years, the members of this group have endured thousands of sometimes difficult and intrusive questions for little more than

the knowledge that they are involved in an important research project. Their assistance has added immensely to academic and policy research across a wide array of disciplines, helping

those involved to increase their knowledge, elaborate new theory, and develop new ways of understanding critical issues in the social sciences. □

Notes

¹ Under special arrangement, researchers can access certain geographic information about respondents in order to link environmental variables to the records of those respondents. Personally identifying information, however, is never available to researchers; it is protected by law through the Confidential Information Protection and Statistical Efficiency Act.

² Much of the summary information that follows can be found on the NLSY79 Web site at <http://www.bls.gov/nls/nlsy79.htm> or in the NLSY79 *User's Guide* (Bureau of Labor Statistics, 2001).

³ For a more thorough discussion of this issue, see Harry A. Krashinsky, "Do Marital Status and Computer Usage Really Change the Wage Structure?" *Journal of Human Resources*, summer 2004, pp. 774–91.

⁴ Two subsamples comprising roughly 3,000 respondents were dropped from the sample in 1985 and 1991 to reduce costs. The current sample size is 9,964 initial respondents, of which 346 were deceased at the time of the 2002 survey.

⁵ Randall J. Olsen, "The problem of respondent attrition: survey methodology is key," this issue, pp. 63–70.

⁶ The data are available at no cost on the Internet at <http://www.bls.gov/nls/nlsy79.htm>.

⁷ James A. Walker, "Antecedents and predecessors of the NLSY79: paving the course," this issue, pp. 8–14.

⁸ Kenneth I. Wolpin, "Education data in the NLSY79: a premiere research tool," this issue, pp. 15–20.

⁹ Julie A. Yates, "The transition from school to work: education and work experiences," this issue, pp. 21–32.

¹⁰ Audrey Light, "Job mobility and wage growth: evidence from the NLSY79," this issue, pp. 33–39.

¹¹ Robert W. Fairlie, "Self-employment, entrepreneurship, and the NLSY79," this issue, pp. 40–47.

¹² Harley J. Frazis and James R. Spletzer, "Worker training: what we've learned from the NLSY79," this issue, pp. 48–58.

¹³ Lawrence L. Wu and Jui-Chung Allen Li, "Children of the NLSY79: a unique data resource," this issue, pp. 59–62.

¹⁴ On the Internet at <http://www.chrr.ohio-state.edu/nls-bib/>.

¹⁵ See, for example, "Employment of Teenagers during the School Year and Summer," news release 04–217, on the Internet at <http://www.bls.gov/news.release/pdf/nlsyth.pdf>.