Volume Title: Health and Labor Force Participation over the Life Cycle: Evidence from the Past

Volume Author/Editor: Dora L. Costa, editor

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-11618-2

Volume URL: http://www.nber.org/books/cost03-1

Conference Date: February 2-3, 2001

Publication Date: January 2003

Title: Reflections on the Early Indicators Project.A Partial History

Author: Larry T. Wimmer

URL: http://www.nber.org/chapters/c9626

1

Reflections on the Early Indicators Project

A Partial History

Larry T. Wimmer

While many events have clearly defined starting points, the origin of a research project is often more difficult to identify. It is more likely a function of whom you ask and the point of reference used. So it is with the beginning date of the project "Early Indicators of Later Work Levels, Disease, and Death" (EI). It might seem logical to date the project from the year our proposal (Fogel 1991) first received funding from the National Institutes of Health (NIH) and the National Science Foundation (NSF). That award, however, came five years after our initial application (1986), when we were politely told, "An interesting idea, but we are not convinced that you can actually collect these records. When you can demonstrate feasibility come back." I suspect that some on the panel did not expect to see us again. In the interim, at the urging of Bob Fogel, we completed the collection software that was of concern to the panel and collected a sample of twenty companies in order to demonstrate the feasibility of our collection procedures. That twenty-company sample quickly became the basis of several significant research papers. Approval of the project and its funding initiated seven years of intense collection. Even before that first application (1986), support from the National Bureau of Economic Research (NBER) deserves much of the credit for the "start" of the EI project. As early as 1981, NBER found the idea sufficiently promising to contribute advanced funding as part of their ongoing support of the Development of the American Economy (DAE), which had begun in 1979.

In my mind, however, I believe it is reasonable to date the origin of the EI project back to 1972! In that year Bob Fogel and Stan Engerman were proposing the collection of probate data in order to study American slav-

Larry T. Wimmer is professor of economics at Brigham Young University.

ery. While visiting Emory University, Bob went to the archives in Atlanta to examine the records that were expected to provide a critical source of data on slave prices. He was told that the original probate records were often spread throughout county archives across the South, making collection expensive and oversight very difficult. As if that were not sufficiently discouraging, he was also informed that many of the records were too fragile to allow public use. "However, you might consider using the microfilmed probate records collected and available in one location at the Family History Library of the Church of Jesus Christ of Latter-day Saints in Salt Lake City." My recollection is that within minutes Bob was on the phone to Clayne Pope and me asking, "What do you know about the Family History Library?" Both of us, former students of Bob's, were at Brigham Young University just south of Salt Lake City. As I recall, our answers were approximately the same. "The library is where my mother has done family genealogy work, but other than that I have no idea. But, we will find out!" Earlier, Alice Hanson Jones, while completing her dissertation at Chicago on colonial wealth, made extensive use of the library; knowing that I was from Utah, she told me what "helpful people" they were at the library. Unfortunately, her comments had not made much of an impression upon my mind. In retrospect, our answers to Bob's question seem unbelievably naive as we look back upon almost thirty years of our own work and that of many other social scientists using the immense microfilm collection of personal, church, city, county, state, and federal records found in the Family History Library.

The year 1972 dates the start of collaborative work and common interests that culminated in our joint EI project. Initially, Bob and Stan used the probate records of the Family History Library for age-specific slave prices, an important contribution to their "new" data in *Time on the Cross* (Fogel and Engerman 1974). From the outset, Bob seemed interested in almost every aspect of the history and collection of the library. Eventually, the Family History Library recognized him during its centennial year in 1994 for his contributions and ongoing association. Meanwhile, Clayne, Jim Kearl, and I were using the archival data to study wealth determinants and distributions in early Utah (Kearl, Pope, and Wimmer 1980). Bob and Stan's work on the height of slaves and its correlation with their mortality experience suggested the value of height data as another means of estimating overall health and standard of living. Subsequently Bob, Stan, and Roderick Floud (of the London Guildhall University) began investigating other archives for height data as a proxy for net nutritional intake. Pope used the library's published family histories to begin a reconstruction of the mortality experience of nineteenth-century U.S. populations. The work of these several authors merged in 1978 with a study on "The Economics of Mortality in North America, 1650–1910" (Fogel et al. 1978).

Much of this early work suggested a deterioration in height, health, and life expectancy during the mid-nineteenth century while wages continued to rise. This perplexing puzzle led to the search for further data that might shed light on socioeconomic and health conditions during this critical period. It was in the search for such data that the extensive collection of military records in the National Archives involving Civil War recruits and veterans came to our attention. These military records, combined with the census manuscripts and published family histories, suggested the possibility of a surprisingly complete prospective study of aging among Northern white males during the specific time period in question. These data, linked across individuals and across multiple records, yield information starting with the national origin, wealth, and occupation of the parents of our young recruits in 1861; identify each battle, disease, and hospitalization of a recruit during his wartime service; provide a documented record throughout the remainder of the veteran's life as he entered the massive pension system stemming from the Civil War; and finally conclude with later-life family structure, living circumstances, and employment found in the 1900 and 1910 federal census records.

The early phase of the project involved five years of development and testing of alternative software and collection methods, followed by seven years of actual data collection. It was clear from the outset that such a study required collaboration across different academic fields. Thus, concurrent with the collection, a number of demographers, economists, and medical researchers began examining the data for a wide range of issues. These issues include the influence of height and other socioeconomic and biomedical factors on the development of specific infectious and chronic diseases; labor force participation at middle and late ages; and elapsed time to death. The initial team included Bob Fogel (economist, University of Chicago), James Trussell (demographer, Princeton), Nevin Scrimshaw (medicine, Massachusetts Institute of Technology [MIT] and Harvard), Irwin Rosenberg (medicine, Tufts), Michael Haines (economist, Colgate), and Clayne Pope and me (economists, Brigham Young University [BYU]). To this list eventually were added a number of other senior investigators, including Robert Mittendorf (medicine, University of Chicago) and a growing number of graduate students of Bob's from Chicago. Many of the latter have gone on to play major roles in the direction of the project, even becoming senior investigators themselves in future proposals—this list includes Dora Costa, Sven Wilson, Chulhee Lee, and John Kim.

The EI project proposed the collection of a life-cycle sample of 39,616 men mustered into 331 randomly selected companies of the Union Army. The sample is drawn from eight different federal record sources and is supported by additional information regarding local health, water conditions, and incidence of disease. Four of the eight records are the federal censuses of 1850, 1860, 1900, and 1910. The military records constitute the core of the sample, and include the Military Service Records, the Carded Medical

Records, the Pension Records, and the Surgeons' Certificates. The Military Service Records (MSR) contain information on each recruit before enlistment (location, occupation, and physical characteristics such as height, etc.), plus a daily muster record including health, battles, wounds, hospitalization, desertion, POW status, cause of death, or muster-out information. The Carded Medical Records (CMR) contain information from field and regimental hospitals, including length of stay, reason for admission, and condition or disposition upon release. The most valuable records and those making this study unique are the Pension Records (PEN) and the Surgeons' Certificates (SCRT) from the federal pension system. These records begin with the introduction of a veteran into the pension system whenever an initial claim is made and, after a lifetime of claims, affidavits, documents, letters, counterclaims, etc., conclude in most cases with evidence relating to the veteran's death—a death certificate or accompanying letters confirming the time and cause of death. It is not uncommon for the pension to continue beyond the death of the veteran through claims stemming from the veteran's widow or family. The PEN records frequently contain several hundred documents that, unlike the census manuscripts, exist for the explicit purpose of authenticating every aspect of a veteran's claim relative to his true identity, age, military service, past and present residence, previous and current employment, general and specific health-related conditions, and fitness for manual labor—plus general economic circumstances and later-life health, retirement, and family structure.

It was immediately clear that the records of the federal pension system constituted a very promising data source that might be used to answer a wide range of questions regarding health, migration, and labor force participation for a large segment of our population from the mid-nineteenth century through the first quarter of the twentieth century. These records provide us with an important benchmark on infectious and chronic diseases before our modern understanding of germ theory, before widespread public health programs, and before the introduction of modern intervention into disease treatment—a benchmark against which to judge the enormous improvements in medicine and life expectancy taken for granted in the twentieth century. These data reveal much about labor force participation among an aging population a generation before the introduction of Social Security and other pension programs. First, however, we were confronted with a number of roadblocks that had to be overcome for the project to succeed.

One of our first challenges was that of devising a system for linking each recruit across all eight records covering a period of as much as eighty years. Little was known then of linking except within communities where migration meant leaving the sample. Complicating our task was the sheer size of the sample, plus a large number of common names, considerable interstate migration, and frequent name changes or use of aliases. What today is commonplace was then a serious set of questions: What form should a unique identification number take? (We even experimented with unique letters and combinations of numbers and letters.) Would it be sufficient, and if so, how should such a number be devised? The task of linking seemed possible only because of the location and other identifying information found in the military pension records. It was not self-evident that we could practically or reasonably produce a historical data set large enough to make inferences specific to age, location, occupation, and disease.

One of the most serious challenges involved our need to produce an interactive software system capable of linking the different purposes among skilled programmers, professional researchers, and the people who would input the data. We were faced with problems that today are part of any commercially available software but then were major hurdles for us. No interactive, commercial packages were available that could give us real-time feedback or could handle the size of our data set. Our earliest collections were written down by hand at the archives, sent to BYU to be keypunched, and subsequently entered into the university's mainframe for further analysis both here and at the University of Chicago. Such procedures were extremely time consuming and expensive, and multiplied the probability of introducing errors into the data set. The introduction of "new," thirteen-pound portable computers was promising for collection at source—earlier laptops had been judged to be too expensive and the screens too difficult to read! In this way, the EI project bridges the interval from keypunched cards to laptop computers that were more powerful than our university's early mainframe! Mark Showalter's association with the project illustrates another example of the time and cost of developing our own software. As an undergraduate at BYU he helped with our first efforts at software development. Subsequently, Mark has completed his Ph.D. at MIT and been a colleague in our department for ten years. Randy Campbell, Shawn Jordan, and Steve Shreeve took over software development from Mark and were ultimately responsible for the software used throughout the collection phase of the project.

That we might be able to use laptop computers and develop our own interactive software to link across multiple years and records answered only one of many difficult questions. The much larger question involved the complicated nature of the records themselves. While we had collected census data for the Utah wealth project, searching for almost 40,000 recruits spreading quickly across the United States presented a formidable challenge. In addition, information found in each federal census year differs from that of previous years. Nevertheless, as daunting as these records seemed initially, census collection became the least of our problems. Janet Bassett, a genealogist working with our programmers at BYU, was responsible for the development of a series of fixed-field collection screens for each census year using our interactive software. Subsequently, she supervised the training of our student teams who actually performed the tasks of searching, verifying, linking, and collecting the census records. With adequate money and student time we were confident that we could collect the census records, although questions remained regarding retrieval and linkage rates that could be answered only with time and experience.

The MSR and CMR presented us with a different set of problems. All collection involving these records had to be done on site at the National Archives, where we were fortunate to find generous administrators and exceptionally skilled and experienced staff. Once on site we found that, unlike with the census records, locating the military records was a minor problem. These data often appeared in a fixed format; however, the format itself changes frequently and contains several open-ended responses.

After several months of working with these records, we were able to decide upon fixed-field collection screens that enabled data-entry persons to identify and collect the relevant information from each record and iteration of forms. Julene Bassett and Noelle Yetter, two former BYU students working for us at the archives, helped achieve this second success. There are a total of eleven collection screens involving the MSR and CMR, including information regarding the recruit's age, residence, and occupation before enlistment, plus military and medical data from the date of enlistment to his departure resulting from his death, going AWOL, or being mustered out.

The PEN and SCRT records with their hundreds of free-form letters, affidavits, and documents presented by far the greatest software and collection challenges. Initially, Clayne and I feared that the task associated with the PEN records might be insurmountable, that student data-entry persons would simply have no idea what to include and what to leave out of such a mass of information. Even if we could construct a manageable set of collection screens there was the real possibility that it might require such extensive supervision, take so long, and cost so much that it would make collection impractical. In addition, we worried about consistency and error rates associated with collection. Bob assured "us" that "we" could solve these problems! After a year of living with the PEN records and a number of false starts, we had a set of eight fixed-field collection screens, with accompanying backups and expansion screens containing more than 3,200 variables covering the life cycles of these recruits from before enlistment to their death. These screens enabled carefully trained and supervised students to turn page by page through the pension records, identifying the desired information and retrieving the relevant data from each document. At the National Archives, Noelle Yetter provided the commitment, continuity, and consistency that made this collection possible. She trained, supervised, performed de novo testing of error rates, scolded, and encouraged almost 100 students from BYU, working in teams of eight to twelve at a time, over the ensuing seven-year period.

Bob had been right about the feasibility of collecting the PEN records—although Clayne and I reminded him that his "optimism" led us to substantially underestimate the time and effort in bringing about that outcome. Celebration necessarily waited upon solving what Clayne and I were convinced was the greatest obstacle of all, the Surgeons' Certificates (SCRT)! How could we expect undergraduates with no medical background or training to collect data from century-old records, using archaic nineteenth- and early twentieth-century medical terminology? The typical veteran experienced an average of almost five examinations during his time within the pension system, with some having over thirty. Each exam is the result of an appearance by the veteran before three pension-board-certified physicians, and stems from his initial claim or subsequent petition for changes in his classification. These examinations cover a wide range of health-related conditions, from accidents and wounds either during or after the war to infectious and chronic diseases.

Clayne and I warned Bob that unlike the census, MSR, CMR, or PEN, about which we had previously expressed our doubts, in the case of the SCRT we might truly be facing a hopeless task. Not surprisingly, Bob insisted that it was possible and even offered to have Chicago take on this task. The first set of SCRT collection screens we received consisted of individual, fixed-field screens for each disease. In a fixed-field format each disease set must anticipate all possible responses that might be encountered for that specific disease. As a result, some collection sets contained as many as fifty to sixty separate screens per disease. We tested these screens using several of our best data-entry persons, and found that they typically became lost early in the detail of a single disease. After a very large investment of time, and through the joint efforts of our medical personnel, programmers, and experienced collection supervisors, we produced what may be the major accomplishment of the collection phase of the project. The impossible ultimately became an impressive set of thirty-nine screens combining fixed-field, variable-width, and open-ended comments that has enabled the collection of these incredibly complicated and yet extraordinarily valuable SCRT. Each screen includes the flexibility of multiple backup screens for additional information, and can be amended as new information and conditions are encountered. If each variable is found only once there are a total of over 2,300 variables—counting is more probabilistic in the case of SCRT since they include open-ended questions. Our success with the SCRT depended very heavily upon the efforts of Nevin Scrimshaw and Irwin Rosenberg, two of our senior medical investigators. Their input was absolutely essential, but as is so often the case, the tedious hours of writing, testing, and re-rewriting screens over this period fell upon students: Julene Bassett and Sharon Nielsen (two former BYU students) and Louis Nguyen (then a medical student at Chicago). Subsequently, Julene and Sharon, and finally Brant Williams, supervised almost every stage of the collection project. Louis, now a physician at the Barnes Jewish Hospital, St. Louis, remains an active participant in the project and in the analysis of the data.

During the seven years of collection, over 200 students collected 303 companies of census records, MSR, CMR, PEN, and SCRT involving 35,571 recruits of the Union Army. A subsequent proposal has been approved which will complete the full 331 companies; add samples from the 1870 and 1880 censuses, a sample of 6,000 Black recruits, and a sample of 10,000 males rejected for military service—as well as add information from private family histories to the Civil War sample (Fogel 2001). The decision was made to conclude this initial collection by recollecting the original twenty-company sample as a quality check against our data collection procedures. Other quality controls consist of over 100 automated checks upon upper and lower bounds of numeric values, acceptable dates, place names, and occupations. Two-field checks are used to compare date intervals involving ages at marriage, death, etc. These built-in, automated checks are in addition to training, supervision, weekly supervisors' meetings, and de novo testing of 5 to 10 percent of the data to track error rates. These quality control instruments and error rates are reported in the data user's manuals prepared by the Center for Population Economics (CPE) at the University of Chicago and the Department of Economics at BYU. Finally, all the data were visually inspected by our BYU project supervisor for any apparent outliers before our phase of the collection was considered complete and the raw data sent to the CPE.

Of the 35,571 initial recruits in the current EI sample, 98 percent (34,775) were linked to the MSR and 85 percent (30,286) to the CMR (Fogel et al. 2001, 70). The size of the sample is suggested by the 1,230 and 1,495 variables found in the relatively small MSR and CMR (CPE 2000, 1). Of those veterans who survived the war and therefore were at risk of being found in the pension system, 79 percent (24,185) have a PEN (Fogel et al., 170). Of those with a PEN, 69 percent have at least one SCRT, with an average of 4.63 examinations per veteran. The SCRT contain 81,877 observations on 2,312 variables for the 16,713 veterans with SCRT (CPE 1999, 1–3). After the liberalizing Pension Law of 1890, an increasing number of veterans enter the system; and of these, 83 percent appear with at least one SCRT (Fogel et al., 115). The average age at which a veteran entered the system was 47.3 years and the average longevity after the first examination is 24.8 years (Fogel et al., 116–17).

Sixty-two percent of the recruits are linked to at least one census record. The lowest linkage rates were those associated with the early census years, 36 percent for 1850 and 41 percent for 1860. Such low rates result primarily from two factors: First, many of the recruits were immigrants who had not yet arrived in 1850; and second, census indexes for those early years exist only for heads of households. During those early years, most of our recruits were, of course, children living within the households of their parents. For those at risk of being found in the 1900 and 1910 census records, defined as those known not to have died prior to the census year, we achieve very high linkage rates of 82 and 70 percent (Fogel et al., 2001, 170). Detailed descriptions plus photocopy examples of each of the eight data sources, examples of the collection screens associated with each record, and a listing and identification of all variables can be found in the data user's manuals available from the CPE.

As each company of digitized records is received by the CPE at Chicago, another equally challenging and important set of procedures take place before the "Early Indicators of Later Work Levels, Disease, and Death" becomes a public-use record. I slight that part of the history not because it is less important, but because it is a different history, much of it waiting to be written as that final phase of the project is completed. It is the work at CPE that will make the efforts at BYU increasingly meaningful and accessible to scholars as yet another large number of students and their supervisors are responsible for the final cleaning and processing of the data. Further software development has been required for cleaning, testing, and coding the original data before the final step, the preparation of a public-use tape. In the past much of this work has been done by Min-Woon Song, John Kim, and Dietrich Kappe under the supervision of Dora Costa, Chris Acito, Julene Bassett, Sven Wilson, and Peter Viechnicki. The current managing director of research at the CPE is Joseph Burton. One can find further information on the project and the public-use tape at http://www.cpe.uchicago.edu.

"Collaboration," as it pertains to the EI project, involves far more than providing helpful suggestions or contributing to a joint report. Collaboration has required major commitments of time and effort by fifteen senior investigators from eight universities at almost every stage of the project, and has included not only research, but also the shared tedium of constructing, reading, and revising screens again and again. Collaboration has occupied the time of numerous graduate students who have gone on to become scholars and senior investigators in their own right, and, last but surely not least, benefitted from more than 200 students and full-time employees without whom this project literally could not have happened.

References

Center for Population Economics (CPE), Graduate School of Business, University of Chicago, and Department of Economics, Brigham Young University. 1999. *Public use tape on the aging of veterans of the Union Army: Data user's manual. Surgeons' Certificates 1860–1940.* Version S-1. Chicago: CPE.

———. 2000. *Public use tape of the aging of veterans of the Union Army: Data user's manual.* Military, Pension, and Medical Records 1820–1940, Version M-5. Chicago: CPE.

Fogel, Robert W., with Dora Costa, Charles Holmes, Matthew Kahn, Diane Lauderdale, Chulhee Lee, Louis Nguyen, Clayne Pope, Paul Rathouz, Irwin Rosenberg, Nevin Scrimshaw, Chen Song, Werner Troesken, and Sven Wilson. 2001.Early indicators of later work levels, disease, and death. Grant submitted to the National Institutes of Health. Center for Population Economics, Graduate School of Business, University of Chicago. Typescript.

Fogel, Robert W., and Stanley L. Engerman. 1974. *Time on the cross: The economics of American negro slavery.* Boston: Little, Brown.

Fogel, Robert W., Stanley L. Engerman, James Trussell, Roderick Floud, Clayne L. Pope, and Larry T. Wimmer. 1978. The economics of mortality in North America, 1650–1910: A description of a research project. *Historical Methods* 11 (2): 75–109.

Fogel, Robert W., with Michael Haines, Clayne Pope, Irwin Rosenberg, Nevin Scrimshaw, James Trussell, and Larry Wimmer. 1991. Aging of Union Army men: A longitudinal study, 1830–1940. Grant submitted to the National Institute of Health. Center for Population Economics, Graduate School of Business, University of Chicago. Typescript.

Kearl, J. R., Clayne L. Pope, and Larry T. Wimmer. 1980. Household wealth in a settlement economy: Utah, 1850–1870. *Journal of Economic History* 40 (3): 477–96.