Volume Title: Education, Income, and Human Behavior

Volume Author/Editor: F. Thomas Juster, ed.

Volume Publisher: NBER

Volume ISBN: 0-07-010068-3

Volume URL: http://www.nber.org/books/just75-1

Publication Date: 1975

Chapter Title: Appendix A: Basic Data

Chapter Author: F. Thomas Juster

Chapter URL: http://www.nber.org/chapters/c3705

Chapter pages in book: (p. 397 - 404)

Appendix A: Basic Data

by F. Thomas Juster

The chapters in this volume attempt to draw inferences about behavior from sets of observations about individuals—i.e., from samples of microdata. Many of the data sources used are relatively standard—the 1960 census 1/1,000 sample, the 1960–1961 BLS Survey of Consumer Expenditures, etc.—but a number of the samples analyzed in the volume represent nonrandom samples of individuals for whom a rich collection of information happened to be available. Much of the microdata that fit this category have been generated by researchers involved in the substantive work reported here and elsewhere. Because these samples are generally not representative of the population and are not widely known, they are described in some detail so that the reader can better understand and interpret the results reported above.

NBER-TH SAMPLE

These chapters represent one of the first uses of what will come to be, in my judgment, one of the major bodies of data available for analyzing the returns to the quantity and quality of education. The sample has an interesting history. During World War II, the Army Air Force accepted volunteers for pilot, navigator, and bombardier training programs. The volunteers, who numbered some 500,000, had to pass the aviation cadet qualifying test with a score equivalent to that of the average college sophomore. Qualifiers were then given a battery of 17 tests measuring such abilities as mathematical and reading skills, physical coordination, reaction to stress, and spatial perception. Although these tests were modified during the war, a standard set of tests was used for 75,000 men during the period from July to September 1943.

In 1955 Professors Robert Thorndike and Elizabeth Hagen, of Columbia University Teachers College, undertook a study to determine the validity of these tests in predicting subsequent vocational

success. They selected a random sample of 17,000 of the 75,000 individuals tested during the July-to-September period. A very large fraction of the 17,000 people responded to the questionnaire—roughly 70 percent. Results of the original survey are published in Thorndike and Hagen (1959).

Thorndike and Hagen have shown that there were no significant differences in test scores between the civilian respondents in 1955 and the 75,000 volunteers tested on the same battery in 1943. When compared with the United States male population aged 18 to 26 in 1943, the air cadet group was more highly educated and recorded higher test scores; also, the tested group consisted of people willing to volunteer, which may be an important source of difference between the sample and the general population.

In 1968, the NBER contacted Professor Thorndike and learned that much of the basic information collected in the 1955 survey was still in existence. The mailing addresses for the entire sample of 10,000 were also available. The Thorndike-Hagen survey contains information on 1955 earnings, earnings for jobs held between separation from the service and the survey date, and some additional data on education and family background. However, the earnings data extended only to 1955; for many sample members only a few years of regular full-time earnings data are available after the completion of formal schooling. Hence the NBER decided to conduct a resurvey of the Thorndike-Hagen respondents.

Of the 10,000 (actually 9,600) respondents for whom we had 1955 addresses and who were in the Thorndike-Hagen 1955 follow-up study, we have managed to obtain detailed information from over 5,000 people on work and earnings history, educational attainment, family background, political and social attitudes, leisure-time activities, financial situation, and so forth.

The return rate for respondents who could be located is actually close to 70 percent in the NBER follow-up. Of the original 9,600, 300 had died by 1969; an additional 1,800 could not be located despite five separate follow-up attempts and an updating of addresses via the insurance and disability files of the Veterans Administration. Our basic sample was thus more like 7,500, with close to 5,100 completed forms being received.

The most accurate data on earnings from this sample are presumably those which relate to 1955 (from the Thorndike-Hagen study) and to 1969 (the actual mailing date of the NBER follow-

up); data for other years are subject to error resulting from recall bias. Hence most of the chapters in this volume concentrate on the 1955 or, more generally, the 1969 earnings data.

SAMPLE BIAS

As indicated above, the NBER-TH sample has more education than the population as a whole, as well as higher scores on an IQ-type test. Mean education in the sample is about 16 years, with a standard deviation of two years. Mean IQ is about 112 on a test for which the population mean is 100, and the standard deviation is 15. The sample is presumed to be almost entirely white as a result of segregation policies at the time and is heavily entrepreneurial: fully 20 percent of sample respondents are self-employed. Respondents are presumably less risk-averse than the population as a whole, as reflected both by their inclusion in a volunteer program and by the heavy proportion of entrepreneurially minded individuals. Mean 1969 income is roughly $18,000. Other relevant characteristics of the sample are described in the individual chapters in which the data are used. These include the chapter on education and screening by Taubman and Wales and the chapters by Hause, Wachtel, Solmon, and Beaton.

In the chapter by Hause, which uses the NBER-TH sample along with others, a number of observations were deleted. Hause eliminated independent proprietors, doctors, lawyers, teachers, and pilots and restricted the sample to those who were aged 44 to 47 in 1969. In addition, Hause eliminated respondents reporting poor health, those whose 1969 earnings were less than $500, and those whose 1969 income was more than three standard deviations from the mean, using the log of earnings for each educational level as the basis for the distributions. The final sample size in the Hause chapter was 2,316.

The Rogers Sample

The Rogers sample is based on responses to a 1966 survey designed and carried out by D. C. Rogers. The modal group consists of Connecticut eighth graders tested for IQ in 1935. The age distribution is tight, with a standard deviation of 1.2 years. All earnings data are retrospective, obtained from the questionnaire. The 1965 figure is intended to be a reasonably precise measure of total earnings for the year. The 1960, 1955, and 1950 figures are full-time equivalent earnings based on inflated salary or wage-rate recall information.

The original sample contained 364 observations. By eliminating

those reporting zero salary or wage for any year, those not working full time in 1965, those with a severe handicap, and three extreme observations (which were more than three standard deviations from the corresponding schooling means), the final sample size in the Hause paper is reduced to 343. These observations were rejected in order to reduce the extreme heteroscedasticity of individual earnings data that makes it difficult to estimate parameters of interest in small samples.

The Project Talent Sample

The Project Talent subsample is based on the responses of some 14,000 male high school juniors who took the Project Talent battery of tests in 1960 and who indicated positive earnings in 1966. For the calculations in the Hause chapter, respondents were eliminated who were still attending school, who worked only part time in 1966, who were farmers or in the military, or who reported poor health in 1960. Nonwhites were removed for separate analysis. For each of the five educational levels in the remaining observations, the mean and standard deviation of the log of 1966 earnings were computed; observations lying more than 2.75 standard deviations beyond the mean were discarded. The first group of criteria removed individuals who were not full-time members of the civilian labor force and specific groups whose earnings were heteroscedastic or subject to special influences. The second criterion eliminated observations in the extreme tails of the log-earnings distribution. This way of treating the data further reduces heteroscedasticity and is probably a low-cost way of improving the efficiency of the estimates (relative to no adjustments). The effective sample size was 8,840.

Missing independent variables were obtained either by assignment of modal class for discrete, nonordered variables or by estimation from subregressions using a flexible program written by A. L. Norman. No observation with more than five missing independent variables was used in subsequent calculations.

The Husén Sample

The Husén data are based on male third graders in Malmö, Sweden, who were given a series of four aptitude tests in 1938. Additional information was obtained from school and social records and a 1964 questionnaire (for which the response rate exceeded 80 percent). Information on earnings was obtained for 1968, 1964, 1959, 1954, and 1949 directly from archives containing a summary of data from individual income taxes. Thus these earnings are realized

earnings rather than the full-time equivalent earnings reported in most of the other samples. No information was available on weeks worked per year or hours worked per week except for a questionnaire item that distinguished part-time and full-time workers in 1964. A rejection criterion for log of earnings exceeding 2.75 standard deviations of the corresponding mean (by schooling level) was applied and iterated once. Only respondents who answered the questionnaire were included in the analysis. The effective sample size is 455.

The "continental" schooling system, in which relatively few obtain high levels of formal schooling, prevailed when the Malmö third graders were tested.

The Consumers Union Sample

The Consumers Union panel was originated by the NBER for a study of consumer buying intentions (see Juster, 1964) and had also been used for a previous study of savings behavior (see Cagan, 1965). Panel members are above the national average in income and education. They are also meticulous, as indicated by their membership in Consumers Union and by their willingness to fill out long and complicated questionnaires. The response rates are exceptionally high for a mail survey—80 percent of the original panel responded to the first questionnaire.

The sample is certainly not representative of all United States households. Tables A-1 and A-2 compare the sample used in the Solmon chapter with all United States households, using education and income. College graduates predominate among sample households, and the income classes below $5,000 are small (although the income distribution is more similar to that of total United States households if the less-than-$3,000 income class is excluded). There

TABLE A-1 ***Comparison of education: Consumers Union sample and all United States households, 1957–1958, percent***

Education	*Consumers Union sample**	*All U.S. households*†
High school graduate or less	17.9	82.1
Some college	22.4	8.8
College graduate or more	59.7	9.2

* Excluding self-employed and not employed, incomplete questionnaires, and households with unusual gains or losses over $1,000 or with savings greater in absolute amount than 49 percent of income.

† Based on a sample survey of the labor force 18 to 64 years old in March 1957, U.S. Bureau of the Census (1959, p. 109).

TABLE A-2
Comparison of income levels: Consumers Union sample and all United States households, 1958–1959, percent

Income level	*Consumers Union sample**	*U.S. households†*	
		All	*All excluding incomes less than $3,000*
Less than $3,000	0.5	33.0	
3,000–3,999	1.7	11.1	16.6
4,000–4,999	4.7	12.4	18.5
5,000–9,999	59.2	35.1	52.4
10,000–14,999	26.3	6.4	9.6
15,000–24,999	6.3	1.6	2.4
25,000 and over	1.2	0.4	0.6

* Excluding self-employed and not employed, incomplete questionnaires, and households with unusual gains or losses over $1,000 or with savings greater in absolute amount than 49 percent of income.

† Based on a sample survey of families and unrelated individuals in 1959 (U.S. Bureau of the Census, 1960, Table 5).

are also a disproportionately large number of teachers and government workers and a relatively small number of wage earners. Most of these distributional limitations may be avoided by breaking down the sample by income, education, occupation, and other characteristics. For example, although the sample has an average savings-income ratio much higher than that for all households, the difference in saving nearly disappears when differences in income are taken into account.

The data used for the Solmon study were obtained from the last of four questionnaires sent to the Consumers Union group. A doctoral student at Columbia University, Carl Jordan, processed the basic questionnaire data and was left with 6,291 observations; he then eliminated 64 observations containing errors, leaving 6,227. For the families on the original Jordan tape, some variables useful for the current study were obtained from previous questionnaires. It was necessary to eliminate 82 of these 6,291 observations for various reasons. Finally, the Jordan tape containing 6,227 observations was merged with the tape containing the additional variables (6,209 observations), and a tape with 6,056 observations was produced; the number of observations was smaller because only families appearing on both tapes were retained, and an additional 90 observations had to be eliminated because an income variable, supposedly identical on the two tapes, did not match.

An elaborate screening program was written in order to increase

the accuracy of the savings estimates. Savings in 1959 was defined as the sum of changes in the value of assets between the end of 1958 and the end of 1959, minus the sum of changes in debt between 1958 and 1959, excluding capital gains or losses. Hence if an asset (or debt) was reported at the end of 1959 but nothing was reported for the end of 1958, the total amount of assets (debts) at the end of 1959 was presumably obtained during that year.

Since nonresponse could not be distinguished from a true zero, a screening program eliminated observations if an asset value for any component of financial and property saving appeared for 1959 but not for 1958. Exceptions were made for cases in which, despite the omission of the 1958 asset value, the respondent indicated that he had used personal records to respond—the implication being that in such a case a zero (blank) in 1958 really meant zero asset value. In other cases, a zero (blank) was presumed to indicate no response, and the observation was eliminated. A similar procedure was used if 1959 debt was reported but not 1958 debt.

This screening procedure was carried out for each component of savings; consequently, 1,818 observations were dropped, leaving 4,238. Too many observations were probably dropped, in that some people not indicating record use and having reported no assets or debts in a given category at the end of 1958 had, indeed, no 1958 asset or debt and had saved (dissaved) the entire end-1959 value during 1959. (However, if a zero appeared in both 1958 and 1959, the observation was not excluded.) The extra caution was employed in the screening procedure because it seemed better to reduce the sample size rather than run the risk of incorporating sizable errors in the savings data.

Thus the basic data tape used in the study contained 4,238 observations. Subsequently, additional observations were dropped for a number of reasons. We eliminated, for instance, those whose response to the educational-attainment question was unclear. Also eliminated were those families reporting husband's income or total family income of zero, as well as those with full savings (financial, property, and on-the-job training) equal to zero, since it was assumed that uniform zeros really meant nonresponse. The latter assumption would be invalid for any one asset or debt category, but the data used by Solmon obtained savings by aggregating across a large number of categories; it is hard to believe that a zero sum is a real number in such cases.

In nearly all cases extreme savings-income ratios reflect either

unusual financial circumstances or errors; hence families reporting absolute values of the savings-income ratio in excess of 0.5 were eliminated, reducing the sample size from 4,238 to 3,387.

Several other questions arose with regard to pruning the sample further, in particular, the question of whether subgroups with special characteristics should be eliminated or analyzed separately. Three such groups were considered: (1) independent professionals and business proprietors (456), (2) 1959 house buyers (281), and (3) those with incomes under $3,000 or over $50,000 (20). (The figures in parentheses indicate the number of families in each group.)

In general, the effects on savings of groups 1 and 2 were studied by the use of dummy variables indicating whether the family head was self-employed or had purchased a home. The third group was small enough to be ignored.

References

Cagan, P.: *The Effect of Pension Plans on Aggregate Saving: Evidence from a Sample Survey,* National Bureau of Economic Research, New York, 1965.

Juster, F. T.: *Anticipations and Purchases: An Analysis of Consumer Behavior,* Princeton University Press for National Bureau of Economic Research, Princeton, N.J., 1964.

Thorndike, Robert L., and Elizabeth P. Hagen: *Ten Thousand Careers,* John Wiley & Sons, Inc., New York, 1959.

U.S. Bureau of the Census: *Statistical Abstract of the United States: 1959* (80th ed.), Washington, 1959.

U.S. Bureau of the Census: *Current Population Reports,* ser. P-60, no. 33, Jan. 15, 1960.