MICRODATA REQUIREMENTS AND PUBLIC POLICY DESIGNS*

BY F. THOMAS JUSTER

Existing microdata sets are of limited use in the examination of important public policy questions. However, the usefulness of such sets can be improved by adding additional information. An example is the NBER extension of the Thorndike–Hagen sample. Such microdata is necessary to sort out the influence of a complex set of factors which, in this case, affect the pay-off to education. The author concludes that the benefits from the systematic collection of such data far outweigh the costs.

INTRODUCTION

Two types of questions are relevant for examination of the relation between economic microdata and public policy. First, what are the interesting questions that can be wholly or partially answered with the aid of existing microdata sets? Second, what are the questions that policy-makers really need to have answered, and can economists make significant progress in finding answers to those questions with existing microdata sets?

Although everyone has his own set of interesting and relevant research questions that bear on public policy decisions, the ones that are of special interest to me are likely to be of interest to others as well. Some important questions to which I do not believe fully satisfactory answers exist, i.e., questions in the second category above, are:

1. A hardy perennial on which a great deal of effort has been expended with only modest success: How are consumers likely to divide their income between spending and liquid asset accumulations during the next quarter? The next six months? The next year?

2. How do (and will) consumers divide their time between job-market and other types of activities and what are the factors that determine their choice?

3. How does the combination of home environment, school environment, and genetic endowment operate to produce, first, school performance, and second, job-market performance? And a subsidiary question: To what extent and under what circumstances can low levels of one or more of these sets of input factors be offset by high levels of the others?

4. What are the major socio-economic determinants of changes in birth rates over time and of differences in human capital investment per child, both over time and among families at a given point in time?

5. What are the forces that determine urban and suburban growth and decay, what are the factors that distinguish deteriorating from stable neighborhoods, and what are the short-term and long-term consequences of programs designed to improve urban environments?

* Paper prepared for the Workshop on the Use of Microdata Sets in Economic Analysis, sponsored by the NBER (October 22–23, 1970), revised April 1971.

6. How do initial perceptions, new information, and experience combine to form new perceptions for phenomena like expectations, aspirations, and socio-economic attitudes?

7. What are the objective facts of household social and economic behavior—income level and distribution, savings, assets, durable goods purchases, unemployment rates, wage and salary earnings, prices paid, and so forth?

Whether or not we have adequate answers to any of these questions is to some extent a matter of disagreement. Obviously we can say something about each. But my view of the matter is that we know a good deal less than is needed to formulate public policies that stand a fair chance of producing the results desired by the makers of policy.[1]

Let me take the last question for illustration, since all of us are accustomed to working with sets of data in which each of these variables is measured with what is assumed to represent a reasonable degree of precision. Although we certainly do not lack measures of individual or family income and family savings, I know of no reason to suppose that the available measures meet the standards of accuracy required by the sophisticated models in which they are used. The income relevant for the explanation of economic behavior is not necessarily income as reported to the Internal Revenue Service, nor income as reported to an interviewer on a survey. And the potential difficulties of measurement tend to be larger when families have multiple income sources or when they are at the upper or lower ends of the income distribution, although the problems tend to be present throughout the entire spectrum. Perhaps the best measure of our difficulty is to recognize that it is not possible to measure income and consumption, or consumption and saving, or saving and income without being able to measure all three, since they are connected by a well-known tautological relationship. If the economics profession agrees on anything, it is that we cannot measure savings accurately at the micro level (and perhaps at the macro level as well), and our frequent attempts to measure income and consumption for the same family suggest that there must be large errors in either or both since the implied estimates of savings tend to look very dubious.

For durable goods purchases, especially automobiles and houses, the benchmark measures of total purchases derived from business sources do not encourage us to suppose that the household purchase data obtained from surveys are accurate within close limits, although the problem here may be a comoination of conceptual difficulties (what is a household purchase) and relatively large sampling errors because of the "rare event" nature of such purchases. For unemployment, the problem is mainly conceptual—who is in the labor force according to some analytically satisfactory definition. For wage and salary earnings, our only regular source of data is establishment-based information on manufacturing: otherwise there is a vast and quite unsatisfactory void. For prices, one would like to do better than we now do with respect to actual transaction prices, quality adjusted measures of price, and differences in the transaction prices paid by socio-economic

[1] The reasons for this situation are unclear. One line of argument turns on the relative size of professional rewards from generating primary data as contrasted to the building of sophisticated models designed to use existing data. See F. T. Juster, "Microdata, Economic Research, and the Production of Economic Knowledge," *American Economic Review*, May 1970.

population groups. Much of this information has been traditionally obtained from establishment rather than household sources: all of it could come in principle, be obtained from households, and there are some major advantages from so doing in cases where the establishment basis has been traditional.

Existing Microdata Sets

In general terms, economic researchers interested in the analysis of household behavior have access to two broad types of microdata sets. The first are the general purpose periodic surveys conducted under government sponsorship and funding, e.g., the Current Population Survey (CPS), the Quarterly Household Survey (QHS), the Bureau of Labor Statistics Consumer Expenditure Survey (CES), and the Decennial Census. These data sets can be used to analyze some of the specific questions in the above list: CPS can be used to analyze labor force participation and income distribution questions; QHS, which is the vehicle for both the Survey of Consumer Buying Expectations (CBE) and the Survey of Residential Alterations and Repairs (SORAR), can be used to analyze questions about expectations and purchases of selected consumer durables, income, income expectations, and expenditures on home repairs and alterations. Both surveys are available with a high degree of frequency, CPS monthly and QHS quarterly. The CES and the Census have lower periodicity but are apt to contain a wider range of potentially valuable explanatory variables.

With one exception, the periodic government-sponsored surveys are confined to the measurement of objective behavioral variables and a sharply limited number at that. As a result, neither CPS nor QHS can be used for more than a narrow range of analytically interesting questions. Moreover, they have been of limited availability to outside researchers, although that situation has been improving.

The BLS expenditures surveys have been widely used for analysis of consumption and savings behavior, and their chief limitation (aside from measurement errors) lies in the fact that they are single-time cross-section data sets with no possibility of analyzing economic processes through time for identical households. The Decennial Census data have the great advantage of very large sample size, hence researchers can analyze the influence of variables like region, city size, and race for the relatively narrow range of analytically interesting variables that are obtained. As with BLS, the Census is a single-time cross-section, hence process analysis is not possible.

The second type of available microdata sets comprise special purpose ones that often have a research orientation. With few exceptions, these are also limited to single-time cross-sections. Such sets include the widely used Survey Research Center data on consumers' financial status, attitudes and expectations (as well as a whole host of SRC surveys devoted to other problems), the Survey of Economic Opportunity (SEO), the Federal Reserve Board's Survey of Financial Characteristics, the Census Bureau's Consumer Anticipations Survey (CAS), the Wisconsin and IRS tax file data sets (which in the case of the former at least were for many years distinguished more for the headaches presented to researchers trying to use them than for the insights obtained from analyzing them), and numerous special purpose surveys concerned with transportation, land use, education (e.g., Project

Talent, the Wolfle-Smith and Rogers samples, the Chicago and Bay Areas and Penn-Jersey Regional Transportation Studies, and the Detroit Transportation and Land Use Study).

I should note two special-purpose microdata sets developed by the NBER for specific research objectives. The first is the Consumers Union Panel (CUP), a highly interesting but also highly atypical set of data which contains several years' observations on a wide range of financial, economic, and demographic data for a sample ranging between six to fifteen hundred thousand households (depending upon the number of surveys one wants to use); the other is a recently completed survey of lifetime earnings and educational history for some 5,000 males, designed for use in a study of the returns to investment in education. The latter, the NBER-TH sample, is discussed below in more detail.

A major distinction between the periodic government-sponsored surveys and the much larger collection of special-purpose research oriented surveys is that the former are basically designed to provide data inputs into a governmental information system. The basic justification for CPS, for example, is that it provides measures of labor force participation, employment and unemployment which can be distributed by age, sex, race, geographic region, etc. The focus of CPS is not on providing a vehicle with which one can *explain* unemployment rates, but rather on providing a vehicle with which one can *measure* them. CBE, similarly, provides some useful data on consumer purchases and expected purchases of durables. But there is insufficient data on the survey to permit one to do much in the way of explaining either the purchases or the expectations. The basic justification for CBE is the potential usefulness of the expected purchase data in short-term forecasting.

The special-purpose surveys, in contrast, are ordinarily designed to answer a specific range of research questions, although they can often be used for other purposes as well. Thus, although the SRC periodic surveys are designed mainly to shed light on consumer spending and saving behavior, they can be used (sometimes by adding some variables, sometimes as is) to answer questions dealing with returns to formal education, the relation between income and health, and the relation between residential and work place location. In short, by the judicious inclusion of additional variables on these surveys, problems of a rather diverse sort can often be fruitfully examined.

Improving Existing Microdata Sets

The possibility of expanding the usefulness of microdata sets by adding additional information is one that is often overlooked, and which in some cases can be shown to yield very high returns. Needless to say, the possibility exists only in data sets where the identification and location of respondents has been maintained as part of the basic record. Let me illustrate what can be done in this area by examining a specific case, where modest investment in a follow-up survey promises to yield a very high return.

About two years ago, the Carnegie Commission on the Future of Higher Education made a grant to the NBER for the purpose of examining the economic benefits from higher education (benefits being understood to have either a positive

or negative sign). One part of the Carnegie project was designed to filter out the net returns to investment in higher education by attempting to measure returns standardizing for the influence of "innate ability" on earnings. Estimates of the return to educational investment abound in the literature: virtually all studies have been forced to make crude and very indirect "guestimates" of the bias due to the omission of ability from the estimated relationship. In reviewing the prospects for examining the ability-education-income relation, we discovered several bodies of data that had not been fully exploited and that could be used for this purpose. One was the often-used Wolfle–Smith sample in which IQ test scores and education were available for the 1930's with an income follow-up in 1953. Many studies have used the published tabulations from the Wolfle–Smith study, but it turned out that the worksheet tabulations, containing much more detail, were still available and could be used.

Another and more interesting discovery was the existence of a sample obtained around 1955 by Professors Thorndike and Hagen (Columbia University Teachers' College) in connection with a study of occupational choice. Thorndike and Hagen started with a sample of some 17,000 young men who had taken a series of U.S. Air Force aptitude tests in the second half of 1943. The sample was relatively homogeneous in age, all had high-school education or the equivalent when they took this series of tests, and all had been accepted by the Air Force for an Aircrew Training Program.

Thorndike and Hagen invested a large amount of resources in conducting a mail follow-up survey of this sample, focussed on an explanation of occupational choice. The follow-up was greatly facilitated by the Veterans Administration, which provided current addresses for sample members. Several mailings, Credit Bureau files, and other sources were used to maximize the response rate. Eventually, Thorndike–Hagen ended up with about 9,700 responses from the original 17,000 sample—a quite respectable response rate for surveys of this sort. Among other variables, they obtained data on post World War II education, job history and characteristics, and monthly earnings on successive jobs.

In looking over the potential usefulness of these data, it became clear that, although the 1955 earnings estimates would permit analysis of the net effects of educational attainment on earnings, standardized for the influence of the Aptitude Test Scores as a measure of innate ability, the data set would be enormously more valuable if the earnings data were further away from the completion data of formal education and if there were more details on earnings history. A substantial fraction of the TH sample went on to complete college, and many went to graduate school: hence for about half the sample, formal educational training was not completed until 1949 or after, leaving at most a six-year period of labor force experience on which to base estimates of financial returns. Fortunately, Thorndike and Hagen are meticulous researchers: not only had they retained the original cards, which they graciously made available to us, but also the 1955 address at which respondents had been located. Hence we at the NBER decided that it was worthwhile to attempt a second follow-up survey, starting out with the 1955 mailing addresses.

A follow-up survey centered around earnings history was sent to the 9,700 Thorndike–Hagen respondents. Beginning in mid-1969, six successive mailings

were made. We received excellent cooperation from the Veterans Administration, who processed about 5,000 "undeliverable" mailing addresses in an attempt to secure a more updated address. Many respondents appear to have retained their government insurance, hence were still in the active VA file.

Our success in obtaining responses to the follow-up questionnaire has been quite remarkable, in my judgment. The first two mailings were made from the initial list of 1955 addresses, and went to all 9,700 respondents; we received close to 2,500 completed forms. The next step was to take the address unknown responses, have the VA process these through their Insurance and Disability files, and send out questionnaires to respondents who showed up with a current address different from the 1955 one. After a search of both files, we sent out some 5,000 questionnaires in a third and fourth wave, from which we obtained another 2,000 responses. Our final attempt was based on the use of telephone directories for selected cities in which previously unlocatable (or reluctant) respondents lived according to our best information. Interestingly enough, we turned up close to 1,000 new addresses in this way, suggesting that a great many people are mobile only within cities provided the cities are large enough. Thus our fifth mailing went to all these new addresses. The final mailing went to all previous nonrespondents: these last two waves yielded another 650 interviews, giving a grand total of approximately 5,100.

Of the original 9,700 odd respondents, roughly 1,500 appear to have been unlocatable, in that all of our attempts to contact them met with a post office rejection as "undeliverable." Another 300 of the original 1955 respondents were deceased by 1968, leaving a total potential sample size of about 7,800. Thus the response rate was about 65 percent for those respondents whom we could locate (more precisely, who might have been located, since the Post Office is not infallible in returning "address unknown" mail).

The resulting data set is probably the most valuable single collection of information in existence for analysis of the association between ability, educational attainment, and earnings. As do all sets of data, it suffers from certain defects that cannot be remedied: the basic sample represents the upper half of the ability distribution (at least it was designed to do so by the Air Force); it includes virtually no blacks, or at least we assume that this is the case given the original sample; and it contains a large proportion of entrepreneurial individuals, as reflected by a nonprofessional self-employment rate of close to 20 percent in the responses obtained. Thus one cannot use this sample to analyze the ability-education-income relation for the lower half of the ability distribution, nor for minority groups—nor, obviously, for women!

On the other hand, one takes what one can find in this business. We do have earnings information which covers the period from first full-time job to the prime earning years—the typical respondent was in his mid-forties in 1968. We have a reasonable educational spread—about a quarter of the sample have just a high school education, another quarter have some college training but no degree, another quarter have just a college degree, and the last quarter have some graduate training. The income history is extensive for many respondents, and includes more than five job changes with beginning and ending salary for each job. We have complete information on the respondent's educational background, including the name and location of the high school attended (from which one can

make an inference about schooling quality), and the names of colleges and universities attended (from which one can obtain direct estimates of schooling quality). We also have enough information about family background to standardize for some of the influence of these variables. Finally, we have a battery of twenty test scores on the basis of which "innate ability" can probably be measured with at least as great accuracy as in any other set of data. Moreover, one can examine the relation between ability, education and earnings for those who chose self-employment rather than a salaried occupation. I know of no other data set, with the exception of the Malmö (Sweden) sample, which contained information of the sort contained in the NBER-TH sample—earnings in the prime of life rather than shortly after graduation from school, a measure of ability which predates much of the schooling received by respondents, and a complete history of educational attainment.

I cannot provide you with extensive information on the results of the analysis, although a good bit of it is now available in manuscript form. I will note simply that, as others have found, formal education certainly pays regardless of ability level, although it clearly seems to pay more for those with greater ability. Interestingly enough, ability has a very modest payoff if one is a teacher, and a very high payoff if one is self-employed but lacks a college degree. For reasons that are yet unclear, self-employed respondents with a college degree seem to do about as well if they are very bright or very dull—or anywhere in between. And it certainly pays to be a doctor or a lawyer, especially the former!

In passing, I should note that one of the totally unique features of the sample is the presence of a large number of "physical ability" measures—finger dexterity, rotary pursuit, two-hand coordination, aiming stress, etc. From the Air Force's point of view, it obviously mattered whether someone could shoot straight or indeed could shoot at all! We are currently using this information to try to determine the rate of return to finger dexterity for doctors: if we have one or two surgeons in our list, we expect to report that it is much better to be a dexterous surgeon than one who is "all thumbs." And to save possible embarrassment among workshop participants, I will refrain from reporting the relation between earnings and ability for Ph.D. holders—or even whether the relation is positive or negative!

Microdata Requirements

The appropriate scientific underpinnings for many public policy problems requires the use of microdata sets able to distinguish the influence of a complex set of determining factors whose impact on behavior is apt to be both nonlinear and interactive, different for micro units of varying socio-economic characteristics and different in the lag structure relating changes in circumstances to changes in behavior. Whether or not the micro world is as untidy as this is of course a matter to be determined empirically: the only evidence we have which bears on this question is the failure of models embodying much simpler assumption to produce consistently reliable predictions about behavior.

Moreover, not only is there likely to be a good bit of complexity with respect to the functional form of relationships for different types of micro units, but the

relevant variables are not all objectively measurable ones. One of the things that economists have learned from other social science disciplines is that the influence of objective changes in economic circumstances has its impact on behavior only after passing through a filter of subjective perceptions, expectations, plans, goals, and so forth.[2] My reading of the empirical evidence is that these types of variables cannot be ignored if one is attempting to specify a realistic behavioral model. Unfortunately, these are precisely the kinds of variables that the standard microdata sets are unlikely to contain. Hence, merging standard microdata sets may still leave a wide gap in our ability to specify an adequate behavior relationship.

What seems to me a good illustration of this point is the attempt to explain the surprisingly expansionist behavior of consumers in the aftermath of the 1968 surtax by referring to the temporary nature of the tax: many economists have argued that the tax had little influence on spending because it was viewed as temporary and thus would have had little impact on "permanent" income. My reading of the evidence is that this interpretation is incorrect, or at least seriously incomplete: not only was the behavior of consumers less surprising than many have suggested,[3] but the independent evidence from survey data is that the vast bulk of consumers did not view the tax as temporary. And it is the subjective perception of reality that matters rather than the objective facts, except insofar as the latter eventually have their influence on the former.

One can conceive of three possible ways in which the microdata requirements for public policy decisions could be met: first, existing microdata sets can be merged in an effort to broaden the information base contained in any one; second, existing microdata sets can be augmented by obtaining new information designed to fill in important gaps; third, one could decide that our present microdata sets are hopelessly inadequate and cannot be made appreciably less so, and thus consider the creation of a "basic research" microdata set.

The first two solutions can obviously achieve useful results for some problems, and are of course markedly less expensive than the third. The first, merging of existing microdata sets, has the obvious merit of maximizing the value of existing information.[4] It has two serious drawbacks, one substantive and one administrative. For most analyses of behavior, one wants a collection of variables measured for identical households rather than a synthetic variable estimated from a class of households and assigned on the basis of some common set of characteristics. Thus identical households must be merged to obtain optimally useful microdata sets. But most microdata sets have very few micro units in common, those that do may or may not have enough identifying characteristics to permit a merge, and when the latter condition is met (as with any sample that has Social Security numbers and the basic Social Security earnings file, for example) considerations of privacy constitute a formidable stumbling block.

The second solution—adding new data to existing microdata sets for the purpose of resolving a particular problem—is quite promising. Even in the most

[2] The relevance of George Katona's work at the Survey Research Center is obvious.

[3] See Arthur Okun's paper, "The Personal Tax Surcharge and Consumer Demand—1968–70" in *Brookings Papers on Economic Activity 1: 1971* (forthcoming).

[4] For discussion of the possibilities here, see "The Macro Accounts and Microdata Sets," Nancy D. Ruggles, NBER and Richard Ruggles, Yale University and NBER, American Statistical Association, *1970 Proceedings of the Business and Economic Statistics Section*, pp. 208–213.

rigidly defined such case, where one has to add information for the identical household in the original data set, there are probably a great many samples where this is possible and where the returns are substantial. I have discussed one above—the NBER-TH sample—but our experience with investigating possible data sets that could be used to examine the ability-education-income relation suggests that the addition of new information to existing data sets is often feasible. I will cite one other striking case as a further illustration of the possibilities here.

For some years now the National Institute of Health has been conducting an extensive survey on the characteristics of pregnant women and the ensuing child, starting with pre-natal data on the child and continuing throughout the first eight years of life. This sample, which numbers some 50,000 cases, contains information (of unknown reliability) on the characteristics and performance of the children, on the medical history and circumstances surrounding birth and early childhood, and on the socio-economic background of the mother. Present plans call for data gathering to be discontinued when the child reaches age eight. Quite obviously, many of the unanswered questions about the relative contributions of genetic and cultural factors to child development can be much more accurately analyzed with this basic sample than with any other—provided information on the performance of these youngsters continues to be obtained. The costs would of course be high, but considering the quality of the investment that would be salvaged (i.e., the potential scientific value of continuing the data gathering process), the rate of return might be very high.

The third possibility—creation of a "basic research" micro sample— is one that warrants serious exploration. What I have in mind is a large scale sample that represents a continuous microcosm of the population, changing only with births, deaths, and new family formation. Sample members would be compensated for their time, and the compensation should be enough to overcome the distaste for invasion of privacy that one would inevitably find in some (small) fraction of the population. Standards of accuracy would be high, access to records would be part of the quid pro quo for compensation, and consistency checks would be the accepted technique for insuring accuracy. For a number of reasons the responsible organization ought to be nongovernmental, with government users having the same rights of access and constraints as with any other user. The information base would be sufficient to permit economists—and probably other social scientists as well—to examine a wide range of significant problems, although that would only be the case after the sample had been in existence for a number of years. In principle, the panel would be viewed as having an infinite life span.

Let me note some critical substantive problems before commenting on the major problem—which is, of course, the cost. The usefulness of such a research sample clearly depends on its being representative of the population. Is it possible to have a continuous panel that remains representative? To the extent that the problem is mobility, the answer is yes, but at a very high cost. To the extent that the problem is panel bias other than that arising from mobility, the answer is unclear. I would guess that true panel bias is a function of the interview frequency rather than the total number of interviews: that is, panel bias is likely to be serious if one interviews people every month for a year, but may not be serious if one interviews people every year for twelve years. But that is a matter for empirical

determination, not *a priori* theorizing. It does imply that a continuous panel needs to be monitored and checked continually for bias, and that procedures for replacement may have to be worked out.

Another potential problem is the ability of economists and other social scientists to handle the sheer bulk and complexity of the implied data file in such a way as to insure reasonably rapid access to the data. Ten years ago that problem would have been insurmountable; an attempt to implement a grandiose scheme like this would probably have resulted in an equally grandiose and expensive fiasco. I would judge from the papers presented at this workshop that economists could manage this problem now, and could do so in a reasonably efficient way.

What would be gained from the establishment of a research panel? The minimum possible gain would be a flow of basic information about actual behavior characterized by a very high standard of accuracy. Much of the information could be used to feed into a system of social and economic accounts for the household sector—which represents no small gain, in my view. The maximum possible gain would be the eventual possibility of being able to predict, within tolerable error limits, some of the consequences (both immediate and long-term) of alternative public policies.

Finally, what about the costs: In a word, high. Just how high is a matter of conjecture, but I suggest that a ball park number is a factor of 10 above what we tend to think of as an expensive set of microdata. My guess is that we are talking about $5–10 million per year for a minimum ten year period. Economists are not accustomed to think in terms of these orders of magnitude, but I suggest we take a page from the books of our brethren in the physical sciences. Particle accelerators and astronomical observatories that cost in the tens of millions are not uncommon, and they are judged to be worth their cost. Yet one is simply a way to generate observations, and the other is a way to measure observable physical phenomena.

A note of caution should be added. Physical scientists have justified their demands for costly research tools by citing their ability to produce results, i.e., to predict events that have not yet taken place and thus to demonstrate true scientific competence. Economists might be able to persuade someone to give them an initial opportunity to do likewise, but they would have to show some output after a reasonable trial period to justify a continued effort.

National Bureau of Economic Research