Tax Modelling and the Policy Environment of the 1990's

By Ralph B. Bristol, Jr.*

Questions posed by policymakers to their staffs obey a form of Parkinson's Law: However great the ability of the staff to provide answers, the questions will always go just a little bit farther. In seeking to predict the kinds of questions policymakers will ask, therefore, we have only to predict the ability of staff to answer questions, and then add 10 percent. This paper will examine some half-dozen areas of such supply and demand for information, and attempt some predictions. It will focus primarily on the demands of policymakers and the shortcomings of our existing models and data sources, leaving for subsequent papers the task of predicting just how we will change our systems to meet these new demands and overcome these shortcomings. The six areas to be discussed are: 1. longitudinality—changing from a focus on annual data to a longer time perspective; 2. timeliness—shorter deadlines for data availability; 3. different units of analysis—families or households rather than just tax returns; 4. on-line data accessibility; 5. matching or linking tax records to different data sources; and 6. public access and privacy/disclosure problems.

LONGITUDINALITY

Tax policies almost always refer to a particular time unit, namely a year. The specific demarcation may change—fiscal years are permitted, and quite common for corporations—but almost all legislation refers to a specific 12-month accounting period. The exceptions to this have generally consisted of providing for "carryovers" which derive from limits placed on tax provisions. Thus, when a ceiling was placed on the amount of long-term capital losses which could be deducted from ordinary income, a "carryover" was provided so that any excess losses over the limit could be claimed in subsequent years. Similarly, limits on the amount of charitable contributions which can be deducted from taxable income in any given year are combined with an allowance for carryovers to future years. In the case of business "net operating losses" (NOL's), both carryovers and carrybacks are provided for.

In recent years there have been more and more tax provisions which involve a further widening of the tax-period horizon. "Once-in-a-lifetime" limits were introduced for the exclusion of (some) capital gains resulting from the sale of a personal residence. When energy credits were introduced in 1978, a cumulative ceiling was provided for. Gift taxes come into force only after a lifetime exclusion has been exhausted. Many tax credits involve longitudinality: non-refundable ones may permit any unused credit to be carried forward to another year. Finally, income averaging itself, first introduced in 1964, explicitly bases tax calculations on a multi-year period, although the restrictions are such that this provision of law has never been used by more than a tiny minority of taxpayers.

It seems clear that the future will see more and more tax provisions involving longitudinality. One-time only and cumulative-limit provisions seem popular with legislators—such limits appear to reduce revenue losses and prevent tax abuses. Clearly, then, if we are to be in a position to provide policy advice and revenue estimates, we must have longitudinal information—observations on identical tax units over a period of time longer than just the traditional 12 months.

At the same time that tax legislation has been lengthening the relevant time period for observations, developments in the field of economics have also been calling for a change in focus. Not that this has happened overnight—the basic theoretical work by Ando-Modigliani and Friedman was done in the 1950's—but the lack of empirical knowledge has made it possible to ignore those advances until fairly recently. Today, it is unmistakable that either a taxpayer's stage in his life-cycle, or his permanent income is a much more relevant measure of his economic power, his economic well-being, than the mere flow of his receipts (or accruals) over any particular 12-month period.

There are strong life-cycle effects on the "tax life" of an individual. When young, an individual appears in the tax system only as an exemption for his parents. Even when he enters the labor force, exemption levels and filing requirements may keep him off the tax rolls. Then for a while he may enjoy a tax "double exemption" status, claiming his own exemption while his parents also continue to do so. Later in life, home ownership and the acquisition of consumer durables through credit purchases introduce an individual to new areas of the tax code. Finally, retirement with taxable pensions, nontaxable social security benefits, and the drawing-down of assets make the taxpayer a member of yet another special-interest group of taxpayers.

Future tax changes may make longitudinality more important or less important. If we move toward flat-rate taxes,

* Reprinted from the *Multi-National Tax Modelling Proceedings*, Revenue Canada Taxation, September 1985, with the permission of the author. Dr. Bristol currently teaches at the University of New Hampshire. He was a senior staff economist in the Office of Tax Analysis, U.S. Department of the Treasury, at the time this article was written.

longitudinality will be much less important. Highly progressive rates place a harsh penalty on incomes which fluctuate from year to year—lowered tax rates in "bad" years do not offset the high rates paid in "good" years, so the average effective tax rate over time is higher than if the same total income had been received "smoothly" throughout the period. If we move closer to a proportional tax, this penalty on varying incomes will diminish, but will it still be important? Our knowledge of just how people's incomes fluctuate over time is still quite rudimentary, so we just do not know the magnitude of this effect.

On the other hand, if future tax changes are in the direction of consumption-based taxes or value-added taxes, longitudinal data will become critically important. Indeed, it is the personal opinion of the author, that the United States will never implement a consumption tax, merely because of the difficulties in achieving equity across age groups. Specifically, the typical economic unit consumes less that it earns *before* retirement, and then spends more than it earns *after* retirement. It would be politically disastrous to say to people of middle-age or older, "All right, now that you've paid income tax all your life, from now on, instead of taxing your *income* (which is going down), we'll tax your *spending* (which is staying up)!" Well, regardless of whether everyone agrees with this political forecast, one can see the importance of having longitudinal observations when analyzing fundamental tax policy issues such as whether income or consumption is the appropriate base for taxation.

Even apart from future tax policy changes, we need longitudinal data for analyzing our *present* tax system. For example, consider the taxation of capital gains. We really do not know much about the distribution of realizations over time for identical units. It seems to me that the policy implications are quite different if, on the one hand, most people realize gains at only a few points in their lives (e.g., selling a home or a business, or cashing in assets post-retirement) or if, on the other hand, they typically realize gains every single year (e.g., stock market speculators). The empirical answer is, of course, that some taxpayers fall in first category, and others in the second. Their relative magnitudes, however, are unknown, and no amount of data analysis of single-shot, one-year tax returns can shed any light on this matter. Thus, whether to analyze existing tax systems or to be ready to analyze future tax systems, it is imperative that we acquire more longitudinal information on taxpayers.

TIMELINESS

Policymakers are never satisfied with the responsiveness of our data systems. They resent being presented with outdated data. At this season, the closing months of Calendar Year 1985, they begin asking us questions concerning the impact of tax policies which became effective January 1st. After all, they reason, the policies have been in effect for almost three-quarters of a year, so "what's happenin', baby?" First, we have to tell them that we do not have the slightest idea of what is happening in 1985, then we have to break the news to them, as gently as possible, that we do not even have any (microdata) information on Calendar Year 1984! Such staff responses are generally met with incredulity by policymakers newly arrived in the Government. We must carefully explain to them the time-sequence of events.

In the United States, individuals must file by April 15th. However, for anyone requesting it, there is an automatic four-month extension period, running until August 15th. (As an aside, which returns do you think request that extension, low-income returns with simple, 1040A schedules? Or are they more likely to be complicated returns, filled out by CPA's, attorneys and other tax practitioners, replete with extra schedules and "accompanying documentation"?) Once all the returns—some 100 million of them—come into the IRS Service Centers, what should be done with them? That is, what activities should be given the highest priority? Not surprisingly, statistical analysis is "low on the totem pole". The highest priority is given to money—both coming in and going out. After all, the cost of a one-day delay in handling a $100 million can be measured very precisely—at 12-percent interest, it amounts to over $30,000. Not surprisingly, Service Center directors are encouraged and exhorted (not to say bullied) to get those checks out of the envelopes and into the banks. Of almost equal importance is the matter of refunds (I would remind you that over 70 million returns show an overpayment of tax which means that the filers are entitled to refunds). When taxpayers file returns to get their money back from the Government, they tend to be impatient. Thus, again not surprisingly, Service Center directors are given deadlines for "clearing their books"—getting refund checks out to taxpayers before the complaining letters start coming in.

So where does this leave Statistics? Just where it always is, an orphan. Unfortunately, the informal but effective motto of the Internal Revenue Service (IRS) is: "Our job is to collect *taxes*, not tax statistics!" The collection of data for what is often referred to as "purely statistical purposes" is regarded as an unnecessary cost, a burden imposed on the normal and ordinary activities of the taxcollecting system. Thus, only after the money has been collected and the refund checks mailed, will administrators consider diverting some of their resources to what is viewed as the sterile and unproductive task of what I call "policy statistics." (To be sure, some statistics are kept and treasured by the revenue processing system itself. These tend to be production statistics such as number of returns processed, amount of dollars received, number of refund checks written, and number of overtime hours expended. These

statistics are seen as "helping us do our job." Statistics for tax policy are not viewed in such a kindly light.)

The statistics generated by the revenue processing system tend not to be very useful for policy analysis. Dollars of revenue, dollars of refunds, and number of returns are of some interest, of course, but policy analysis is more likely to focus on such matters as why the tax was paid, what the underlying income was, just which tax provisions (or tax schedules) applied, and so forth. To obtain this information usually requires a second handling of the returns. (As an aside, many people seem to think that once a return is mailed to the IRS, it automatically becomes part of some gigantic data bank, immediately accessible to any tax analyst [or snoop] with a computer terminal).

In actuality, of course, every single bit of information must be key-punched by hand before it can be used for any purpose at all. (Optical readers are changing this). Furthermore, information in this raw form cannot be used without further processing. The taxpayer may have put numbers on the wrong line or he may have made a mistake in arithmetic. To "clean up" the data, returns must go through a rather elaborate process of data editing, verification, and consistency checking. Only after all this has been done, are the data finally in a form suitable for analysis.

In the United States, the Internal Revenue Service's preliminary or "early cut-off" Statistics of Income (SOI) for individuals represents returns processed by the end of September. These become available to policymakers by early December, while the "final" statistics take another six months. Incidentally, when we say "final", we do not mean that further corrections are unneeded; we just mean that no new returns will be added to the sample and no new information will be added to the return data. Further data improvement must come from information internal to the return itself.

All this, naturally, can be explained to a policymaker (if you are lucky enough to have one willing to listen), but after you have gone through such an elaborate explanation, what is his response likely to be?

"Yes, yes. But can't you do something to speed things up a little?"

UNIT OF ANALYSIS

The traditional unit of analysis for tax policy studies is, not surprisingly, the tax return. This is our basic data source—it requires practically no estimation or imputation, it changes appropriately when the tax law changes, and its definitions conform to the terms used in legislation. Unfortunately, it is accurate to state that this is *never* the true focus of interest. Consider matters of equity or tax burden. Identifying "low-income" groups on the basis of tax return information is quite misleading. First there is the question of whether the income reported is that of an individual, a couple, or a large family. Nor is family size the only problem. For example, consider the bottom income group in the 1981 U.S. Statistics of Income, those 18 million returns with Adjusted Gross Income (AGI) under $5,000. These surely appear to be below anyone's "poverty line." Yet closer analysis (using information not available on the tax returns themselves) discloses the fact that 40 percent of these returns are filled out by taxpayers under the age of 20! Undoubtedly some of these returns represent taxpayers mired in poverty, but there can be little doubt that the majority of these teenagers are not true "economic units" whose welfare should be of concern, but rather are a subset of some *other* economic unit (family) whose income and economic status may be very different from what appears on that teenager's tax return.

Even if our concern is not with equity but rather with economic behavior and efficiency, what we want to examine is not tax returns, but some other unit. The appropriate or relevant unit must be defined in terms of some kind of *behavior.* Census and survey workers have of necessity developed many alternative concepts which prove useful in different situations: households, families, spending units, dwelling units. What these definitions have in common is some notion of *sharing* or *pooling*: individuals will pool their incomes or their spending or perhaps merely their housing bills. Which of these economic units is most appropriate depends upon the particular policy analysis we are conducting.

Attempting to combine, or perhaps I should say re-align, these 100 million returns into something like 90 million "families" or "households" turns out to be quite a problem. Apart from a mailing address, there is usually nothing on a return which provides an indication (in computer terminology, a "pointer") as to which other return or returns this individual should be combined with. The most common household will be represented by a joint return, filled out by husband and wife, with perhaps other dependents, usually their children. In addition to these "standard" family members, there may be additional income earners (or consumption spenders) who may themselves be tax return filers, or they may be non-filers, or they may be non-filers who don't show up at all (except perhaps as claimed dependents) on tax records. Examples of such income earners are children with part-time jobs whose income is below the filing requirement, and elderly people (usually relatives) living in the home who may appear on the tax rolls because their income is tax-exempt (social security recipients are the most common example).

In an attempt to overcome these data shortcomings, the Treasury Department's Office of Tax Analysis has devel-

oped its Merge Model. This represents a combination of 50,000 sample households interviewed in the Census Bureau's Current Population Survey (CPS) and 80,000 sample tax returns developed as the Statistics of Income Tax Model [TM]. From the 1981 Model, for example, we first extrapolated the 1981 SOI to 1983 levels, using a special algorithm which has been developed at the Treasury Department. Twenty-eight targets are picked, involving distributions across income classes; numbers of returns; and dollar amounts of different types of income, such as wages, interest, rents, etc. The returns in the sample are then reweighted so that they add up, in the aggregate, to the pre-specified targets. This is done in such a way that the change in the individual weights is minimized.

Next, we align the CPS and SOI files; this involves assuring that the two files have the same number of units filing tax returns and the same number of units reporting each type of income. We apply a tax-calculator to the CPS and then adjust the resulting discrepancies. For example, almost everyone reporting rental income on the CPS reports positive net rental income, whereas tax filers in the SOI were twice as likely to report negative rental income as positive income.

Once the two files are calibrated so that they appear to be representing the same population, they are merged. This is done through the use of a special transportation algorithm. This uses a penalty function consisting of ten variables such as family size, wage income, property income and home ownership. The algorithm links together families who are as "alike" as possible. Weights frequently have to be split in this process, and the result is a merge file of some 200,000 records.

The shortcomings of the Merge Model are obvious: it is a "soft" or "statistical" match, and we cannot be certain that the family units and tax returns are, in fact, correctly matched. The two samples are not very well aligned: the CPS has many low-income units but not many high-income ones, while the SOI is exactly the opposite, excluding entirely units with income below a certain level ("nonfilers"), and being quite rich in high-income returns. (In the 1981 merge, 300 "returns" in the highest income class of the CPS had to be matched or linked with 33,714 such returns in the SOI. At the other end of the income distribution, 4,277 low-income SOI records had to be matched with 17,647 CPS records.)

ON-LINE ACCESSIBILITY

There is an increasing demand for on-line accessibility of data by policymakers. In part a product of the computer age and a result of the rapid multiplication of desk-top terminals, this demand represents a combination of Section 2, above (shorter deadlines) and Section 6, below (access).

Simulation models have boomed in popularity since the development of computers, and nowhere has this been clearer than in the area of fiscal policy analysis. Macroeconomic models were first in the field, as computers made it possible to manipulate first handfuls, then dozens, and now hundreds of econometric equations. The ability to "play God" and answer "What if?" questions is irresistible and, nowadays, widespread.

Following close on the heels of the macro-models, came the micro- or cross-section models. While less demanding in terms of mathematical and econometric complexity, these are much more demanding in terms of computational power and in terms of data. When you are simulating the behavior of individual economic units, whether business firms or personal taxpayers (I shall be referring primarily to the latter), you need a huge number of them (to take account of the random variability of their behavior). It was not until the early 1960's that such models of taxpayers were used by the U.S. Treasury, and then the sample sizes were of the order of 10,000 returns. One computer simulation might take several hours of clock time. Tax analysts today have difficulty imagining what it was like to examine proposed tax programs in the absence of simulations by what is now succinctly known as the Tax Model. (Incidentally, we now work with sample sizes of about 75,000, and simulation runs require less than 15 minutes of computer time. Clock time is another story, which I will not go into.)

To date, these Tax Model simulations have been the exclusive property of the computer people, the "high priests" of the operation. They have the ability, and the responsibility, of translating tax policy questions into "simulateable" tax policies. By this I mean that they must not only convert things into a language that the computer can understand, but must also filter out inconsistent policies and be alert to all of the sophisticated and easily overlooked intricate interactions of the tax code (e.g., Does this new provision change any taxpayers from itemizers (of deductions) to non-itemizers? How does it affect the "minimum tax" or the "alternative minimum tax?" Does it affect any taxpayer's "excess investment interest?" How does it change if a taxpayer is income-averaging?) Computer languages are getting increasingly user-friendly, but the same cannot be said of tax laws and regulations!

A modern tax policymaker, with a computer terminal sitting on his desk, wants to be able to make Tax Model runs himself. He is not going to be terribly patient about listening to qualifications and caveats about what the model is and is not designed to do. He wants an answer! In developing our tax models, we must take into account not only the shortcomings of our statistical data, but also the problems posed by this new generation of computer operators.

In some cases, the demand for on-line accessibility means displaying specific, identifiable tax returns (e.g.,

what sort of taxes has Chrysler paid over the last five years? what is the loss carryover position of the top five steel companies?). From a staff point of view, such demands raise nightmares in terms of privacy problems (e.g., does this finance minister realize how explosive this information might be?), in terms of data management problems (e.g., our corporate file may be sorted by years, not by company name), in terms of data comparability (e.g., how do you warn a user that, because of a merger, this year's data are not comparable with those of previous years?), in terms of timeliness (e.g., this company's fiscal year is such that it will not even have to file for another six months), and in terms of completeness (e.g., the company did not even fill out that particular schedule).

The availability of tax statistics to what might be termed "non-tax statisticians" highlights all of the weaknesses and problems of our tax statistics—the dirty little secrets like missing returns, taxpayer errors, incorrect edits, and faulty imputations. Tax policymakers (at least in the United States) are political appointees whose background is invariably devoid of statistical training. Once on the job, their time is very limited, and it just is not realistic to expect them to become educated and sophisticated concerning tax statistics. Historically, such considerations led to the development of the "permanent civil service" structure. This was all well and good in an earlier, perhaps more gracious, age, but it is not clear that mandarins and on-line, real-time computers can coexist. We (as mandarins) may deplore some of the things policymakers do with our data, but we have got to realize that they are going to grab the numbers, and it is yet another challenge, another set of demands, we must bear in mind in developing our models and our data systems.

LINKING AND MATCHING RECORDS

Perhaps the first question we should ask about linking or matching records is, why do it at all? The answer is that the instrument we are working with—the tax return—lacks certain information that we need. We lack some information merely because of missing responses or missing returns, but my focus here is not on that kind of omission, but rather the complete absence of some (non-tax) information from returns. Tax returns, after all, concern taxes, and information which does not directly affect an individual's tax liability will generally be omitted from the return, beyond a bare minimum of identifying material essential to processing the return. There is tremendous pressure in the United States for even greater shortening of the tax form: both the Paperwork Reduction Act of 1980 and the general concern with Government intrusion on privacy mean that tax returns in the future are liable to contain less, not more, information than they do today. In brief, it will get worse before it gets better.

What other information do we need, beyond what is included in the tax return? We have already discussed, in Sections 1 and 3 above, the need for longitudinal information and information on different analysis units. In general, any "new" tax proposals will, almost without exception, involve new variables, ones we have not been observing on tax returns. In order that we be able to discuss and analyze such proposals, we must be prepared ahead of time.

How can we obtain additional information on taxpayers? For one thing, we cannot approach them directly, say by conducting sample survey interviews, because in the United States that is legally considered to be in the nature of a tax audit. Because we possess certain identifying information on each taxpayer (most notably his social security number, name, and address), we can learn certain things from other (governmental) sources. For example, in the United States, thanks to the cooperation of the Social Security Administration, we have been able to append to our Tax Model sample information on each taxpayer's (and spouse's) age, sex, and social security benefit status. This has proved invaluable to us, but it does involve overcoming some reluctance on the part of the Social Security Administration. After all, it is not their mission to provide statistics on taxpayers, and they have severe obligations to protect the privacy of social security participants. (We have been able to obtain their cooperation, because [a] we have been able to demonstrate that the information available from such a matched file is useful for administering the Social Security Act itself, and [b] we have developed very careful security provisions for the data, ensuring they will be used for policy analysis, not for tax law enforcement).

Most of the information we seek is demographic or economic, and if you stop and think, IRS and the Social Security Administration are just about the only agencies which obtain this information on individuals (as opposed to aggregated data). The one clear exception is the Census Bureau, but this agency has its own unique rules and regulations concerning disclosure of data and does not appear to be a promising source. Other Government agencies might be sources for information on such variables as unemployment compensation, fringe benefits, retirement plans (both costs and benefits), and various vital statistics, but we have not yet attempted matching for any of these.

The alternatives to exact matches are "statistical" or "soft" matches. These involve linking our data base with another data base, usually a survey conducted for some other purpose. We can thus take its observations on the variables we care about, and use them to impute values to our sample of tax returns. To take an example, suppose we wanted to include on our file information on the number of hours worked, and suppose we had available a labor-force survey with this variable for a sample representing the same

population as our tax file. One extreme (and undesirable) procedure would be to calculate the *average* number of hours worked for everyone in the labor force survey and "impute" or assign that average value to every one of our taxpayers. Thinking about such a procedure reminds us that micro-models, by definition, are *not* just interested in averages or other measures of central tendency, but rather are interested in the full dispersion and heterogeneity of individuals. We can obviously do better in this instance than merely imputing the average value. If we have information on age of taxpayers, for example, we know that the very young and the very old probably should have zero hours assigned to them. We would probably want to assign working hours only to returns which reported non-zero wage income, and so forth. In fact, what we would want to do is take *all* the observed variables common to both the survey and our tax file, and use these variables to "pair off" or match the two sets of units and thus make the most appropriate assignments.

This entire area of matching files is a fairly new and somewhat controversial field. Some people feel that such linking is not worthwhile, that you cannot "get something for nothing." The necessary assumptions as to just what is independent of what, and just what is correlated with what, are a little tricky and sometimes obscure. Very little work has been done validating match procedures, primarily because of the difficulties involved in defining a "good" or "successful" match.

The Treasury Department, on a biennial basis, matches its Statistics of Income sample with the Current Population Survey sample, as described above in the section on unit of analysis. This is done by using a very sophisticated transportation algorithm developed for the Office of Tax Analysis. Is it successful? All we can say is that it seems to give reasonable results, that it gives useful results, and that there are no apparent flaws in the procedure. I wish I could be a little stronger in its defense, but we simply do not yet know enough about validation.

To summarize, because we will need more and more information about taxpayers in the future, and because we will not be allowed to burden them with the questions necessary to elicit this information, we will have to turn to other cross-section information sources. Unless we are willing to give up the richness of micro-models, the only way we can synthesize information from multiple sources is by some technique of matching or linking units. We are doing a lot of this now, but we will have to do even more in the future, and we do not yet know enough about it.

PUBLIC ACCESS AND DISCLOSURE

Individual tax returns have generally been protected from public scrutiny in the United States. Strict procedures now govern the handling of returns and of any computer tapes containing tax return information. Even if specific identifiers (name, address, social security number) are removed, the return is still considered to be confidential. As long as the only public release of tax statistics was the published Statistics of Income series, there were few problems of disclosure. Certain standards had to be met, such as the "Rule of 3," meaning that no cell could be published for data above the state level if there were not at least three returns in the cell. Because most of the statistics were so aggregated, however, disclosure issues were seldom a problem.

With the development of microeconomic modelling, however, new problems arose. As long as computers could not handle more than a few hundred variables, the sort of crude cross-tabulations available in the SOI volumes were adequate. Once researchers could handle thousands of records, the pressures grew for the release of "scrubbed" or "sanitized" tax return (microdata) information. The distribution of taxpayers across variables, rather than just their average or typical values, became of interest to analysts outside the Government as well as inside. How could the interests of these researchers be reconciled with the need for confidentiality of individual returns?

The practical compromise that has been made is that, as long as a sample SOI return has a high enough weight, that is, there are a lot of such returns in the country, there is no problem in releasing tax return data as long as specific identifiers have been removed. The implicit assumption is that even if you study such a return, there are enough other returns in the population just like that one that it would be impossible to make a positive identification. The problem arises with low-weight sample returns. Specifically, all returns above a particular income level (typically $200,000, but varying by year) are sampled at a 100-percent rate, i.e., if an individual has a high enough income, you can be sure his return is included in the sample. This creates the possibility of "hunting" for a specific return.

Considering all the information on tax returns such as types of income and types of deductions, it would seem feasible to conduct such a hunt. It should be noted that disclosure implies not just the fact that information from a tax return is made public, but also whether or not a person even filed a tax return. In fact, a few years ago, a newspaperman claimed to have made some positive identifications of individuals on the basis of the publicly-released SOI. (There is some doubt as to whether he actually succeeded.) The SOI Division has been conducting a number of research studies in this area.

One study focused on the feasibility of identifying an individual's tax return on the basis of publicly-available information. Some business publications publish the sala-

ries and bonuses of top corporation executives—could this be used to spot such individuals? It turns out that deferred compensation, stock options, and other forms of income manipulation (mostly motivated, it might be noted, by the-desire for tax avoidance) introduce so much "noise" into the translation of compensation into taxable income, that it is almost impossible to recognize individuals on the basis of their tax returns.

The study also used other publicly-available information in its tests, such as court-ordered alimony and child-support payments, and real estate tax bills. Again, the complexity of the tax returns filed by high-income individuals tended to obscure much of the information (e.g., itemized deductions for real estate taxes would typically cover multiple holdings). However, there did seem to be a chance of positive identification using this information.

How, then, can information be made available in public-use files which is sufficiently detailed to be of use to researchers, yet still not permit invasions of privacy? Several measures have been proposed and utilized. Some information is just completely erased, such as taxpayer's name, address, and social security number. Some data fields (that is, variables) may be rounded. Thus, if particular income sources and itemized deductions are rounded to the nearest thousand dollars, this will make it much harder to identify individuals, but may still provide rich enough detail for researchers. Another technique is that of "collapsing cells:" all returns in a particular cell (i.e., sharing certain characteristic), are added together and each one is then assigned the *average* value of everyone in that cell. This approach is not popular with researchers because it destroys just the variablity which micro-models seek to exploit. A variant of this is the "moving average" approach in which the collapsed cell changes for each individual. Thus, for no individual is the correct value shown, but the correct overall average and most of the variability are retained. Finally, individual values may be obscured by adding random noise with an expected value of zero, thus again obscuring the individual's value, but keeping the correct overall average and as much variability as desired. All of these techniques raise the possibility of creating internal inconsistencies within the return, that is, components of income may no longer add up to total income, and itemized deductions may not equal the sum of the parts.

Research in this area is being actively pursued by the Internal Revenue Service as well as by other agencies which create public use files (e.g., the Commerce Department and the Department of Health and Human Services). The task is rather difficult, because it is like trying to develop an "unbreakable" cryptographic code: there is no way you can "prove" it is invulnerable. All you can do is show that various techniques of breaking it do not work, but there is no way of proving that there does not exist some other, as yet unheard of, technique that will do the job. This rather unsatisfying conclusion is the best we can hope for in the area of guaranteeing confidentiality.

To repeat what was said before, we can, if we wish, assure against disclosure, but only by locking up all the returns and not creating any public use files at all. However, unless you subscribe to the belief that all wisdom resides in the Government, it is vital to get these data out to interested researchers. The appeal of micro-econometric modelling and cross section studies is the richness of detail, the scope of variability across thousands of economic units. For better or worse, taxes are a vital part of our economic lives, and it is important that we learn more about just what they do to us and to our economy. We must not degrade our data bases any more than absolutely necessary to protect taxpayer privacy.

In the final analysis, the important factor is public confidence. As long as the Government promises people it will protect their tax returns from disclosure as it did in 1974 and 1976, by enacting the Privacy and the Tax Reform Acts, respectively, it must keep that promise. Even though 1984 (the calendar year, that is, is now behind us, there remains a very real fear of "big brother" type abuse of large computerized data banks, and the IRS is high on the list of the most feared governmental institutions. However, we should note that the honesty and resistance of IRS in standing up to political pressures and keeping tax returns inviolable was one of the few bright spots in Watergate.