EXACT MATCHING OF MICRO DATA SETS IN SOCIAL RESEARCH: BENEFITS AND PROBLEMS

Robert Boruch, Northwestern University
Ernst Stromsdorfer, Washington State University

1. INTRODUCTION

The first objective here is to review some applied social research projects that have benefited from exact matching. The examples are merely illustrative but stem from a variety of disciplines.

The second objective is to discuss the negative aspects of matching. In particular, our argument is that, by espousing the opportunity to match too ardently, we may constrain or misdirect our ability to respond to other research issues and problems. An issue of special interest here is obtaining unbiased estimates of the effects of manpower projects.

The idea of matching records in the interest of science has a long pedigree. For instance, R.A. Fisher lectured at a Zurich public health congress in 1929, arguing the usefulness of public records supplemented by (and presumably linked with) family data, in human genetics research (Box, 1978, p. 237). Earlier, Alexander Graham Bell exploited geneological records, administrative records on marriages, census results and others, apparently linking some sources, to sustain his familial studies of deafness (Bruce, 1973; Bell, 1906).

2. HOW AND WHY HAS MATCHING BEEN HELPFUL

The fundamental reasons that matching has been useful do not differ appreciably from those implied by the above examples. Nor do the reasons differ much across the social and behavioral sciences. The following illustrations are taken from Boruch and Cecil (1979); unless otherwise noted, specific references are given there.

2.1 Matching to Understand Phenomena and Avoid Egregious Error

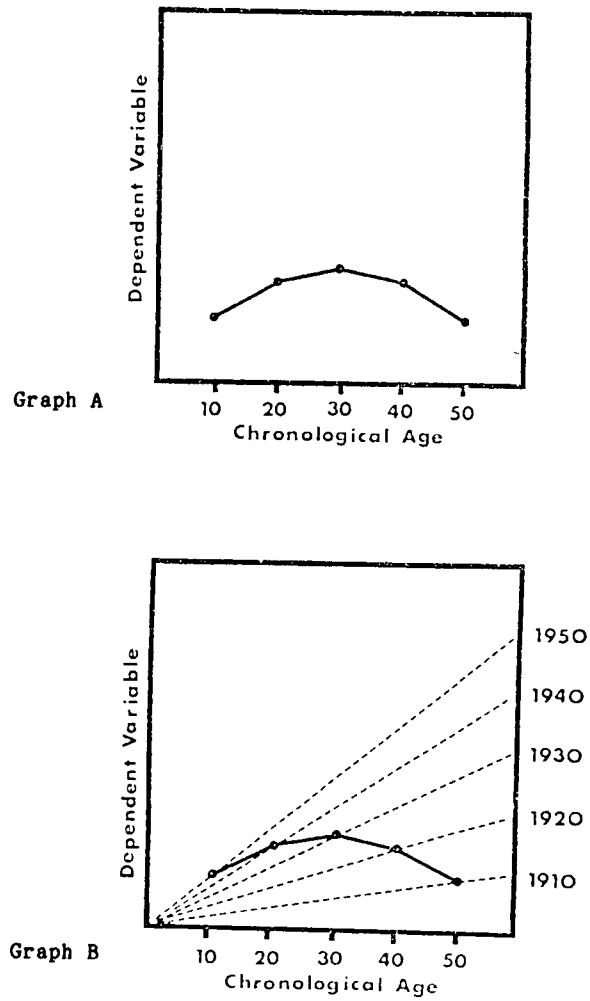
In psychology, for example, graphs of the sort used in Figure 1A were commonly used during the 1940's and 50's to describe the gradual increase in IQ with age, an IQ plateau and gradual decrease in IQ with age. The data are based on cross-sectional surveys.

The ability to match, as in linking individuals' records obtained at one point in time to those collected at another to generate longitudinal files, yielded an entirely different picture of behavior. This, given in Figure 1B, tells us that earlier declines in IQ are an artifact of cross-sectional studies and that cohort differences are important and account for the misleading interpretations of the earlier data.

Lest you think the example confined to a quantitatively naive discipline, consider an economic example. Table 1, based on simple cross-sectional surveys, suggests that a graph similar to Type A is appropriate for earnings data as well as IQ data. Such earnings data were commonly used during the 60's to describe increases, plateau, and gradual decline in income. Table 2 gives cohort earnings obtained in longitudinal surveys, matching on individuals. It shows a different picture, one that is less dramatic and more similar to the Type B figure.

Studies that try to separate genetic and environmental influences in schizophrenia are bound to be more controversial. But they are important and worth pursuing... So, for example,

Figure 1. Confounding of Age and Cohort Differences in Cross-sectional Research.



From: Boruch, R.F., and Cecil, J.S. Assuring the Confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press, 1979.

Table 1.--Estimates of Mean Annual Income in Dollars for Men Aged 25-64
(Data is based on independent samples taken in 1947, 1948, and 1949.)

Year	Age			
	25-34	35-44	45-54	55-64
1947	2,704	3,344	3,329	2,795
1948	2,898	3,508	3,378	2,946
1949	2,842	3,281	3,331	2,777

From: Boruch, R.F., and Cecil, J.S. Assuring the Confidentiality of Social Research Data Philadelphia: University of Pennsylvania Press, 1979.

Table 2.--Estimates of Mean Annual Income in Dollars Over Ten-Year Intervals for Six Cohorts

Year	Ages		
	25-34	35-44	45-54
1. 1947	2,704 (1947)	5,300 (1957)	8,342 (1967)
2. 1948	2,898 (1948)	5,433 (1958)	8,967 (1968)
3. 1949	2,842 (1949)	5,926 (1959)	9,873 (1969)
Year	Ages		
	35-44	45-54	55-64
4. 1947	3,344 (1947)	5,227 (1957)	7,004 (1967)
5. 1948	3,508 (1948)	5,345 (1958)	7,828 (1968)
6. 1949	3,281 (1949)	5,587 (1959)	8,405 (1969)

Note: Each cohort was surveyed every ten years. The first cohort, for example, contains individuals who were 25-34 years of age in 1947 and had an average income of $2704; in 1967, when they were 45-54 years of age, their mean income was $8342.

From: Boruch, R.F., and Cecil, J.S. Assuring the Confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press, 1979.

Danish-U.S. collaboration supported by the National Institute of Mental Health (NIMH) has involved intensive record matching to determine how children born of schizophrenic parents fare when they are adopted and reared by non-schizophrenic, foster parents. Matching among records of hospitals, surveys, and psychiatric systems was required to execute the research. The work appears to confirm a genetic component in that incidence of schizophrenia among such children is higher than its incidence among adopted children born of nonschizophrenic parents, including children adopted by schizophrenic parents.

That use of matched records can improve scientific analysis seems clear from studies of the economic impact of education. Paul Samuelson, for example, has argued that returns on higher education are substantial. Christopher Jencks has analyzed various survey data sets to argue that the returns are marginal. Fagerlind used Swedish data that were better than data available to either Samuelson or Jencks: matching individual records from military screening; birth registries, tax registries on earnings of the respondent, census records on occupational mobility. These analyses favor Samuelson's theory.

Neither the schizophrenic study nor the Samuelson-Jencks-Fagerlind work is unambiguous, of course. There has been considerable debate about the models exploited in each. The main point is that improvements in data, notably through linkage of records from a variety of sources, can enhance the analyst's ability to explore ideas and test hypotheses. The "sources" may be additional survey panels in a longitudinal design. Or they may be administrative records that are at least as good as survey data.

2.2 Matching to Avoid Aggregation Error and Ecological Fallacy

We often compute correlations between X and Y based on aggregate data, being cautious, of course, in generalizing to the individual level. The opportunity to match individual records often gives us the opportunity to entirely avoid the problems and caution engendered by aggregation.

One of the oldest illustrations is still the most dramatic. At a particular point in time, the correlation between literacy rate and color (black vs. white) computed on the basis of nine census regions in the United States was .95. When the data are aggregated by State instead of region, the correlation becomes .77. Finally, access to individual records led to a correlation of .20.

2.3 Matching Records in Randomized Tests of Social and Education Programs

In Middlestart education programs at Oberlin College, for instance, a series of experiments was undertaken to understand whether precollege programs worked for promising but poor adolescents. The evaluators relied on randomization to assure statistically unbiased estimates of long-run program effect. They relied on records matched among surveys, high school records, and standardized precollege records to avoid the problem of low validity in student reports of grades, and to enhance the statistical power of the tests.

Randomized field experiments, designed to understand how one can increase compliance with food stamp registration rules, have been mounted by the Office of Analysis and Evaluation of the U.S. Department of Agriculture's Food and Nutrition Service (1984). These tests depend on matches of records among participant reports and records of State Employment Security agencies and the Food Stamp Agency. Results show remarkable decreases in food stamp costs and employment benefits for certain innovative approaches to compliance assurance.

Police research is relevant, too, of course. In the Minneapolis Domestic Violence Experiments,

the object was to understand how to handle domestic violence effectively, for example, immediate arrest versus referral to social services, within limits. Undertaken by the Police Foundation, the experiment involved matching among police patrolman records, precinct arrest records, and the experimenters' records. Arrest, incidentally, seems to work in the sense of reducing subsequent incidence of domestic violence (Sherman and Berk, 1984).

Motor vehicle research is pertinent to matching, too. Work done some years ago by the Insurance Institute for Highway Safety, for example, involved linking an experimenter's observations on vehicle registration, the drivers' seat belt use, and advertisements on the topic, to motor vehicle records that contained data on the drivers' residence area. The residence area match with the other information made it possible to determine how effective alternative TV commercials, directed to different areas, were in encouraging seat belt use.

Program Implementation and Validity of Reporting

The New Jersey Negative Income Tax Experiments attended to the potential problem of overpaying welfare recipients. This set a standard for validity studies in later experiments. Overpayment of benefits in such experiments was critical insofar as (a) other sources of assistance were available to participants in the experiment, and (b) they might receive such assistance illegitimately through error (welfare rules are complicated) or deceit (crime is still a bastion of the free enterprise system). All participants reported their income based on recall. Matching these reports with administrative records helped to assure reasonable implementation of the program and to assess quality of reporting.

For example, welfare audits were created to reduce or prevent the problems: these depended heavily on the experimenters' ability to match research records with records of welfare boards. Internal Revenue Service (IRS) W-2 forms were required of families and permitted comparisons between IRS-reported income and income reported to the experiment. (Underreports of income to the experiment relative to IRS appear to have been less than 15 per cent). The Social Security Administration (SSA) cooperated by taking the experimental data, matching to its own records on individuals, and providing aggregate earnings data (not individual records) to permit estimates of underreporting of earnings in the experiment (Kershaw and Fair, 1979). (The SSA comparison suggests that about 80% of families underreport to researchers by 15% or less even when they have incentives to misreport.)

In the Seattle and Denver Income Maintenance Experiments (SIME/DIME), research records were matched to public agency records on food stamp purchase, rent subsidy, and wages. The experiment produced some small surprises through evidence that public records on rent support and food stamps were less accurate than respondents' reports in the experiments, evidence that was later strengthened by independent investigation. Underreporting of wages appeared in the expected direction based on matches with IRS records (Halsey, 1980).

In the New Jersey Negative Income Tax Experiments, Mercer County Welfare Board records were used in a pilot test to determine composition, work history, and residential mobility of families that attrited from the experiment and could not be interviewed without great difficulty. More generally, the attrited families in five cities were traced through post office change-of-address cards, motor vehicle registration agencies, welfare boards, prisons, and community groups. Apparently, face-to-face interviews with former neighbors were most productive (Kershaw and Fair, 1979).

The use of administrative records to trace attriters and assess misreporting in all the income maintenance experiments is an important but underexamined topic. The experiments themselves were well run, relative to any pragmatic standard. They cover a sufficient number of sites to tantalize any scholar with an interest in regional differences in record accuracy, misreporting models and so on. Sample sizes for validity studies were small, however. This may account partly for the disinterest of scholars. Still, it is a bit distressing to some that otherwise thoughtful commentators such as Hausman and Wise (1985) fail to recognize the policy import of misreporting and the methodological contributions of randomized tests of economic programs to this area.

2.4 Matching and Testing New Ways to Elicit Information

Innovative ways to elicit information, such as randomized response, need to be tested despite their cleverness. We are unaware of any individual match studies in this arena. But studies that compare marginals or point estimates for individuals on whom both responses and archival records are available are done.

So, for example, Bradburn, Locander and Sudman found that a randomized response method worked at times to reduce response distortion on sensitive topics such as drunk driving arrests. The basis for comparison was administrative records on the same individuals, e.g., arrest records. Individual records were not matched; comparisons are based on marginal counts or averages. But matching in this and related research is possible in principle. And it may be useful insofar as it helps us to understand how response distortion varies with sensitivity of the traits that are being examined and characteristics of individual respondents.

A fascinating example of a near match study on reporting energy use to the Census Bureau was given by Tippett (1984) in recent 1984 Proceedings of the ASA. Her experiment involved encouraging utility companies to send a randomly

assigned group of individuals a statement of the year's utility bills. A randomly assigned comparison group was not sent the statement. The statements were sent prior to the 1980 census to understand whether providing such records could enhance quality of respondents' reports of utility costs to Census. Both groups overstated costs; the "primed" group overstated costs appreciably less than the control group. Again, matching could be helpful in understanding how degree of reporting error varies with the true state of the individual.

2.5 Matching Records to Understand Validity of Response and Inferential Errors

We know that error in measurement of a response variable degrades statistical power. More important, it can lead to invidious biases in covariance analyses based on fallibly measured covariates. That is, the analyses can make programs look useless when their effects are in fact slightly positive, and can make programs look harmful when indeed they are merely useless (Riecken et al., 1974). The recent work by Andersen, Kasper, Frankel and their colleagues (1979) on total survey error clarifies the effect of imperfections in observational studies generally.

The point is that understanding validity of the measures is important in applied social research, especially policy research, as well as in basic work. Matching studies undertaken in education and supported by the National Institute of Education and the National Center for Education Statistics, for instance, show that females are appreciably more accurate than males in responding to questions about their own grades and coursework, and more accurate in reporting on income and education levels of parents. There are race differences as well as gender differences in respondents' ability and willingness to furnish information. Failure to recognize these differential validities can lead to errors in understanding which programs work and for whom. Matching helps us to avoid those errors merely by showing which subgroup differences in reporting quality may account for differences in performance.

Imperfect measures of employment and occupation can produce similar biases in explanatory models of income gain and other response variables. Matching studies of the sort undertaken by Mathiowetz and Duncan (1984) in which private employer records are linked to survey records of the Panel Study on Income Dynamics are not common. But they have potential for revising ideas about error structure. Errors in retrospective reporting on employment and occupation seem to depend less on time or recency than on salience of events in a particular month (e.g., a raise) and task difficulty (e.g., a single unemployment spell vs. multiple spells). Gender and race differences in reporting error are reduced when these variables are taken into account.

3. WHEN BENEFITS OF MATCHING ARE NEGATIVE OR AT LEAST NOT SO CLEAR

Having the option to capitalize on existing records and to match so as to obtain a better file is important because the idea and the relevant technology have been so useful. For instance, the 1984 Proceedings of the ASA, Section on Survey Research Methods contains over 30 articles that concern exact matching methods or analysis or depend heavily on matching for conclusions (validation studies, capture-recapture, others). Unlike the 1984 Proceedings, the 1978 Proceedings of the same section contained no sessions on using administrative records in conjunction with surveys or on quality control of statistical systems (partly through linkage).

The Interagency Linkage Study participants --Internal Revenue Service, Census, and Social Security Administration--deserve special credit for advances in this arena. Other agencies have worked at least as vigorously and as often, however, e.g., the National Center for Education Statistics and the National Center for Health Statistics. And a good many research projects undertaken with support of the U.S. Department of Labor's Employment and Training Administration, the National Institute of Justice, the National Center for Health Services Research (and the Department of Health and Human Services more generally) have made use of matching where it has been useful and legally possible to match.

Matching is a seductive option, however. That is, we may capitalize on matching existing records to obtain estimators that are efficient and cheaply produced, but wrong. They are wrong at times partly on account of the administrative system in which matching must take place. They are wrong partly because the matched data (observational data more generally) are inappropriate despite their accessibility and ostensible relevance.

Consider a recent case, one in which the role of matching is important.

3.1 The Case at Hand

Estimating the effect of manpower employment and training programs in this country is a significant policy issue. Since 1965 or so, most estimates have been based on observational data, i.e., sample surveys. Two kinds of observational data are most relevant here--the Continuous Longitudinal Manpower Survey (CLMS) and the Current Population Survey (CPS). Both are based on large, well-designed samples. Both have been augmented by matching respondent records with social security (SSA) earnings records.

The CLMS-SSA match works as follows. The Bureau of the Census, under agreement with the Department of Labor, designs the CLMS probability sample and collects the data. The record on each individual includes identifying information and social security number. A list of respondent SSA numbers is given to the SSA which then searches

SSA files for records on the relevant individuals. The SSA records include the social security number, earnings, birth year, six letters of surname, and other bits of information. These SSA records are then given to Census for matching to the CLMS survey records under an interagency agreement that assures confidentiality of both sets of files. Census matches the records, deletes identifying information and geographic area related characteristics. The geographic data are deleted to prevent deductive disclosure.

Recently, the U.S. Department of Labor contracted for two kinds of analyses bearing on the impact of manpower programs and based on these files. In the first kind, different, well regarded contractors were asked to use such data to estimate the effects of training programs (Westat, 1984; Dickinson, et al., 1984; Bassi, et al., 1984). In the second kind of study, estimates based on observational survey data, similarly constructed, were compared to estimates yielded by randomized field experiments. In particular, the models used on CLMS and CPS data were used to construct quasi-experimental comparison groups. The performance of these comparison groups was compared to randomized control groups generated in the National Supported Work Demonstration (Fraker & Maynard, 1985).

The results of three independent analysts generating models and using them to estimate program effects based on CLMS and CPS data yielded the following results:

(a) Effects of training on earnings are positive and significant, especially for females and all post Comprehensive Employment and Training Act follow-up years (Westat, 1984, p. 61).

(b) Effects on earnings for men are not generally significant; effects on women's earnings are significant (Bassi, et al., 1984, p. xv).

(c) Effects on earnings for men tend to be significant and negative, but effects on women are positive and significant but small (Dickinsen, et al., 1984, p. xiii).

We have oversimplified here, of course. "Significance" is emphasized too much and the statements are misleadingly blunt. But the conclusions are as they appear in the final reports.

Comparing estimates of control group performance similarly constructed to estimates of control group behavior based on randomized experiments had the following results: depending on the particular model and matching strategy used, estimated effects on earnings range from minus 2000% of "true" earnings to plus 50% of "true" earnings, "true" being estimated from the randomized trial.

These results should be a bit disconcerting. They are indeed puzzling and potentially embarrassing. The Labor Department deserves praise for scholarship in disclosing the puzzle and for its political fortitude in willingness to tolerate potential embarrassment.

More to the point, what are the reasons for the discrepancies? Sampling variations may account for some of the differences. But it is not likely to account for all. In the next section, the reasons engendered by another line of argument are discussed, in the interest of understanding the strength and weakness of the argument.

3.2 Line of Argument

The critic can propose that part of the reason for discrepant results lies in relying---

(a) solely on observational data, matched or otherwise, and
(b) on models whose validity is untestable with the data at hand.

Critics who are more blunt may further suggest that the CPS, SSA, and CLMS are used because they are available and seemingly appropriate and not because they are sufficient.

Finally, the administrative system in which matching occurs demands that one give up some opportunities that should not be given up if the object is to produce good estimates of program effects.

To illuminate the contentions, consider SSA earnings matches with observational data from surveys. Problems similar to ones discussed here occur in other contexts. The material that follows is based on thoughtful reports by Bassi, et al. (1984), Dickinson, et al. (1984), and Westat (1984), that is, the producers of the estimates of manpower program effects.

State Identifiers and Areas as Missing Data

Welfare laws differ appreciably among states. These laws determine who gets welfare and how much they get. It makes sense to incorporate such data into any analysis of the way a federal employment program is used by the poor and what the impact of the program is. Local labor market information is also crucial to thoughtful analyses of why people do or do not get jobs as a consequence of programs.

Yet such information is absent from public use microdata files that are released after matching records. The result is that the economist must be content with data that are bound to generate estimates of program effect that are likely to be biased. That is, important major variables are left out of the left hand side of explanatory equations because they are deleted from public use files or remain unmeasurable variables. The incompleteness of the model is responsible for biased estimates of effect.

Why are they left out of such files? Because their inclusion will permit deductive disclosure. That is, it becomes possible to deduce the identity of anonymous respondents if

information about geographic area is supplied. The Census, for example, cannot countenance the possibility of deductive disclosure of information that it has collected, and invokes Title 13 to justify its position. Census perspective on this matter is important not only for this case: The Bureau "performs a major portion of its survey work on a reimbursable basis for other Federal agencies" (Cox, et al., p. 1, 1985). It is important as a survey agency and as a model of virtue in this respect.

Exclusion of relevant data seems to us to be the most serious consequence of our use of Census-SSA in data collection and matching. From such a matching system, we cannot produce credible estimates without the appropriate variables.

Earnings not Covered by SSA

Many public sector jobs are not covered by SSA reporting. Insofar as the employment and training program leads to jobs that are public sector and not covered, two problems occur. When earnings are a dependent variable, estimates of impact will be understated when the comparison groups jobs are more likely to be SSA covered. When earnings are used as a covariate, e.g., "prior base year," estimates of program impact will be biased because the covariate is fallible.

One way to assess the problem is by looking at interview-based earnings reports and SSA earnings, of course. Dickinson, et al. (1984) did so. They found substantial error in CLMS interview reports, e.g., 33% of CLMS respondents who said they did not work in 1977 had positive SSA earnings reported. The rate for CPS is about 10%. We still have a dilemma: SSA is clearly better than self-reports of earnings, although they are imperfect.

SSA earnings data are also truncated at both ends. For example, the maximum earnings subject to SSA tax is the maximum recorded earnings level. Dickinson, et al. (1984) examined interview earnings and SSA cap earnings to find no appreciable difference between analyses using each. i.e., estimates of program effect are about the same (p. 98).

Updatedness: A Possibly Tractable Problem

As of 1983-84, the period of DOL analyses of interest here, 1979 SSA records merged with CPS and CLMS data are incomplete. That is, not all 1979 SSA earnings for members of these samples were available. A "zero" entry for the missing data means we cannot tell how much missing data there is. Bias cannot be estimated. Still, this problem seems tractable.

Program Participation not Measured: A Possibly Tractable Problem

The CPS does not now measure participation in employment programs. Consequently, a public use file will not permit construction of a comparison group that is "uncontaminated." Among youth in the CPS comparison group, for example, it has been estimated that between 1975-78 30% entered CETA. So the contamination issue seems important. It, too, seems tractable but not without substantial effort.

Alignment Problems

According to Dickinson, et al. (1984), in Westat's analysis of the FY76 cohort, SSA earnings in calendar year 1975 were used to match individuals, despite the fact that calendar year 1975 earnings included up to six months of post-enrollment earnings for some CLMS members, (p. 35). Dickinson, et al., used calendar year cohorts rather than fiscal year cohorts. The disadvantage is in potentially missing the preprogram drop in earnings.

4. RESTATEMENT OF THE PROBLEMS AND POSSIBLE SOLUTIONS

4.1 Core Problems

There are two kinds of problems implicit in the case just presented. The first concerns reliance solely on surveys coupled to administrative records to understand relative effects of programs. Problems engendered by relying on such data affects not only efforts to estimate impact of manpower training programs, of course. They also appear in health services research, psychiatric and mental health services evaluations, assessments of court procedures, tax compliance, and police procedures (Riecken, et al., 1974). We attribute the problems partly to the seductiveness of matching and partly to the more dangerous problem of untestable models.

The second kind of problem stems from our inability to use all the data in ways that permit confidence that the analysis is statistically unbiased. Denial of access to micro-records on account of deductive disclosure affects research by Bureau of Labor Statistics (Plewes, 1985) as well as the DOL Employment and Training Administration, by the National Institute of Justice (e.g., in victimization studies), and others. The issue is also likely to affect newer statistical programs, e.g., the Survey of Income and Program Participation (David, 1984). We attribute this problem to the administrative environment in which matching technology must be exploited.

4.2 Resolving the First Kind of Problem and Exacerbating the Second

A scientifically reasonable solution to the first kind of problem is to actively experiment. That is, we need to run randomized trials of projects, project components, or project variations. The research policy option that seems worth exploring is routinely adjoining randomized experiments to the longitudinal studies and/or record files that are matched. See for instance, the Hollister, et al. (1985) report on evaluating the effectiveness of youth employment programs.

Exercising the option of randomized experiments can exacerbate the second problem, i.e., of deductive disclosure. That is, experiments generally involve a smaller number of individuals than national probability samples and more detailed information on each individual. This makes deductive disclosure easier. It also makes it difficult to adopt sampling rates as a partial index of likelihood of deductive disclosure (Cox, et al., 1985). If an agency with restrictive rules is involved in data collection then no public use tapes with sufficient detail will be released and no sensible competing analyses will be done.

Apart from the information demands of randomized experiments, the demand for microdata is increasing. Cox, et al. (1985) recognize that this increase has strong implications for Census policy on disclosure and they provide a thoughtful analysis.

4.3 Resolving the Second Problem

The possible resolutions to the disclosure problems are of at least three kinds: procedural, statutory, and empirical. The following options illustrate each.

Avoiding Restrictive Agencies

One may stay away from agencies that have data worth matching but that also have restrictive disclosure policies. Indeed, it is not hard to argue that private agencies are as capable of producing good data with equal privacy protection for the respondent and fewer constraints on the research than a government agency. The case is especially arguable for controversial topics of research such as AIDS, but it is also relevant here (Boruch, 1984).

Still, doing without micro-records from agencies such as the Census Bureau, Social Security Administration, or others, and doing without their capacity to serve as a broker for linking records from independent sources, is not an attractive prospect. We may gratuitously abandon opportunities to do socially useful and reliable research by foregoing collaboration with such agencies. So it is sensible to consider other options in addition to this one.

Proactive Change in Law and Policy

Alteration of law and more feasibly the interpretation of law is possible and seems desirable. The battles for statistical enclaves suggest, however, that this war will not be won easily, if at all. Still, sensible work has been done and some progress in clarifying issues has been made (Alexander, 1983). Assaults on Census's stewardship of Title 13 seem not to have been productive, for example (Plewes, 1985). Still, working toward legitimate reinterpretation of law seems an effort worth making, especially if more empirical research can be brought to bear on the issue of perceived risks of disclosure to populations. This brings us to the next option.

Empirical Research

Research on the role that privacy and consent have in record matching contexts seems sensible. How much the assurance of confidentiality means to respondents and how it influences the cooperation rate has received some attention from empiricists. For example, randomized field tests have been run under the auspices of the NAS Committee on National Statistics to understand whether people attend to assurances about privacy (Panel on Privacy and Confidentiality, as Factors in Survey Response, 1979). We agree with Thomas Plewes (1985) of the Bureau of Labor Statistics (BLS) in urging that more related work needs to be done.

In particular, obtaining respondent consent to disclose and link records for research purposes is an avenue for resolving deductive disclosure/confidentiality problems at Census, SSA, and elsewhere. We are aware of no good field experiments to determine effective strategies to elicit consent or their consequences. The BLS has been successful, according to Plewes, in eliciting consent for disclosure of its data to the Department of Agriculture, for instance, so that better sampling frames for forms could be developed. But this evidence is anecdotal and few hard data from controlled trials are available.

Both Cox, et al. (1985) at Census and Plewes (1985) at BLS recognize that public perceptions of government agencies are important in this context. That is, public confidence in government affects cooperation in surveys and resultant public data.

This chain of reasoning is plausible. But our agreement is a matter of intuition, not hard evidence. Moreover, the politicians' view of the idea and its implications for a bureaucracy and votes seem important. Neither the Census Bureau nor BLS (nor other agencies) can work on this tangle of issues with impunity, at least not always. Academic researchers have some responsibility to do so if they expect to have access to good data. We know of very few who are involved in such work, e.g., Flaherty, Hanis, and Mitchell (1979) in Canada, Mochmann and Muller (1979) and Damman and Simitis (1977) in Germany.

Research: Analytic

The Department of Labor's support of competing analyses, and of comparisons of the results of randomized tests to the results of nonrandomized assessments, is admirable. Research in the same spirit on matching and disclosure is warranted.

The thoughtful observer ought to admire the work by Nancy Spruill and Joe Gastwirth (1982) on microaggregation and masked data and work by George Duncan and Diane Lambert (1985) on disclosure limited dissemination. Their analysis helps to actualize a balance between privacy needs and the need to assure quality of released data. The thoughtful observer will also recognize, however, that not much work has been

done on the costs, traps, flaws, and benefits of using the suggestions of these analysts. We ought to know more about these issues. And so we ought to invest some resources routinely in the design of side studies to illuminate the limits on the utility of their work.

The importance of this matter stems partly from the fact that the effects of social programs in tax compliance, police, training, and employment effects are usually small. Expecting small effects, we should then be better able to anticipate the effects of micro-aggregation, random perturbation (contamination), random rounding, collapsing, and other strategies used to transform data so as to make it suitable for public use. All such tactics are used by the Census and other agencies to protect individual (and at times institutional) privacy (Cox, et al., 1985). But very little has been published about their implications for the validity of inferences based on analyses of such public use data.

Administrative Procedures

Suppose that we create a matching system under which public use tapes that are first expurgated or "adjusted" to reduce deductive disclosure problems are used for crude analyses. These analyses are eventually verified using the unexpurgated records by the agency that maintains the more detailed micro-records. The procedure achieves a balance between privacy concerns and scientific demands for quality in analysis.

But it demands substantial resources, i.e., a sequential system of crude analyses, based on public use tapes, followed closely by confirmatory analyses, based on within-agency analysis of micro-records. Still, the option seems worth considering especially because the procedure seems generalizable, e.g., to matching economic variables in the Survey of Income and Program Participation (David, 1984).

For example, 1976 Annual Housing Survey data on energy use were matched on geographic area to local utility company data. Census created the file. To protect against deductive disclosure, the Census adjusted the accuracy of energy use data "prior to release to guard against the possibility that the utility companies could uniquely identify individuals on the released file from their reported cost data" (Cox et al., 1985, p. 22). The adjustment involved random perturbation (that can be accommodated up to a point in analyses, given the perturbation parameters) and rounding. We are unaware of any formal benefit-cost analysis of this case. We believe that some sort of evaluation of such cases should be undertaken and published.

5. REPRISE AND CONCLUSION

There is no doubt that matching can be and has been useful in a variety of social research projects. Moreover, the analytic work on the topic by Pelligi and Sunter (1969) and others is remarkable for its thoughtfulness. The technology for matching, considered apart from the matching system (organization and data), has stimulated fascinating research by academic and bureaucratic scholars. But solutions to the problem of getting the benefit of matching without reducing interpretability of data are not yet clear.

The ingeniousness of a matching algorithm is one thing. The system in which the algorithm is applied is quite another. It is clear that the administrative environment of the matching system can lead to invidious problems in analysis at the policy level. The problems lie not so much in matching technology as in other elements of the matching system: the data and rules under which it was collected, the institutional vehicle for matching and the rules governing it, and the procedures one uses to understand the errors we make based on analyses of matched data. The problems are severe enough to warrant the serious concern of applied statisticians and social scientists. Unless attention is dedicated to the matter we will do far less than we should for science, society, and the profession.

ACKNOWLEDGMENTS

The background research for this paper was supported by the National Science Foundation during 1984-85 to the Social Science Research Council (Measurement and Methodology Division, NSF) and 1985-86 to Northwestern University (RII-8418179 Ethics and Values in Science and Technology, NSF). We are indebted to Robert Pearson of the Social Science Research Council and Richard Beatty of Westat for background information. Readers should be grateful, as I am, to Theodore Clemence of the Census Bureau for graciously providing references on Alexander Bell.

REFERENCES

Alexander, L. Proposed legislation to improve statistical and research access to federal records. In R.F. Boruch and J.S. Cecil (Eds.), Solutions to ethical and legal problems in social research. New York: Academic Press, 1983, 273-292.

Andersen R., Kasper, J., Frankel, M.R. and Associates. Total survey error. San Francisco: Jossey Bass, 1979.

Bassi, L.J., Simms, M.C., Burbridge, L.C., and Betsey, C.L. Measuring the effect of CETA on youth and the economically disadvantaged. Washington, D.C.: The Urban Institute, April 1984.

Bell, A.G. The deaf. In: U.S. Department of Commerce and Labor, Bureau of the Census. Special Reports: The blind and the deaf, 1900. Washington, D.C.: U.S. Government Printing Office, 1906.

Boruch, R.F., and Cecil, J.S. Assuring the confidentiality of social research data. Philadelphia: University of Pennsylvania Press, 1979.

Boruch, R.F. Should private agencies maintain federal research data? IRB, 1984, 6(6), 8-9.

Box, J.F. R. A. Fisher: The life of a scientist. New York: Wiley, 1978.

Bruce, R.V. Alexander Graham Bell and the conquest of solitude. Boston: Little, Brown, and Company, 1973.

Cox, L.G., Johnson, B., McDonald, S.K., Nelson, D., and Vazquez, V. Confidentiality issues at the Census Bureau. Presented at the first Annual Census Bureau Research Conference. Reston, Virginia, March 20-23, 1985.

Damman, U., and Simitis, S. Bundesdatenschutzgesetz. Baden-Baden: Nomos Verlagsgesellschaft, 1977.

David, M. Discussion. Proceedings of the American Statistical Association: Social Statistics Section. Washington, D.C.: ASA, 1984, pp. 534-536.

Dickinson, K.P., Johnson, T.R., and West, R.W. An analysis of the impact of CETA programs on components of earnings and other outcomes. Menlo Park, CA: SRI International, November 1984.

Duncan, G.T., and Lambert, D. Disclosure limited data dissemination. Journal of the American Statistical Association. 1985, in press.

Fellegi, I.P., and Sunter, A.B. A theory for record linkage. Journal of the American Statistical Association, 1969, 64, 1183-1210.

Flaherty, D.G., Hanis, E.H., and Mitchell, S.P. Privacy and access to government data for research. London: Mansell, 1979.

Fraker, T., and Maynard, R. The use of comparison group designs in evaluations of employment related programs. Princeton, N.J.: Mathematica Policy Research, 1985.

Halsey, H.I. Data validation. Chapter 2 of P.K. Robins, R.G. Spiegelman, S. Weiner, and J.G. Bell (Eds.) A guaranteed annual income: evidence from a social experiment. New York: Academic, 1980, pp. 33-55.

Hausman, J.A. and Wise, D.A. (Eds.) Social experimentation. University of Chicago Press, 1985.

Hollister, R. and others (Eds.) Report of the Committee on Youth Employment Programs. Washington, D.C.: National Academy of Sciences, 1985.

Kershaw, D. and Fair, J. The New Jersey income maintenance experiment: Operations surveys, and administration, Volume I, New York: Academic Press, 1979.

Locander, W., Sudman, S. and Bradburn, N.M. An investigation of interview method, threat, and response distortion. Journal of the American Statistical Association, 1976, 71, 269-275.

Mathiowetz, N.A., and Duncan, G.J. Temporal patterns of response errors in retrospective reports of unemployment and occupation. Proceedings of the American Statistical Association: Section on Survey Research Methods. Washington, D.C.: 1984, 652-657.

Mochmann, E., and Muller, P.J. (Eds.), Data protection and social science research. Frankfurt/New York: Campus Verlag. 1979.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J.Strugnell, A. and Abbatt, J.D. Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. Computers in Biology and Medicine, 1983, 13(3), 157-169.

Panel on Privacy and Confidentiality as Factors in Survey Response. Committee on National Statistics. Privacy and confidentiality as factors in survey response. Washington, D.C.: National Academy of Sciences, 1979.

Plewes, T.J. Confidentiality principles and practice. Paper presented at the first Annual Census Bureau Research Conference. Reston, Virginia, March 20-23, 1985 (available from the author, Bureau of Labor Statistics).

Riecken, H.W., and others. Social experimentation: A method for planning and evaluating social programs. New York: Academic, 1974.

Sherman, L.W., and Berk, R.A. The specific deterrent effects of arrest for domestic assault. American Sociological Review, 1984, 49, 261-272.

Spruill, N.L., and Gastwirth, J. On the estimation of the correlation coefficient with grouped data. Journal of the American Statistical Association, 1982, 77, 614-620.

Tippett, J.A. An experimental project to improve reporting of selected costs in the 1980 census. Proceedings of the American Statistical Association: Survey Research Methods Section. Washington, D.C.: ASA, 1984, 323-328.

U.S. Department of Agriculture, Food and Nutrition, Office of Analysis and Evaluation. Food stamp work registration and job search demonstration. (Contract No. 533198085). Alexandria, VA: DOA, 1984.

Westat, Inc. Continuous Longitudinal Manpower Survey: Net impact results. Rockville, MD: Westat, April 1984.