International symposium on linked employer-employee data

*A recent conference addressed
the benefits and challenges
of constructing data bases
that link firm and worker characteristics*

John Haltiwanger,
Julia Lane,
Jim Spletzer,
Jules Theeuwes,
and
Ken Troske

Employers and employees interact with each other every day in the workplace. However, the data collected and processed by government statistical agencies do not generally reflect that interaction. Historically, statistical agencies have collected information about the activities of workers from household surveys and information about employers from business surveys. While these surveys provide a wealth of information about the economy and society, there remains a large information gap, given the need to understand the interaction of employers and employees.

This need has been intensified by the dramatic changes in the international economy over the past 20 years. Changing trade patterns, technological developments, and the restructuring of jobs have had an unavoidable impact on workers. The full nature of this impact on earnings inequality, employment security, and worker incomes is not well understood, primarily due to a lack of adequate data.

The May 1998 conference on linked employer-employee data, held in Washington, DC, brought together a wide range of social scientists and statisticians representing more than 20 countries. One important reason for the international scope of the conference is that some other industrialized countries are much more advanced than the United States in constructing such data. Worker-firm linked data sets have been used to analyze the changing structure of the economy in Canada, France, Scandinavia, the Netherlands, and Belgium. Even in those cases, however, the data often are not comparable across countries, thus inhibiting international comparisons. The confer-

ence enabled researchers and statisticians, who too often work in isolation in their own countries, to become aware of the efforts made in neighboring countries.

While it is not difficult to make the case that there is a great analytical need for linked employer-employee data, there are a number of issues that have to be resolved before creating such a resource. Often foremost among these are limited resources and concerns about respondent burden. The most attractive option in many countries is to find some means of combining existing household and business data without having to initiate costly new surveys.

Second, the procedures for linking employer-employee data invariably involve sensitive administrative data. Therefore, the protocol for creating the data sets in secure environments that protect the confidentiality and privacy of survey respondents is a fundamental issue. The good news on these two fronts is that a number of countries have made significant progress in creating linked employer-employee data while protecting confidentiality. The conference offered extensive opportunities to learn about the return to these data creation efforts, as well as new approaches to dealing with confidentiality.

The aims of the conference, therefore, were to demonstrate that many linkage/access/confidentiality issues have been overcome in some countries, to discuss how those lessons can be applied to other industrialized and developing countries, to demonstrate by example the value of research based on linked employer-employee data, and to facilitate international comparisons.

John Haltiwanger
(Census Bureau and
University of
Maryland), Julia
Lane (Census
Bureau and
American University), Jim Spletzer
(Bureau of Labor
Statistics), Jules
Theeuwes (University
of Amsterdam), and
Ken Troske (University of Missouri).

These broad goals led the conference to develop eight main themes:

- Creating employer-employee data sets
- Confidentiality of linked data
- Econometric issues
- Analyzing training and productivity
- Analyzing firms, workers, and wages
- Analyzing firms, jobs, and turnover
- Progran development and policy analysis
- International comparisons

There were between two and four sessions devoted to each of these eight themes, with three papers presented in each session. The conference organizers invited eight rapporteurs to provide a summary and synthesis of the papers presented on each of the themes and report back to the plenary sessions. These reports are provided in the following sections. ☐

Creating employer-employee data sets

John Haltiwanger

Much has been said about the great research potential of employer-employee linked data sets, and many of the conference participants identified challenges to be faced in using such data. Several sessions of the conference focused on yet another challenge: the nuts and bolts of actually creating the data sets. The work needed to actually create linked data sets is extensive, tedious, and painstaking. It is of fundamental importance to realize that, although it is much more glamorous to talk about the new insights to be gained from these data, the insights are not going to come unless the data sets are established—and established well.

In investigating the possibilities, one of the panels considered the use of administrative data for creating employer-employee data sets. The panel focused on the Scandinavian countries, which have some of the best administrative data in the world. For example, in Sweden, researchers can draw on the following registers: the central business register; the population register; the register of education; the register of students; the register of social benefits; the register

John Haltiwanger is visiting scholar at the U.S. Census Bureau and a professor of economics at the University of Maryland.

of jobseekers; income verification records; and tax return forms. There are similar sources in other Scandinavian countries. Of course, the administrative data do not exist simply for linking, but are powerful tools in and of themselves.

The United States does not have such rich administrative data— at least not all in one place. There is a fair amount of data in the Federal system, but it is spread across a number of agencies. Moreover, much of the most interesting administrative data are held at the State level. Some of the analysts participating in the conference had overcome this problem by simply going to the State level. However, State boundaries are pretty artificial. To the extent to which workers cross State lines to work, a State-level approach is not going to be sufficient.

Another approach to developing linked data sets involves Federal-State partnerships. The Bureau of Labor Statistics has a number of successful Federal-State partnerships, including the ES-202 and Mass Layoff Statistics programs. However, evidence presented at the conference suggested that substantial institutional barriers will have to be overcome to bring the administrative data in the United States together, while at the same time protecting privacy and confidentiality.

A related difficulty with U.S. administrative data (and this is true even in the Scandinavian countries) is that the data were not really intended for the purpose of linked data research. This problem can be addressed using yet another approach: surveys with intended matched employer-employee components. Surveys of this type, as outlined by conference participants from the United Kingdom, Canada, and Australia, ask a series of questions of a sample of employers. The researcher then obtains a list of employees from each employer, and permission from the employer is sought to conduct a second-stage survey of the employees. This approach is a top-down method. The alternative is a bottom-up approach, which involves obtaining a list of workers, asking them an extensive series of questions, and then getting their permission to contact their employer for a second-stage survey.

Although the potential of such surveys is great, the logistics for conducting them are complicated. And, which way should the surveys be structured, top down or bottom up? Attrition—what is lost in the second stage—is different depending upon that choice. Conducting the second stage is pretty tricky in either case. If one approaches an employer and asks to survey the employees, one must set up a very precise protocol to ensure that a random sample of workers is selected. In some cases, managers may ask only their best workers to be in the survey. Conversely, asking an employee's permission to talk to the employer is also a sensitive business.

Another interesting problem is deter-

mining who knows what in the following sense: if one puts together a survey and wants to ask some questions of the employer and some questions of the employee, which questions should one ask of each? Sometimes, when both parties are asked the same questions, they give very different answers. Two-stage surveys also face a timeliness issue. If one collects data from the employer at one point, it might be some time before the corresponding information from employees is obtained.

Another limitation of survey instruments that have both employer and employee components is that most of them relate to a single point in time. From the perspective of the labor market analyst, this point-in-time approach may not be a particularly promising way to really get at labor market dynamics. To their credit, the Canadian survey agency is planning for a longitudinal component in its linked data sets but, given the high pace of worker and job flows, this design may be difficult to implement.

On the basis of evidence presented at the conference, this writer believes that the United States needs to pursue a mixed approach. This would involve exploiting the administrative data that we do have to the greatest extent possible. U.S. statistical agencies can build some core administrative record data bases that are comprehensive employer-employee matched data at the national level.

But it is also the case that such data sets in the United States will be fairly sparse. This might be overcome by linking the core administrative data to existing surveys to yield a much bigger "bang for the buck."

More generally, statistical agencies might adopt a plug-and-play approach towards surveys, existing or new. One can plug targeted modules into existing or new surveys—establishment or household—and make sure the two can be linked to the underlying core administrative data. However, there would always be the need for at least a few targeted surveys that have employer and employee components. The United States is going to need to push the administrative data as far as it can; this will require greater coordination across both statistical agencies and sources of the administrative data, and must include both Federal and State agencies. Achieving this coordination, developing the data, providing access to users for statistical and research purposes, and doing this in a manner that protects the privacy and confidentiality will be a great challenge.

Confidentiality of linked data

Gerald Gates

The conference highlighted the importance of confidentiality in linkages of employer and employee data. Papers on the topic consider two relationships—one involving administrative records holders who are asked to share responsibility for protecting their data bases with statistical offices and the other involving statistical offices that are asked to put their trust in nongovernment researchers. The approaches taken reflect the different motives of and consequences for the holders of the data in providing or not providing access to external users. Some important issues raised by the papers include: who should be able to see data collected with public funds; what motivates the holders of these data to share information; what

Gerald Gates is with the Policy Office, U.S. Bureau of the Census.

methods have been developed to protect data and provide access; what are acceptable levels of risk and how do we measure them; and how do countries differ in their approach to this problem?

Confidentiality is arguably the most important right of respondents and the most important obligation of data collectors. Without it, fear will discourage many from participating in government programs or surveys. On the other side of the coin lies access. Collectors of both administrative and statistical data have an obligation to administer their programs efficiently, fairly, and openly. For the statistical office, this means using data that are available, rather than collecting similar information more than once, and requires cooperation from administrative agencies. The statistical office also has an obligation to make data widely available so that they can be fully analyzed.

The conference looked at access and confidentiality from the perspective of linked employer-employee data. Access depends on the legal obligations of the data holder to define and protect confidentiality and to permit legitimate access by others. The importance of confidentiality is apparent. The importance of providing access is less apparent, but equally great. One must remember that these data were collected with public money to serve important public interests. Providing access outside of government allows full use of the data for analyses and, as others can replicate the findings, ensures that such studies are scientifically sound. Once it is agreed that access is important, the question becomes how to provide access within the legal boundaries established by the confidentiality pledge.

Data protection involves establishing barriers to would-be intruders to make reidentification of respondents quite difficult. Some data are more difficult to protect than others. For example, microdata require different protection techniques than do tabular data. Business

micro data are arguably the most difficult to protect due to the known presence of large businesses in the sample. For instance, any user of a survey of manufacturing establishments can assume that it includes firms that dominate their industry. Similarly, files that contain entire populations lose the protections afforded by sampling. Linkages of data from different sources create additional challenges, because special precautions are necessary to protect identities from those offices that hold the source data. Data that represent unique populations or that include unusual characteristics require greater protection. Longitudinality adds to the richness of the data, but also enhances the blueprint for discovering the identity of records. Many of these more-difficult-to-protect scenarios are found in linkages of employer and employee data.

Providing access to confidential data can be accomplished in two ways: putting restrictions on the data with few or no restrictions on access, or putting restrictions on access with few or no restrictions on the data. The papers presented at the conference confirm that there is a range of acceptable methodologies for each approach, depending on the nature of the data and the research to be done. What also is clear is that, while a lot of work is going on in the field of disclosure limitation methodology, the statistical community is just beginning to think about measuring the risk remaining in public data releases and whether that level of risk is acceptable.

The conference papers traced several common themes. First, statistical offices are considering ways to disguise source data rather than add noise to, or suppress, the output. Microaggregation techniques and subsampling are being considered for difficult-to-release business microdata. These approaches have limited appeal for the statistical agency because they must be tailored to each specific research objective. Output containing derived measures, such as covariance matrices, is recommended over raw data because of its built-in protections. Sta-

tistical offices also are looking at the inherent protections in the data created by nonsampling and measurement errors. In addition, statisticians are considering reasonableness criteria that take into account the costs in money and time needed to break the protections. Targeting disclosure limitation efforts at the intruder with unlimited time and resources is unreasonable, and prevents much important research. Other statisticians are eliminating unique cases from the output.

Restricting access to the data is becoming a much more acceptable choice for providing researchers with data. In some cases, it is the only choice. Several years ago, the only option for researchers involved relocating to the site of the statistical office for as much as a year in order to conduct research on the agency's mainframe computer. With the introduction of microcomputers and networks, statistical offices have begun to offer more flexible options. Some examples are licensing users to process the data at their own site; providing users with test files to use in writing their own programs, which the statistical office will run on the confidential data; and establishing regional secure sites in which researchers may work with the confidential data. In each case, the output is sanitized to protect confidentiality. The key to these arrangements is having the legal authority to share confidential data and establishing an atmosphere in which security is paramount.

The most consistent message from conference participants was that research access to data is important and can be accomplished when a mutual trust is achieved. The administrative office must trust the statistical office and the statistical office must trust the researcher. Winning this trust involves more than promises, however. Statistical offices have strong confidentiality protections in law and can provide needed reassurance to administrative offices. What remains to be addressed is an understanding of motives and consequences that respects the interests of all parties, and a

clear knowledge of government's obligations to its citizens. Much of what is actually done results from a meeting of minds, not a legal mandate.

In the relationship between the statistical office and the researcher, legal protections are not uniformly available. Often, the statistical office is asked to trust the researcher who, it is argued, has no incentive to breach confidentiality and, in fact, would be harmed by lost access to future data. On the other hand, researchers do not necessarily have experience in handling sensitive data—an experience that makes taking extra precautions second nature. Some authors cite the need for strong codes of ethics for the statistical analysts. Some argue for legal contracts. Still others insist that laws should protect the data, regardless of who holds them. It is becoming clear that promises of a professional, subject to banishment for intentional or inadvertent breaches, are not sufficient to data providers.

In any case, this writer is encouraged to see that the techniques to provide access to data collected under a pledge of confidentiality are being shared around the world. Countries choose different options for data sharing, based primarily on cultural and organizational differences. Perhaps the most striking difference among the countries represented at the conference involved the ability to access administrative data for statistical purposes. In some countries, access is routine. In others, it is a difficult and time-consuming process. Some of this difficulty is due to decentralization of statistical functions. Some is due to motives for and consequences of sharing. Another point of departure involves sharing across borders. With the implementation of the European Union Directive on Transborder Data Flows later this year, and a similar directive coming out of the Council of Europe, the future of collaboration among the world's statistical offices is unclear. What is clear is that that some important advances are taking place in solving tough disclosure problems . □

Econometric issues

John Abowd

Several of the conference papers dealt with the econometric issues surrounding linked employer-employee data sets. One of the major themes of these papers is that using the information from entity matches improves the modeling of the target entity. There were two very interesting papers on using the same or similar information collected from both the employer and the employee to assess the quality of different measures. Others used matched employer and employee data to model outcomes that depend on the characteristic of both sides of the labor market. The final pair of papers address the design of statistical samples using the matched files.

For the benefit of the non-Americans in the audience, the main emphases of the papers that dealt with improved statistical modeling were measurements of employment flows and other outcomes using U.S. unemployment insurance (UI) data. The reason why there is so much interest in the American UI records is that they are the only data set that remotely resembles the kinds of administrative data files that many Europeans have been using routinely for research purposes for the better part of this decade. In addition, most of the U.S. work on employer-employee data sets has been done using linked unemployment insurance data.

Basically, the studies measuring unemployment duration and attrition or measuring compensation costs rely on linked data to provide critical missing information that eliminates important and untestable assumptions. The elimination of important and untestable assumptions is the value added, and the reason why linked data are critically

important. In fact, virtually all of the analytical attempts that try to get around linked-data types of problems without using linked data fail.

The employment flow studies all use the unemployment insurance system data from the ES-202 program, which is the Bureau of Labor Statistics collection of a short list of firm-level unemployment insurance-related variables: the amount of unemployment insurance-related wages that the firm paid and the size of the establishment on a particular date in the month. BLS uses these data for a variety of purposes, including updating its sampling frame for establishments, and also has begun to promote research using the ES-202 data. These data have frequently been used for measuring employment flows. In fact, studies based on these data generally demonstrate that employment flow calculations are plagued by false births and deaths of firms. Those studies link information from individual wage and benefit histories—different researchers use different sources—and then compare the employment counts or the unemployment counts thus obtained with information reported on the ES-202 form of the appropriate firm. In this way, the analysts get a handle on where the false births and deaths were occurring.

In the papers on attrition and unemployment duration presented at the conference, the frame for the survey is the unemployment insurance administrative data. This data set provides researchers with the universe of people who suffer an unemployment spell, at a particular point in time, and hence an uncensored sample of the durations. The analysts complement this information with the survey findings, and the survey's measure of unemployment durations. About half of the observations in the survey are

censored. The symposium paper shows that the linked data permit identification of the nonresponse bias in the estimation of the unemployment duration or the unemployment exit hazard, if you will, in the survey data. Furthermore, this procedure seems to perform better than any other in resolving the attrition bias. Basically, the results suggest that certain categories of attriters, the ones who do not show up in the survey but do show up in the administrative data, had systematically different unemployment duration spells, exactly what one would expect, and that there is no available instrument for correcting this problem.

There is an ongoing controversy about the role of computers and the wage structure—how do the earnings of workers who use computers compare to those of workers who do not. One approach to this problem that was explored at the symposium uses a data set whose structure is quite similar to that of data that have been collected by the Bureau of Labor Statistics and other U.S. agencies. Essentially, this involves surveying a few workers per firm, assuming that the original frame of the sample was designed to be representative of workers. The link is then made back to the employing firms, which have been identified as part of the sampling procedure, and, on the basis of this information, the analyst can go to an administrative data source and get additional information on the employer-employee relationship. So long as the analyst has some firms contributing two or three workers to the survey sample, he or she can actually use these linked data to improve the quality of variables that are measured at the level of the individual worker.

Another technique discussed at the conference was measuring year correction. Studies based on this technique used nonclassical measurement year models to show that linked information coming from the employer greatly improves the measurement year and education variables in one case, and pension information in another. Analysts actually have known for about a decade that,

John Abowd is professor of economics at Cornell University.

in order for pension information to be truly accurate, it has to be obtained from the employer. Evidence presented at the conference confirms that observation, although it suggests that employees do know more about their pensions than they used to.

Some of the papers focused on the value of using characteristics from both sides of the labor market in analyses of the employer-employee relationship. In his own paper, this writer used French data to model the differences between individual- and firm-level heterogeneity in wage determination. Another paper considered unemployment inflows. A fundamental conclusion of these studies and of others that are done on matched data—primarily matched data in European settings—is that there is an enormous amount of heterogeneity that is not explainable by the observables, even though the usual culprits—education, experience, location, and year—are accounted for. It is also not explainable by the characteristics of the firm.

The symposium papers identified the areas for future research in terms of heterogeneity due to the workers and heterogeneity due to the firms. The only way to tackle these issues, as it turns out, is with matched employer-employee data. In fact, for many of these applications, the data have to be longitudinal along at least one dimension.

The papers in the final session that this writer observed provided novice users of linked data with a pretty good set of guidelines for managing likelihood functions and sampling, so that the end result will be representative of the thing the analyst wants it to be representative of. That is not always an easy task to do, so these papers are recommended reading. ☐

Analyzing training and productivity

Lisa M. Lynch

Productivity, training, and the impact of technological change on the relative demand for skilled workers are areas of inquiry that benefit enormously from the use of linked employer-employee data. The symposium papers on these topics all extend our understanding of the dynamics between firms and workers across countries.

In the session on productivity, results of studies using data from three countries (Belgium, Norway, and Italy) suggest that employers in each of these countries have settled into multiple equilibria of wage policies and productivity strategies. For example, some Norwegian firms apparently have sucessfully adopted a low-wage, low-productivity strategy, while others have opted for a high-wage, high-productivity equilibrium. In Belgium, some firms have chosen to compress wages and have low employee turnover, while others allow greater inequality within the firm along with higher turnover. By matching microdata on firms and workers, the authors are able to document this heterogeneity of practices within countries. What is missing from these analyses is more detailed information about firms' specific human resource management (training, incentive-based pay, employee involvement in decisionmaking), and employee characteristics such as training received, previous actual work experience, and family structure. In addition, none of the samples used in these papers is from a nationally representative survey of employers.

In the session on training, the use of matched data sets helped to move the analysis beyond measuring the incidence of training within a country. A paper on the Netherlands provides evidence of the significant return to investments in training for productivity in manufacturing, based on a measure of the stock of worker skills, rather than just training incidence. A paper on the United States shows the degree of complementarity between training and other high performance workplace practices. The final paper, using data from a new longitudinal Workplace and Employee Survey (WES) by Statistics Canada, is particularly interesting because it has detailed information on both establishments and workers, with outcome measures (such as value added) for a representative national sample of employers. This survey will serve as an important model for other countries in improving linked employer employee data.

Another session dealt with the role of technological change and the relative demand for skilled labor. Linked employer-employee data allowed researchers in Canada to look at whether technology has deskilled or upskilled workplaces. They find strong evidence that "high" technology adoption has upskilled the workplace. A paper on Finland's manufacturing sector suggests that one of the reasons why relative wages between skilled and unskilled workers did not widen in the 1980s and 1990s may be that the relative supply of skilled workers was able to keep pace with the relative demand. A paper using Dutch data shows the benefits of matching employee and employer data in order to look

Lisa M. Lynch is professor of economics at Tufts University, and former chief economist, U.S .Department of Labor.

at the flows of workers into and out of firms and jobs by the degree of job complexity. What is missing from most of the data sets cited this session, with the exception of the Canadian data, is information on how workers upgrade skills outside the formal education system, and the nature of the specific skills that are required by the new technologies adopted by employers.

In sum, linking employer-employee data sheds new light on the determinants of productivity, training, and the impact of new technology on the workplace. But relying on existing data sets that have not been designed for the explicit purpose of examining these issues does not guarantee the detailed information needed for economic or policy analysis. So, what would the ideal data set be? It would be a large nationally representative longitudinal matched employer-employee survey. The survey would need to be longitudinal so that analysts could observe changes in workplace practices, such as training and compensation, in the face of changing product demand, technology, and work organization. By collecting information on both workers and employers, analysts

could determine the relative contribution of worker characteristics, versus management practices and product market conditions, on outcomes of interest. Looking at both workers and firms would also yield an understanding of the relative impact of policies, such as welfare reform, that encourage individuals to seek employment and employers to reach out and hire and train former welfare recipients.

Barring a national commitment to design a matched longitudinal survey of employers and employees in the United States, what else could improve some of the existing linked employer-employee data sets? The ongoing efforts to match existing data bases housed at different statistical agencies (both Federal and State) are critical to improving our understanding of the impact of organizational change, trade, and technology on firms and workers. In addition, designing supplemental samples of employers and employees that can be matched with existing data sets to look at issues such as workplace practices and productivity will provide valuable

information. These could be supplemental surveys administered to all survey respondents, such as the Annual Survey of Manufacturers, using as a model the techniques currently used by the Bureau of Labor Statistics in the Current Population Survey. Or researchers could be encouraged to design surveys that are administered to subsamples of existing data sets such as the Census Bureau's Standard Statistical Establishment List. The recent National Employers Survey sponsored by the National Center on the Educational Quality of the Workforce is a model of this approach.

One of the benefits of having an international symposium on linking employer and employee data is that one realizes that there would also be enormous benefits to considering ways to match observations across countries. In particular, as we try to design policies to help workers and firms succeed in an increasingly global economy, it is clear that being able to follow multinational companies as they move production and use suppliers around the world would be useful.   □

Analyzing firms, workers, and wages

Erica L. Groshen

Three sessions of the conference, encompassing nine papers, dealt with the use of linked employer-employee data for the analysis of firms, workers, and wages. The papers cover seven countries and use data from a wide variety of sources. Nevertheless, they have very similar—primarily descriptive—goals. Eight measure how employers' pay levels or structures differ from each other. The

ninth looks particularly at how men's and women's pay varies within and among firms. In essence, all test for the existence and importance of employer heterogeneity: do employers act differently from each other?

To conceptualize how linked data help analysts to understand this basic question, consider the total variation in wages. One portion of variation due to worker characteristics we can measure. This is the level of explanatory power normally achieved in wage regressions on cross-sectional household survey

data. If we knew more attributes, it might be larger. Expanding to longitudinal data allows researchers to measure even more individual wage effects—including those due to unobserved factors, such as motivation or family connections. Similarly, knowing the employer and observing its characteristics allows separation of two more nested sources of variation—the portions due to observed and unobserved employer attributes.

The conference papers ask whether the employer characteristics account for much variation beyond that encompassed by the individual characteristics. The issues include employer sorting by worker attributes—that is, employers tending to choose workers with particular characteristics and pay accordingly. This sorting produces questions of overlapping; when we have information on only one

Erica L. Groshen is an economist, Federal Reserve Bank of New York.

"side" of the labor market, overlaps are attributed solely to the side measured. With information on both sides, the sorting overlaps can be measured. If employers were simple price-takers, the employer variation would be entirely within the larger individual effects. To the extent that employers set wages independently of the market, the employer data add information about variation outside of the individual variation.

Which aspects of employer-employee data allow the researchers to measure employer variation? Even omitting complications such as interaction effects, wage variation has many components. The richer the data at their disposal, the more easily researchers can distinguish among these components. In particular, when the data include individual and employer identifiers, researchers can augment the original data with information gathered elsewhere. As research progresses and new policy issues arise, this flexibility often proves invaluable.

Summarizing broadly, the papers on firms, workers, and wages consistently find that employers do the following:

(1) *Set wages.* Employer wage differentials are an important part of wage dispersion in all cases (eight papers, covering seven countries).
(2) *Sort workers.* Worker heterogeneity accounts for a noticeable share—but far from all—of "raw" employer wage effects.
(3) *Behave systematically.* Employer effects are correlated with observed employer characteristics (size, product market, and so forth).

Overall, these findings make it clear that we can improve our understanding of wages and other labor outcomes by studying employers' activities. To take full advantage, research needs to push in three directions: pursuing policy topics, enhancing employer-employee matched data, and adding theoretical guidance to the investigations.

The finding that wages and other out-comes differ markedly and systematically among employers implies that all labor market questions and most policy issues have a key employer angle. Having established the fact of heterogeneity, the logical next step is to exploit it. In particular, such heterogeneity provides a new means to move policy analysis beyond the ambiguities of time-series studies to a comparison of the activities of different types of firms.

A wealth of new and old topics awaits investigation from the employer angle. The papers discussed here make progress in that direction—addressing questions of rising wage inequality, the male-female wage differential, and economic development. Other potential topics (including welfare-to-work transitions, outsourcing, advance notice legislation, and training methods) will expand our understanding of labor markets. Many issues from other branches of economics also are ripe for investigation, such as mergers and acquisitions, inflation, and macroeconomic policy. □

Analyzing firms, jobs, and turnover

Simon Burgess

Worker turnover, generally refers to the movement of workers around the labor market, between firms, and among the states of employment, unemployment, and inactivity. It has been known for some time that worker turnover and job turnover are "large." The advent of linked employer-employee data allows further insight into this issue, as exemplified by the eight conference papers devoted to the subject.

Simon Burgess is a professor of economics at the University of Bristol.

Labor markets are busy places. Huge numbers of workers are moving between jobs and between employment and unemployment all the time. Firms and jobs are being born, growing, declining, and dying. Even in "sclerotic" European labor markets, turnover figures are very high—far higher than would have been thought 10 years ago. Indeed, comparability worries aside, rates of job turnover appear to be as high in Europe as in North America. It is for labor markets to ensure that this reallocation of jobs and workers proceeds as smoothly and efficiently as possible. This matters for unemployment, aggregate growth, wage dispersion, and perceptions of job insecurity.

Regarding unemployment, the efficiency of the labor market in allocating workers and jobs clearly affects the duration of unemployment and of vacancies. At a deeper level, the reallocation process affects the proportion of job change that is accomplished via unemployment. While somewhat neglected by the literature, worker turnover influences aggregate growth rates by setting the speed with which workers are moved to the most profitable uses and by influencing the incentives for stable job matches. Labor reallocation also matters for earnings dispersion: if different firms pay different amounts to the same worker, then the allocation of workers to firms matters for earnings inequality. Finally, worker turnover appears to influence workers' feelings of "insecurity" (though quite how is not clear).

Almost all existing studies in this field use surveys on firms with a little information on their workers, or surveys on workers with a little information on their employers. But having information on both of the parties is very useful. Forming a job match, the unit for defining worker turnover, is a *joint* decision. While ending a job match is a unilateral decision, the firm's decision about whom to terminate may depend on worker characteristics, and a worker's decision to quit may depend on the employer's characteristics. Thus, one gets a much better understanding of the process of worker turnover with detailed information on both parties.

There are a number of examples in which this is the case. The first involves the links between flows of *jobs* and flows of *workers*. At a macro level, this matters significantly: unemployment for example is clearly about flows of workers, but how does this relate to flows of jobs? European and American job flow rates are similar, but worker flow rates are much higher in the United States. Why, and does this matter for unemployment rates? At a micro level, do firms grow by hiring more or by reducing separations? Or do they shrink by stopping hiring or raising separations? Do some firms do other things? If so, why? Does the pattern change over the business cycle?

Second, one can consider the allocation of workers to firms. Economists believe that this is not random, but driven by the decisions of optimizing actors. What types of workers work at which types of firms? Why? How does this change over time? Are some configurations better (more productive, more profitable) than others?

Third, linked data permit us to empirically examine firms' personnel policies—that is, the decisions by firms as to who to hire, what to pay them, how much turnover to tolerate, and so forth. For all but the smallest firms, some rate of worker turnover is predictable, and, to some extent, within the power of the firm to control. What is the relationship between the firm's choice of wage pre-

mium and its rate of excess worker turnover? The wage premium can be isolated only if one knows the outside opportunities of the firm's workers, which in turn can be known only if the workers' characteristics also are known. One must remember, though, that the firm chooses its work force composition, so that is also a choice variable, along with the premium.

The papers presented at the conference are largely concerned with two of these themes. The one paper that differs a little considers in some detail the definition of a job. The authors aim to carefully document the impact of changing the definition of the employing unit on measured worker turnover. They are particularly concerned to quantify the implications of changes in ownership of the firm.

The remaining papers focus on two issues: (a) the relationship between worker flows and job flows—specifically, which workers leave contracting employers and or which workers join expanding employers, and the impact of employer characteristics on worker mobility; and (b) the link between an establishment's rate of excess worker turnover (churning), and its wage premium.

Two papers are concerned with problem of European unemployment. In particular, they investigate the linkages between unemployment flows and establishment job dynamics. Because employment change obviously implies a minimum number of worker flows, this is mainly about identifying *which* workers move. By correlating an individual's chance of separating from a job with his or her establishment's (employment) growth rate, one may focus on the age profile of mobility. In France, the age-mobility relationship is very steep. Older workers in France are much more insulated from poor performance by their firms than are their counterparts in Sweden. One problem with the interpretation of the study arises because the information on the employer is limited. For example, suppose an establishment shrinks by one job, but experiences five separations, of which four are young workers

and one is an older worker. This would suggest that the young bear a disproportionate burden of the adjustment. But if the four replacement hires are all young, then this paints a different picture. A second problem is that separations are not classified as quits or layoffs. But suppose they were identified: a complete analysis would require further information on the job matches of the employer. This is because the layoff probability, given the employment change, depends on the quit rate of the respondent's coworkers.

Another paper investigates the impact of establishment-level job destruction on worker flows in Denmark. This is portrayed in two parts, the relationship between the separation rate and the job destruction rate, and second, the subsequent unemployment experience of those separating. In fact, the author of the paper finds relatively little impact of job destruction on unemployment. Again, despite having an excellent data set, this study also suffers from the fact that separations are not classified into quits and layoffs. The author deals with this by aggregating establishments into "growing," "stable," and "declining," and assuming that total separations in the "stable" firms can be taken as an approximation of a common quit rate. This rate is then subtracted from total separations in the declining group to yield an estimate of the layoff rate. Clearly crucial for this step is the assumption that the quit rate is independent of the layoff rate and job destruction, which seems questionable. The level of unemployment among workers separating from contracting firms is reported to be no higher than that among those separating from stable firms. But, given the degree of aggregation, it may be that the age make-up of the different populations is very different, and this obscures any real unemployment effect.

The next two papers had access to fuller data sets. One has separations information for the Netherlands that do distinguish between layoffs and quits and relate an individual's chance of layoff to the characteristics of the (soon-to-be-

left) employer and to the individual's own characteristics. Both age and seniority within the firm matter. Also, as theory would predict, the firm's quit rate and layoff costs both reduce layoff rates. The former result confirms the need for detailed information on the employer and employee for this sort of analysis. The latter result provides evidence on the importance of adjustment costs and employment protection legislation.

An investigation of worker and job flows in Sweden shows that many of the standard findings hold also in that country. The core of the new analysis relates to the changing composition of firms' work forces as they expand or contract. Again, because job flows imply some worker flows, this analysis is intended to identify who is most mobile. Because we cannot identify who fills which job, we look at the age and educational composition of the work force to find that contracting establishments tend to lose workers with low educational attainment and growing establishments tend to attract workers with high levels of education.

Next up is a look at the relationship between excess worker turnover (or churning), plant size, and the establishment average wage level in Denmark,

following this relationship over the business cycle. Turnover is higher in plants with lower wages and this relationship is stronger in booms. Also, turnover is higher in small plants. This study uses the actual plant wage, as opposed to the next study, which estimates a wage *premium.*

That paper investigates the relationship between employer's choice of wages and worker turnover in Norwegian data. Data on employees are used to estimate an outside wage and, hence, compute the employer's wage premium. They confirm that higher idiosyncratic wages are associated with lower turnover. The authors also estimate employer-specific wage-seniority profiles, and show that steeper profiles also are associated with lower turnover. They note that these results apply most strongly to larger firms.

Finally, the authors of the last paper use Danish data to compute establishment level turnover and relate these to the characteristics of the work force. Remaining establishment-level idiosyncrasies are explained using establishment characteristics. Then, tenure data are given the same examination. The role of unobservable heterogeneity that can be captured due to the longitudinal element of the data is highlighted.

How should these results be interpreted? Clearly, wage policies are *chosen* by firms, and chosen, among other things, to raise productivity by reducing turnover. So regressing productivity or (excess) turnover on estimated parameters of the firm's personnel policy in effect quantifies the degree to which these policies are successful. But the next step is surely to ask *why* some firms adopt some strategies and other make other choices. That is, to correlate the parameters of the personnel policy with variables describing the employer's environment, such as the technology set available, local labor market conditions, demand volatility, and the trainability of the work force.

To understand unemployment, aggregate growth and earnings dispersion, we need to understand the forces underlying worker turnover. One key element of this is the employer's choice of personnel or wage policy. The *only* feasible way to analyze this is by using linked employer-employee data. We need data on employees working at an establishment to be able to characterize the wage policy. We then need data on the employer to understand why some employers choose particular policies.   □

Program development and policy analysis

Daniel H. Weinberg

In developing linked data sets, there are three different types of matching that can go on—survey to survey, survey to administrative records, and administrative records to administrative records. (Administrative records are defined here as data collected for a separate purpose than research, such as

regulation or enforcement.) Exibit 1 presents an overview of some existing and proposed U.S. data sets created by the Census Bureau and puts the conference papers touching on the use of linked data in policy analysis into the three-way classification.

The first type of match—survey-to-survey—is represented by the Worker-Employer Characteristics Database (WECD). The WECD linked survey data from the 1990 decennial census long

form to employer data from the Annual Survey of Manufactures, the Census of Manufactures, and the Survey of Manufacturing Technology (and possibly others) via use of location information to identify the Standard Statistical Establishment List number. Another attempt will be made in 2000 using both location information and employer name, and probably industry classification. A second U.S. example is a proposed link between the March 2000 Current Population Survey (CPS) and the 2000 Decennial Census, to be used to study nonresponse.

The key issue in linking two survey data sets is the representativeness of the resulting file. One has to have either two very large surveys, or surveys with the

Daniel H. Weinberg is chief, Housing and Household Economic Statistics Division, U.S. Bureau of the Census.

Exhibit 1.	Examples of linked employee-employer data sets

Type of linked data set	Example
Survey data with survey data	**United States.** Worker-Employer Characteristics Database (1990 Census long forms with 1990 Annual Survey of Manufactures and other surveys)
	United States. March 2000 Current Population Survey with the 2000 Decennial Census [proposed]
Survey data with administrative records	**United States.** March 1991 Current Population Survey with 1991 Internal Revenue Service Individual Master File (tax returns)
	France. Annual Social Data Reports with Permanent Dynamic Sample, 1984–89
	United States. Unemployment insurance data with employer layoff survey (Mass Layoff Statistics program), other North Carolina data, 1995–97
	Japan. 1992, 1993, and 1994 Basic Survey on Wage Structure with the 1991 Censuses of Commerce and Establishments and 1992 and 1993 Census of Manufactures
Administrative records with administrative records	**United States.** 2010 decennial census (test in 2000): links Internal Revenue Service tax returns with other data, such as drivers' licenses (supplemented by some field work) [proposed]
	United States. Aid to Families with Dependent Children program with the States welfare-to-work program, and unemployment insurance program, 1992–95 (in cooperation with the State of Missouri)
	United States. Workers' compensation permanent partial disability claims data for 1991–94 with unemployment insurance records for 1989–96 (in cooperation with the State of California.)
	United States. Unemployment insurance records for various years with ES-202 establishment reporting data and other administrative data (in cooperation with the State of Texas)

same (or very similar) sampling frame. The WECD is representative of the first type, but if censuses are considered administrative records systems, none of the six conference papers fits this category.

The second type of match involves linking survey and administrative data. The Census Bureau has linked the March 1991 CPS sample to Internal Revenue Service income information from tax returns via an exact match, using Social Security numbers. A similar project is planned for the March 1997 CPS. Three of the symposium papers also dealt with this type of match.

The final type of matching is done between two (or more) sources of administrative records. The Census Bureau's Administrative Records Research Staff is looking into the possibility of an administrative record census for 2010. A major test of population coverage for such a census is being planned for several sites in conjunction with the 2000 census.

How worthwhile are such matches? The value of such matched files depends in large part on the *accuracy* of such matches, the *coverage* of the relevant population universe, and the *completeness* of the data in the two sources. The

policy *relevance* of the papers depends critically on these three aspects as well.

Investments in overcoming problems with linkage accuracy, coverage, representativeness, and data completeness will all help increase the policy relevance of this kind of research. Given that the four U.S. labor market studies all used the unemployment insurance data, it is clear that further work in this area would benefit from creating a national unemployment insurance data base, allowing researchers to find individuals who either work in a different State, or have moved between observations. ☐

International comparisons

David G. Blanchflower

Six of the papers presented at the symposium dealt with the use of linked data sets for international comparisons. Two described the data collection procedures in individual countries—the first, for the United Kingdom, and the second, for the Czech Republic. Two others addressed the difficulties involved in integrating data on employers and employees from the member countries of the European Union. The final two papers used employer and employee data on job flows and job turnover in formal econometric analyses.

Four of the papers were from representatives of statistical agencies that collect labor market data. Two gave very useful suggestions on how to harmonize data collection across member countries of the European Union. For analysts who are interested in working on comparable microdata—on workers and firms—across countries, the experience provided by the presenters is invaluable. The hope is that statistical agencies in other countries might learn from the suggestions presented and attempt to harmonize their surveys with those in the European countries as well.

A major concern of the authors of each the four papers from the statistical agencies was a series of definitional questions: how to define a job, unemployment, and so forth. They want to be able to assign people and firms to boxes and to generate an integrated system of labor accounts. This is not a term that most researchers will be familiar with. Statistical agencies would like to ensure that their estimates from employer and employee surveys overlap as much as possible. The idea is to assure the public that the published data are believable. It

would seem that problem of tracing very small or new firms to keep up-to-date sampling frames would severely complicate these efforts.

The problem of creating adequate sampling frames was well illustrated in the paper on the Czech Republic. In the early 1990s, the Czech Republic was transforming itself into a capitalist country, and there was such a rapid change in the number and size of firms that it proved almost impossible to generate a meaningful sampling frame. As an illustration, in the 1980s, about 1 percent of the Czech labor force was self-employed, but by 1996 that number had risen to more than 20 percent.

Over time, many other national labor markets also have become more complex. There has been a considerable growth across the whole of the Organization for Economic Cooperation and Development (OECD) in the proportion of young people who both attend school and work. Double-counting individuals who have two jobs or who are employees in their main activity but are self-employed in a second job is likely to cause further problems. This makes establishing a consistent set of labor accounts extremely difficult.

The last pair of papers involved econometric analyses of data. One showed the difficulties of reconciling data on worker flows using establishment, firm, and worker data separately. Useful insights can be obtained by extending the analysis to involve data that match workers to their employer. The final paper used establishment-level data to examine job turnover across a number of OECD countries. Unfortunately, consistent patterns in the data across countries are hard to find, despite gallant attempts to make the country data files comparable. As an example, measured job turnover is actually higher in Sweden, New Zealand,

France, Canada, and Australia than it is in the United States. This is likely to come as a surprise to those who equate the flexibility of U.S. labor markets with high turnover, and the inflexibility of European markets with low turnover.

Perhaps even more surprising is that measures of turnover appear to be uncorrelated with macroeconomic variables such as the unemployment rate. In a number of circumstances the correlation has the opposite sign than would have been predicted. It remains unclear if this finding is driven by inconsistencies in the way the data are derived or if there really are no systematic differences across countries. If one had consistently defined worker-employer matched data across countries, where the data were collected in a consistent way, it might well be possible to show that labor markets in countries with certain kinds of policies worked more efficiently than those with other types of policies. All one can say at the moment is that the jury is still out on whether it is good or bad to have high job or worker turnover.

As far as this writer could tell from the four papers from statistical agencies, they had no obvious need to actually link employer and employee. From the researchers' perspective, establishing a system of labor accounts does not take us very far in understanding how the labor markets actually work. Conferences such as this one present a unique opportunity for the researchers to explain to those who collect the data why they want access to matched employer-employee data, and to demonstrate that there is a payoff in improved understanding from the substantial expenditure that is involved in careful matching. Collecting and presenting the numbers in a consistent way is an essential complement to the work of the researcher, but in itself is not going to help us better understand labor market problems.

How can our understanding of how labor markets work be improved? It is necessary to get at both the firm and the workers' behavior. We need to move beyond inferring motives for particular

David G. Blanchflower is professor of economics at Dartmouth College and a research associate of the National Bureau of Economic Research.

types of behavior and begin to observe the behavior itself. We need to find out what is happening at the workplace by talking directly to the principal actors—the owners, customers, debtors, creditors, managers, and workers. It would be desirable to complement establishment-level data with detailed interviews with managers, and to group establishments together to constitute the firm. The purpose would be to examine the decisionmaking process and the internal and external factors influencing it. Does having a management team trained at Harvard give you better performance than having one trained at Dartmouth's Amos Tuck School or MIT's Sloan School? We have little or no idea. Also, there is an argument for conducting interviews with workers and their families. Apart from things like schooling and industry, we might like to know about their consumption behavior, their use of time,

and their aspirations, so that these kinds of information can be matched to worker and firm data to help us get at behavior.

There is much to be learned from international comparisons, especially in macroeconomics, where the big questions are. During the 1990s, there has been a dramatic rise in the availability of microdata on both individuals and households across a whole array of countries. Many countries have CPS-type surveys; unfortunately only a few make those data available publicly. The Eurobarometer survey series conducted in the member countries of the European Union and the International Social Survey Programme both run the same survey in a variety of countries. Long time runs of these data series are now available (from 1973 in the former case, and from 1985 in the latter). The World Bank's Living Standards Measurement Study and the Luxembourg Income Study are other programs

that provide broadly comparable microdata files across many countries. There are even comparable panel surveys of individuals available across countries, including the National Longitudinal Survey in the United States, the British Household Panel, and the German Socio-Economic Panel. There also are one or two examples of establishment surveys being conducted across countries. The best examples of these are the British Workplace Industrial Relations Survey and the Australian Workplace Industrial Relations Survey, which have similar sample designs and ask broadly similar questions. Participants at the conference learned that matched worker-firm data already are available from France, Denmark, the Netherlands, Italy, the United Kingdom, and Norway. While it remains unclear how internationally comparable these surveys are, a start obviously has been made. □