Chapter 23

EMPIRICAL STRATEGIES IN LABOR ECONOMICS

JOSHUA D. ANGRIST*

MIT and NBER

ALAN B. KRUEGER*

Princeton University and NBER

Contents

* We thank Eric Bettinger, Lucia Breierova, Kristen Harknett, Aaron Siskind, Diane Whitmore, Eric Wang, and Steve Wu for research assistance. For helpful comments and discussions we thank Alberto Abadie, Daron Acemoglu, Jere Behrman, David Card, Angus Deaton, Jeff Kling, Guido Imbens, Chris Mazingo, Steve Pischke, and Cecilia Rouse. Of course, errors and omissions are solely the work of the authors.

Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card

Abstract

This chapter provides an overview of the methodological and practical issues that arise when estimating causal relationships that are of interest to labor economists. The subject matter includes identification, data collection, and measurement problems. Four identification strategies are discussed, and five empirical examples – the effects of schooling, unions, immigration, military service, and class size – illustrate the methodological points. In discussing each example, we adopt an experimentalist perspective that emphasizes the distinction between variables that have causal effects, control variables, and outcome variables. The chapter also discusses secondary datasets, primary data collection strategies, and administrative data. The section on measurement issues focuses on recent empirical examples, presents a summary of empirical findings on the reliability of key labor market data, and briefly reviews the role of survey sampling weights and the allocation of missing values in empirical research.

JEL codes: J00; J31; C10; C81

1. Introduction

Empirical analysis is more common and relies on more diverse sources of data in labor economics than in economics more generally. Table 1, which updates Stafford's (1986, Table 7.2) survey of research in labor economics, bears out this claim. Indeed, almost 80% of recent articles published in labor economics contain some empirical work, and a striking two-thirds analyzed micro data. In the 1970s, micro data became more common in studies of the labor market than time-series data, and by the mid-1990s the use of micro data outnumbered time-series data by a factor of over ten to one. The use of micro and time-series data is more evenly split in other fields of economics.

In addition to using micro data more often, labor economists have come to rely on a wider range of datasets than other economists. The fraction of published papers using data other than what is in standard public-use files reached 38% in the period from 1994 to 1997. The files in the "all other micro datasets" category in Table 1 include primary datasets collected by individual researchers, customized public use files, administrative records, and administrative-survey links. This is noteworthy because about 10 years ago, in his *Handbook of Econometrics* survey of economic data issues, Griliches (1986, p. 1466) observed:

> ... since it is the 'badness' of the data that provides us with our living, perhaps it is not at all surprising that we have shown little interest in improving it, in getting involved in the grubby task of designing and collecting original datasets of our own.

The growing list of papers involving some sort of original data collection suggests this situation may be changing; examples include Freeman and Hall (1986), Ashenfelter and Krueger (1994), Anderson and Meyer (1994), Card and Krueger (1994, 1998), Dominitz and Manski (1997), Imbens et al. (1997), and Angrist (1998).

Labor economics has also come to be distinguished by the use of cutting edge econo-

Table 1
Percent of articles in each category[a]

	Labor economics articles					All fields
	1965–1969	1970–1974	1975–1979	1980–1983	1994–1997	1994–1997
Theory only	14	19	23	29	21	44
Micro data	11	27	45	46	66	28
Panel	1	6	21	18	31	12
Experiment	0	0	2	2	2	3
Cross-section	10	21	21	26	25	9
Micro dataset						
PSID	0	0	6	7	7	2
NLS	0	3	10	6	11	2
CPS	0	1	5	6	8	2
SEO	0	4	4	0	1	0
Census	3	5	2	0	5	1
All other micro datasets	8	14	18	27	38	21
Time series	42	27	18	16	6	19
Census tract	3	2	4	3	0	0
State	7	6	3	3	2	2
Other aggregate cross-section	14	16	8	4	6	6
Secondary data analysis	14	3	3	4	2	2
Total number of articles	106	191	257	205	197	993

[a] Notes: Figures for 1965–1983 are from Stafford (1986). Figures for 1994–1997 are based on the authors' analysis, and pertain to the first half of 1997. Following Stafford, articles are drawn from 8 leading economics journals.

metric and statistical methods. This claim is supported by the observation that outside of time-series econometrics, many and perhaps most innovations in econometric technique and style since the 1970s were motivated largely by research on labor-related topics. These innovations include sample selection models, non-parametric methods for censored data and survival analysis, quantile regression, and the renewed interest in statistical and identification problems related to instrumental variables estimators and quasi-experimental methods.

What do labor economists do with all the data they analyze? A broad distinction can be made between two types of empirical research in labor economics: descriptive analysis and causal inference. Descriptive analysis can establish facts about the labor market that need to be explained by theoretical reasoning and yield new insights into economic trends. The importance of ostensibly mundane descriptive analysis is captured by Sherlock Holmes's admonition that: "It is a capital offense to theorize before all the facts are in." A great deal of important research falls under the descriptive heading, including work on trends in poverty rates, labor force participation, and wage levels. A good

example of descriptive research of major importance is the work documenting the increase in wage dispersion in the 1980s (see e.g., Levy, 1987; Katz and Murphy, 1992; Murphy and Welch, 1992; Juhn et al., 1993). This research has inspired a vigorous search for the causes of changes in the wage distribution.

In contrast with descriptive analysis, causal inference seeks to determine the effects of particular interventions or policies, or to estimate features of the behavioral relationships suggested by economic theory. Causal inference and descriptive analysis are not competing methods; indeed, they are often complementary. In the example mentioned above, compelling evidence that wage dispersion increased in the 1980s inspired a search for causes of these changes. Causal inference is often more difficult than descriptive analysis, and consequently more controversial.

Most labor economists seem to share a common view of the importance of descriptive research, but there are differences in views regarding the role economic theory can or should play in causal modeling. This division is illustrated by the debate over social experimentation (Burtless, 1995; Heckman and Smith, 1995), in contrasting approaches to studying the impact of immigration on the earnings of natives (Card, 1990; Borjas et al., 1997), and in recent symposia illustrating alternative research styles (Angrist, 1995a; Keane and Wolpin, 1997). Research in a structuralist style relies heavily on economic theory to guide empirical work or to make predictions. Keane and Wolpin (1997, p. 111) describe the structural approach as trying to do one of two things: (a) recover the primitives of economic theory (parameters determining preferences and technology); (b) estimate decision rules derived from economic models. Given success in either of these endeavors, it is usually clear how to make causal statements and to generalize from the specific relationships and populations studied in any particular application.

An alternative to structural modeling, often called the quasi-experimental or simply the "experimentalist" approach, also uses economic theory to frame causal questions. But this approach puts front and center the problem of identifying the causal effects from specific events or situations. The problem of generalization of findings is often left to be tackled later, perhaps with the aid of economic theory or informal reasoning. Often this process involves the analysis of additional quasi-experiments, as in recent work on the returns to schooling (see, e.g., the papers surveyed by Card in this volume). In his methodological survey, Meyer (1995) describes quasi-experimental research as "an outburst of work in economics that adopts the language and conceptual framework of randomized experiments." Here, the ideal research design is explicitly taken to be a randomized trial and the observational study is offered as an attempt to approximate the force of evidence generated by an actual experiment.

In either a structural or quasi-experimental framework, the researcher's task is to estimate features of the causal relationships of interest. This chapter focuses on the *empirical strategies* commonly used to estimate features of the causal relationships that are of interest to labor economists. The chapter provides an overview of the methodological and practical issues that arise in implementing an empirical strategy. We use the term empirical strategy broadly, beginning with the statement of a causal question, and extend-

ing to identification strategies and econometric methods, selection of data sources, measurement issues, and sensitivity tests. The choice of topics was guided by our own experiences as empirical researchers and our research interests. As far as econometric methods go, however, our overview is especially selective; for the most part we ignore structural modeling since that topic is well covered elsewhere.[1] Of course, there is considerable overlap between structural and quasi-experimental approaches to causal modeling, especially when it comes to data and measurement issues. The difference is primarily one of emphasis, because structural modeling generally incorporates some assumptions about exogenous variability in certain variables and quasi-experimental analyses require some theoretical assumptions.

The attention we devote to quasi-experimental methods is also motivated by skepticism about the credibility of empirical research in economics. For example, in a critique of the practice of modern econometrics, Lester Thurow (1983, pp. 106–107) argued:

> Economic theory almost never specifies what secondary variables (other than the primary ones under investigation) should be held constant in order to isolate the primary effects. ... When we look at the impact of education on individual earnings, what else should be held constant: IQ, work effort, occupational choice, family background? Economic theory does not say. Yet the coefficients of the primary variables almost always depend on precisely what other variables are entered in the equation to "hold everything else constant."

This view of applied research strikes us as being overly pessimistic, but we agree with the focus on omitted variables. In labor economics, at least, the current popularity of quasi-experiments stems precisely from this concern: because it is typically impossible to adequately control for all relevant variables, it is often desirable to seek situations where it is reasonable to presume that the omitted variables are uncorrelated with the variables of interest. Such situations may arise if the researcher can use random assignment, or if the forces of nature or human institutions provide something close to random assignment.

The next section reviews four identification strategies that are commonly used to answer causal questions in contemporary labor economics. Five empirical examples – the effects of schooling, unions, immigration, military service, and class size – illustrate the methodological points throughout the chapter. In keeping with our experimentalist perspective, we attempt to draw clear distinctions between variables that have causal effects, control variables, and outcome variables in each example.

In Section 3 we turn to a discussion of secondary datasets and primary data collection strategies. The focus here is on data for the United States.[2] Section 3 also offers a brief review of issues that arise when conducting an original survey and suggestions for assem-

[1] See, for example, Heckman and MaCurdy's (1986) Handbook of Econometrics chapter, which "outlines the econometric framework developed by labor economists who have built theoretically motivated models to explain the new data." (p. 1918). We also have little to say about descriptive analysis because descriptive statistics are commonly discussed in statistics courses and books (see, e.g., Tukey, 1977; Tufte, 1992).

bling administrative datasets. Because existing public-use datasets have already been extensively analyzed, primary data collection is likely to be a growth industry for labor economists in the future. Following the discussion of datasets, Section 4 discusses measurement issues, including a brief review of classical models for measurement error and some extensions. Since most of this theoretical material is covered elsewhere, including the Griliches (1986) chapter mentioned previously, our focus is on topics of special interest to labor economists. This section also presents a summary of empirical findings on the reliability of labor market data, and reviews the role of survey sampling weights and the allocation of missing values in empirical research.

2. Identification strategies for causal relationships

> The object of science is the discovery of relations... of which the complex may be deduced from the simple. John Pringle Nichol, 1840 (quoted in Lord Kelvin's class notes).

2.1. The range of causal questions

The most challenging empirical questions in economics involve "what if" statements about counterfactual outcomes. Classic examples of "what if" questions in labor market research concern the effects of career decisions like college attendance, union membership, and military service. Interest in these questions is motivated by immediate policy concerns, theoretical considerations, and problems facing individual decision makers. For example, policy makers would like to know whether military cutbacks will reduce the earnings of minority men who have traditionally seen military service as a major career opportunity. Additionally, many new high school graduates would like to know what the consequences of serving in the military are likely to be for them. Finally, the theory of on-the-job training generates predictions about the relationship between time spent serving in the military and civilian earnings.

Regardless of the motivation for studying the effects of career decisions, the causal relationships at the heart of these questions involve comparisons of counterfactual states of the world. Someone – the government, an individual decision maker, or an academic economist – would like to know what outcomes would have been observed if a variable were manipulated or changed in some way. Lewis's (1986) study of the effects of union wage effects gives a concise description of this type of inference problem (p. 2): "At any given date and set of working conditions, there is for each worker a *pair* of wage figures, one for unionized status and the other for non-union status". Differences in these two

[2] Overviews of data sources for developing countries appear in Deaton's (1995) chapter in The Handbook of Development Economics, Grosh and Glewwe (1996, 1998), and Kremer (1997). We are not aware of a comprehensive survey of micro datasets for labor market research in Europe, though a few sources and studies are referenced in Westergard-Nielsen (1989).

potential outcomes define the causal effects of interest in Lewis's work, which uses regression to estimate the average gap between them.[3]

At first glance, the idea of unobserved potential outcomes seems straightforward, but in practice it is not always clear exactly how to define a counterfactual world. In the case of union status, for example, the counterfactual is likely to be ambiguous. Is the effect defined relative to a world where unionization rates are what they are now, a world where everyone is unionized, a world where everyone in the worker's firm or industry is unionized, or a world where no one is unionized? Simple micro-economic analysis suggests that the answers to these questions differ. This point is at the heart of Lewis's (1986) distinction between union *wage gaps*, which refers to causal effects on individuals, and *wage gains*, which refers to comparisons of equilibria in a world with and without unions. In practice, however, the problem of ambiguous counterfactuals is typically resolved by focusing on the consequences of hypothetical manipulations in the world as is, i.e., assuming there are no general equilibrium effects.[4]

Even if ambiguities in the definition of counterfactual states can be resolved, it is still difficult to learn about differences in counterfactual outcomes because the outcome of one scenario is all that is ever observed for any one unit of observation (e.g., a person, state, or firm). Given this basic difficulty, how do researchers learn about counterfactual states of the world in practice? In many fields, and especially in medical research, the prevailing view is that the best evidence about counterfactuals is generated by randomized trials because randomization ensures that outcomes in the control group really do capture the counterfactual for a treatment group. Thus, Federal guidelines for a new drug application require that efficacy and safety be assessed by randomly assigning the drug being studied or a placebo to treatment and control groups (Center for Drug Evaluation and Research, 1988). Leamer (1982) suggested that the absence of randomization is the main reason why econometric research often appears less convincing than research in other more experimental sciences. Randomized trials are certainly rarer in economics than in medical research, but labor economists are increasingly likely to use randomization to study the effects of labor market interventions (Passell, 1992). In fact, a recent survey of economists by Fuchs et al. (1998) finds that most labor economists place more credence in studies of the effect of government training programs on participants' income if the research design entails random assignment than if the research design is based on structural modeling.

Unfortunately, economists rarely have the opportunity to randomize variables like educational attainment, immigration, or minimum wages. Empirical researchers must therefore rely on observational studies that typically fail to generate the same force of evidence as a randomized experiment. But the object of an observational study, like an experimental study, can still be to make comparisons that provide evidence about causal

[3] See also Rubin (1974, 1977) and Holland (1986) for formal discussions of counterfactual outcomes in causal research.

[4] Lewis's (1963) earlier book discussed causal effects in terms of industries and sectors, and made a distinction between "direct" and "indirect" effects of unions similar to the distinction between wage gaps and wage gains. Heckman et al. (1998) discuss general equilibrium effects that arise in the evaluation of college tuition subsidies.

effects. Observational studies attempt to accomplish this by controlling for observable differences between comparison groups using regression or matching techniques, using pre-post comparisons on the same units of observation to reduce bias from unobserved differences, and by using instrumental variables as a source of quasi-experimental variation. Randomized trials form a conceptual benchmark for assessing the success or failure of observational study designs that make use of these ideas, even when it is clear that it may be impossible or at least impractical to study some questions using random assignment. In almost every observational study, it makes sense to ask whether the research design is a good "natural experiment."[5]

A sampling of causal questions that economists have studied without benefit of a randomized experiment appears in Table 2, which characterizes a few observational studies grouped according to the source of variation used to make causal inferences about a single "causing variable." The distinction between causing variables and control variables in Table 2 is one difference between the discussion in this chapter and traditional econometric texts, which tend to treat all variables symmetrically. The combination of a clearly labeled source of identifying variation in a causal variable and the use of a particular econometric technique to exploit this information is what we call an *identification strategy*. Studies were selected for Table 2 primarily because the source or type of variation that is being used to make causal statements is clearly labeled. The four approaches to identification described in the table are: Control for Confounding Variables, Fixed-effects and Differences-in-differences, Instrumental Variables, and Regression Discontinuity methods. This taxonomy provides an outline for the next section.

2.2. *Identification in regression models*

2.2.1. *Control for confounding variables*

Labor economists have long been concerned with the question of whether the positive association between schooling and earnings is a causal relationship. This question originates partly in the observation that people with more schooling appear to have other characteristics, such as wealthier parents, that are also associated with higher earnings. Also, the theory of human capital identifies unobserved earnings potential or "ability" as one of the principal determinants of educational attainment (see, e.g, Willis and Rosen, 1979). The most common identification strategy in research on schooling (and in economics in general) attempts to reduce bias in naive comparisons by using regression to control

[5] This point is also made by Freeman (1989). The notion that experimentation is an ideal research design for Economics goes back at least to the Cowles Commission. See, for example, Girshick and Haavelmo (1947), who wrote (p. 79): "In economic theory ... the total demand for the commodity may be considered a function of all prices and of total disposable income of all consumers. The ideal method of verifying this hypothesis and obtaining a picture of the demand function involved would be to conduct a large-scale experiment, imposing alternative prices and levels of income on the consumers and studying their reactions." Griliches and Mairesse (1998, p. 404) recently argued that the search for better natural experiments should be a cornerstone of research on production functions.

for variables that are confounded with (i.e., related to) schooling. The typical estimating equation in this context is,

$$Y_i = X_i'\beta_r + \rho_r S_i + e_i, \tag{1}$$

where Y_i is person i's log wage or earnings, X_i is a $k \times 1$ vector of control variables, including measures of ability and family background, S_i is years of educational attainment, and e_i is the regression error. The vector of population parameters is $[\beta_r' \ \rho_r]'$. The "r" subscript on the parameters signifies that these are *regression* coefficients. The question of causality concerns the interpretation of these coefficients. For example, they can always be viewed as providing the best (i.e., minimum-mean-squared-error) linear predictor of Y_i.[6] The best linear predictor need not have causal or behavioral significance; the resulting residual is uncorrelated with the regressors simply because the first-order conditions for the prediction problem are $E[e_i X_i] = 0$ and $E[e_i S_i] = 0$.

Regression estimates from five early studies of the relationship between schooling, ability, and earnings are summarized in Table 3. The first row reports estimates without ability controls while the second row reports estimates that include some kind of test score in the X-vector as a control for ability. Information about the X-variables is given in the rows labeled "ability variable" and "other controls". The first two studies, Ashenfelter and Mooney (1968) and Hansen et al. (1970) use data on individuals at the extremes of the ability distribution (graduate students and military rejects), while the others use more representative samples. Results from the last two studies, Griliches and Mason (1972) and Chamberlain (1978), are reported for models with and without family background controls.

The schooling coefficients in Table 3 are smaller than the coefficient estimates we are used to seeing in studies using more recent data (see, e.g., Card's survey in this volume). This is partly because the association between earnings and schooling has increased, partly because the samples used in the papers summarized in the table include only young men, and partly because the models used for estimation control for age and not potential experience (age-education-6). The latter parameterization leads to larger coefficient estimates since, in a linear model, the schooling coefficient controlling for age is equal to the schooling coefficient controlling for experience minus the experience coefficient. The only specification in Table 2 that controls for potential experience is from Griliches (1977), which also generates the highest estimate in the table (0.065). The corresponding estimate controlling for age is 0.022. The table also shows that controlling for ability and family background generally reduces the magnitude of schooling coefficients, implying that at least some of the association between earnings and schooling in these studies can be attributed to variables other than schooling.

What conditions must be met for regression estimates like those in Table 3 to have a

[6] The best linear predictor is the solution to $\text{Min}_{b,c} E[(Y_i - X_i'b - cS_i)^2]$ (see, e.g., White, 1980; Goldberger, 1991).

Table 2
Identification strategies in observational studies[a]

Type of identifying information	Outcome variable	Causing variable	Estimator	Reference
I. Control for confounding variables				
Control for ability and family background	Wages	Years of schooling	Regression	See Table 3
Control for past outcomes	Employment	Training programs	Regression and matching	Card and Sullivan (1988)
	Earnings		Propensity score matching	Dehejia and Wahba (1995)
			Propensity score matching	Heckman et al. (1997)
Control for military selection criteria	Earnings	Veteran status	Regression and matching	Angrist (1998)
II. Fixed-effects and differences-in-differences				
Panel data/individual changes in status	Wages	Union status	Differencing/ analysis of covariance	Freeman (1984)
	Earnings	Training programs	Differences-in-differences	Ashenfelter and Card (1985)
The Mariel Boatlift	Employment of natives	Numbers of immigrants	Differences-in-differences	Card (1990)
Changes in state laws or rules	Injury duration	Workers' Compensation benefit	Differences-in-differences	Meyer et al. (1995)
	Unemployment	Unemployment	Differences-in-differences	Solon (1985)
	Duration	Insurance benefit	Hazard models	

Change in Federal law	Employment	Anti-discrimination policy	Differences-in-differences	Heckman and Payner (1989)
Twin comparisons	Income	Years of schooling	Differencing	Behrman et al. (1980); Taubman (1976)
	Earnings		Differencing/IV	Ashenfelter and Krueger (1994)
III. Instrumental variables				
Twin births	Schooling	Fertility Teen fertility	2SLS	Rozensweig and Wolpin (1980), Bronars and Grogger (1994)
Twin births Sibling-sex composition	Labor supply	Fertility		Angrist and Evans (1998)
Year of birth Quarter of birth	Wages	Years of schooling	2SLS	Hausman and Taylor (1981) Angrist and Krueger (1991)
Draft lottery Year of birth	Earnings	Veteran status	Two-sample IV	Angrist (1990) Imbens and van der Klaauw (1995)
IV. Regression-discontinuity methods				
Financial aid thresholds	College enrollment	Financial aid	2SLS	van der Klaauw (1996)
Class-size maximum	Test scores	Class size	2SLS	Angrist and Lavy (1998)
Social Security Notch	Labor force participation	Social Security benefits	OLS	Krueger and Pischke (1992)

[a] Notes: The table lists studies classified by type of identification strategy.

Table 3
Effects of ability controls on estimates of the returns to education – early studies

	AM-68 (1)	HWC-70 (2)	G-77 (3)	G-77 (4)	GM-72 (5)	GM-72 (6)	C-78 (7)	C-78 (8)
A. OLS estimates[a]								
No ability controls	0.0205	0.0346	0.022	0.065	0.0508	0.0408	0.041	0.036
	(0.0285)	(0.007)	(0.005)	(0.005)	(0.0039)	(0.0041)	(0.006)	(0.007)
Ability controls	0.0213	0.0171	0.014	0.059	0.0433	0.0365	0.030	0.027
		(0.0074)	(0.005)	(0.005)	(0.0044)	(0.0046)	(0.007)	(0.008)
Ability variable	Math SAT	AFQT	IQ	IQ	AFQT	AFQT	IQ	IQ
Other controls	Field, degree, occupation, experience	Age/race	Age	Experience	Age, length of military service	Age, length of military service, region/urban controls, father's occupation + schooling	Age, race, region/urban controls	Age, race, region/urban controls plus family background variables
Dependant variable	Annual earnings in 1966	Annual income in 1962 (minus transfers)	Hourly wage in 1969	Hourly wage in 1969	Usual weekly earnings in 1964	Usual weekly earnings in 1964	Hourly wage in 1973	Hourly wage in 1973
Data set	Male Ph.D.'s beginning in 1958–1960	Rejected low-AFQT military applicants	NLSYM	NLSYM	1964 CPS veterans	1964 CPS veterans	NLSYM	NLSYM
N	694	2403	1362	1362	1454	1454	897	897
B. Adjusting for endogenous ability and measurement error in schooling[b]								
				0.085				0.132
				(0.009)				(0.049)

[a] Except for column (1), these are OLS schooling coefficients in semilog wage equations. Studies are Ashenfelter and Mooney (1968), Hansen et al. (1970), Griliches (1977), Griliches and Mason (1972), and Chamberlain (1978). Coefficients and standard errors in column (1) were constructed from coefficients on single year-of-education dummies, averaged and divided by average earnings. Studies were selected for the table because they report comparable specifications with and without ability controls.

[b] Estimates include the same controls as in (A). The Griliches estimates use instrumental variables and treat schooling and IQ as endogenous. The Chamberlain estimates are from a structural model that allows for measurement error and unobserved ability. The Chamberlain dependent variable in the structural model is the hourly wage in 1971.

causal interpretation? In this case, causality can be based on an underlying functional relationship that describes what a given individual would earn if he or she obtained different levels of education. This relationship may be person-specific, so we write

$$Y_{S,i} \equiv f_i(S) \tag{2}$$

to denote the potential (or latent) earnings that person i would receive after obtaining S years of education. Note that the function $f_i(S)$ has an i subscript on it while S does not. This highlights the fact that although S is a variable, it is not a random variable. The function $f_i(S)$ tells us what *i would earn* for any value of schooling, S, and not just for the realized value, S_i. In other words, $f_i(S)$ answers "what if" questions. In the context of theoretical models of the relationship between human capital and earnings, the form of $f_i(S)$ may be determined by aspects of individual behavior and/or market forces. With or without an explicit economic model for $f_i(S)$, however, we can think of this function as describing the earnings level of individual i if that person were assigned schooling level S (e.g., in an experiment).

Once the causal relationship of interest, $f_i(S)$, has been defined, it can be linked to the observed association between schooling and earnings. A convenient way to do this is with a linear model:

$$f_i(S) = \beta_0 + \rho S + \eta_i. \tag{3}$$

In addition to being linear, this equation says that the functional relationship of interest is the same for all individuals. Again, S is written without a subscript, because Eq. (3) tells us what person i would earn for any value of S and not just the realized value, S_i. The only individual-specific and random part of $f_i(S)$ is a mean-zero error component, η_i, which captures unobserved factors that determine earnings. In practice, regression estimates have a causal interpretation under weaker functional-form assumptions than this but we postpone a detailed discussion of this point until Section 2.3. Note that the earnings of someone with no schooling at all is just $\beta_0 + \eta_i$ in this model.

Substituting the observed value S_i for S in Eq. (3), we have

$$Y_i = \beta_0 + \rho S_i + \eta_i. \tag{4}$$

This looks like Eq. (1) without covariates, except that Eq. (3) explicitly associates the regression coefficients in Eq. (4) with a causal relationship. The OLS estimate of ρ in Eq. (4) has probability limit

$$C(Y_i, S_i)/V(S_i) = \rho + C(S_i, \eta_i)/V(S_i). \tag{5}$$

The term $C(S_i, \eta_i)/V(S_i)$ is the coefficient from a regression of η_i on S_i, and reflects any correlation between the realized S_i and unobserved individual earnings potential, which in this case is the same as correlation with η_i. If educational attainment were randomly assigned, as in an experiment, then we would have $C(S_i, \eta_i) = 0$ in the linear model. In practice, however, schooling is a consequence of individual decisions and institutional

forces that are likely to generate correlation between η_i and schooling. Consequently, it is not automatic that OLS provides a consistent estimate of the parameter of interest.[7]

Regression strategies attempt to overcome this problem in a very simple way: in addition to the functional form assumption for potential outcomes embodied in (3), the random part of individual earnings potential, η_i, is decomposed into a linear function of the k observable characteristics, X_i, and an error term, ε_i,

$$\eta_i = X_i'\beta + \varepsilon_i, \tag{6a}$$

where β is a vector of population regression coefficients. This means that ε_i and X_i are uncorrelated by construction. The key identifying assumption is that the observable characteristics, X_i, are the *only* reason why η_i and S_i (equivalently, $f_i(S)$ and S_i) are correlated, so

$$\mathrm{E}[S_i\varepsilon_i] = 0. \tag{6b}$$

This is the "selection on observables" assumption discussed by Barnow et al. (1981), where the regressor of interest is assumed to be determined independently of potential outcomes after accounting for a set of observable characteristics.

Continuing to maintain the selection-on-observables assumption, a consequence of (6a) and (6b) is that

$$C(Y_i, S_i)/V(S_i) = \rho + \Gamma_{SX}'\beta, \tag{7}$$

where Γ_{SX} is a $k \times 1$ vector coefficients from a regression of each element of X_i on S_i. Eq. (7) is the well known "omitted variables bias" formula, which relates a bivariate regression coefficient to the coefficient on S_i in a regression that includes additional covariates. If the omitted variables are positively related to earnings ($\beta > 0$) and positively correlated with schooling ($\Gamma_{SX} > 0$), then $C(Y_i, S_i)/V(S_i)$ is larger than the causal effect of schooling, ρ. A second consequence of (6a) and (6b) is that the OLS estimate of ρ_r in Eq. (1) is in fact consistent for the causal parameter, ρ. Note, however, that in this discussion of the problem of causal inference, $\mathrm{E}[S_i\varepsilon_i] = 0$ is an *assumption* about ε_i and S_i, whereas $\mathrm{E}[X_i\varepsilon_i] = 0$ is a statement about covariates that is true by *definition*. This suggests that it is important to distinguish error terms that represent the random parts of models for potential outcomes from mechanical decompositions where the relationship between errors and regressors has no behavioral content.

A key question in any regression study is whether the selection-on-observables assumption is plausible. This assumption clearly makes sense when there is actual random assignment conditional on X_i. Even without random assignment, however, selection-on-observables might be plausible if we know a lot about the process generating the regressor of interest. We might know, for example, that applicants to a particular college or univer-

[7] Econometric textbooks (e.g., Pindyk and Rubinfeld, 1991) sometimes refer to regression models for causal relationships as "true models," but this seems like potentially misleading terminology since non-behavioral descriptive regressions could also be described as being "true".

sity are screened using certain characteristics, but conditional on these characteristics all applicants are acceptable and chosen on a first-come/first-serve basis. This leads to a situation like the one described by Barnow et al. (1981, p. 47), where "Unbiasedness is attainable when the variables that determined the assignment are known, quantified, and included in the equation." Similarly, Angrist (1998) argued that because the military is known to screen applicants on the basis of observed characteristics, comparisons of veteran and non-veteran applicants that adjust for these characteristics have a causal interpretation. The case for selection-on-observables in a generic schooling equation is less clear cut, which is why so much attention has focused on the question of omitted-variables bias in OLS estimates of schooling coefficients.

Regression pitfalls. Schooling is not randomly assigned and, as in many other problems, we do not have detailed institutional knowledge about the process that actually determines assignment. The choice of covariates is therefore crucial. Obvious candidates include any variables that are correlated with both schooling and earnings. Test scores are good candidates because many educational institutions use tests to determine admissions and financial aid. On the other hand, it is doubtful that any particular test score is a perfect control for all the differences in earnings potential between more and less educated individuals. We see this in the fact that adding family background variables like parental income further reduces the size of schooling coefficients. A natural question about any regression control strategy is whether the estimates are highly sensitive to the inclusion of additional control variables. While one should always be wary of drawing causal inferences from observational data, sensitivity of regression results to changes in the set of control variables is an extra reason to wonder whether there might be unobserved covariates that would change the estimates even further.

The previous discussion suggests that Table 3 can be interpreted as showing that there is significant ability bias in OLS estimates of the causal effect of schooling on earnings. On the other hand, a number of concerns less obvious than omitted-variables bias suggest this conclusion may be premature. A theme of the Griliches and Chamberlain papers cited in the table is that the negative impact of ability measures on schooling coefficients is eliminated and even reversed after accounting for two factors: measurement error in the regressor of interest, and the use of endogenous test score controls that are themselves *affected by* schooling.

A standard result in the analysis of measurement error is that if variables are measured with an additive error that is uncorrelated with correctly-measured values, this imparts an attenuation bias that shrinks OLS estimates towards zero (see, e.g., Griliches, 1986; Fuller, 1987, and Section 4). The proportionate reduction is one minus the ratio of the variance of correctly-measured values to the variance of measured values. Furthermore, the inclusion of control variables that are correlated with actual values and uncorrelated with the measurement error tends to aggravate this attenuation bias. The intuition for this result is that the residual variance of true values is reduced by the inclusion of additional control variables while the residual variance of the measurement error is left unchanged. Although

studies of measurement error in education data suggest that only 10% of the variance in measured education is attributable to measurement error, it turns out that the downward bias in regression models with ability and other controls can still be substantial.[8]

A second complication raised in the early literature on regression estimates of the returns to schooling is that variables used to control for ability may be endogenous (see, e.g., Griliches and Mason, 1972, or Chamberlain, 1977). If wages and test scores are *both* outcomes that are affected by schooling, then test scores cannot play the role of an exogenous, pre-determined control variable in a wage equation. To see this, consider a simple example where the causal relationship of interest is (4), and $C(S_i, \eta_i) = 0$ so that a bivariate regression would in fact generate a consistent estimate of the causal effect. Suppose that schooling affects test scores as well as earnings, and that the effect on test scores can be expressed using the model

$$A_i = \gamma_0 + \gamma_1 S_i + \eta_{1i}. \tag{8}$$

This relationship can be interpreted as reflecting the fact that more formal schooling tends to improve test scores (so $\gamma_1 > 0$). We also assume that $C(S_i, \eta_{1i}) = 0$, so that OLS estimates of (8) would be consistent for γ_1. The question is what happens if we add the outcome variable, A_i, to the schooling equation in a mistaken (in this case) attempt to control for ability bias.

Endogeneity of A_i in this context means that η_i and η_{1i} are correlated. Since people who do well on standardized tests probably earn more for reasons other than the fact that they have more schooling, it seems reasonable to assume that $C(\eta_i, \eta_{1i}) > 0$. In this case, the coefficient on S_i in a regression of Y_i on S_i and A_i leads to an inconsistent estimate of the effect of schooling. Evaluation of probability limits shows that the OLS estimate of the schooling coefficient in a model that includes A_i converges to

$$C(Y_i, S_{Ai})/V(S_{Ai}) = \rho - \gamma_1 \varphi_{01}, \tag{9}$$

where S_{Ai} is the residual from a regression of S_i on A_i and φ_{01} is the coefficient from a regression of η_i on η_{1i} (see Appendix A for details). Since $\gamma_1 > 0$ and $\varphi_{01} > 0$, controlling for the endogenous test score variable tends to make the estimate of the returns to schooling smaller, but this is not because of any omitted-variables bias in the equation of interest. Rather it is a consequence of the bias induced by conditioning on an outcome variable.[9]

The problems of measurement error and endogenous regressors generate identification challenges that lead researchers to use methods beyond the simple regression-control framework. The most commonly employed strategies for dealing with these problems

[8] For a detailed elaboration of this point, see Welch (1975) or Griliches (1977), who notes (p. 13): "Clearly, the more variables we put into the equation which are related to the systematic components of schooling, and the better we 'protect' ourselves against various possible biases, the worse we make the errors of measurement problem." We present some new evidence on attenuation and covariates in Section 4.

[9] A similar problem may affect estimates of schooling coefficients in equations that control for occupation. Like test scores and other ability measures, occupation is itself a consequence of schooling that is probably correlated with unobserved earnings potential. For a related discussion of matching estimates, see Rosenbaum (1984).

involve instrumental variables (IV), two-stage least squares (2SLS), and latent-variable models. We briefly mention some 2SLS and latent-variable estimates, but defer a detailed discussion of 2SLS and related IV strategies until Section 2.2.3. The major practical problem in models of this type is to find valid instruments for schooling and ability. Panel B reports Griliches (1977) 2SLS estimates of Eq. (1) treating both schooling and IQ scores as endogenous. The instruments are family background measures and a second ability proxy. Chamberlain (1978) develops an alternate approach that uses panel data to identify the effects of endogenous schooling in a latent-variable model for unobserved ability. Both the Chamberlain (1978) and Griliches (1977) estimates are considerably larger than the corresponding OLS estimates, a finding which led these authors to conclude that the empirical case for a negative ability bias in schooling coefficients is much weaker than the OLS estimates suggest.[10]

2.2.2. *Fixed effects and differences-in-differences*

The main idea behind fixed-effects identification strategies is to use repeated observations on individuals (or families) to control for *unobserved* and unchanging characteristics that are related to both outcomes and causing variables. A classic field of application for fixed-effects models is the attempt to estimate the effect of union status. Suppose, for example, that we would like to know the effect of workers' union status on their wages. That is, for each worker, we imagine that there are two potential outcomes, Y_{0i}, denoting what the worker would earn if not a union member, and Y_{1i} denoting what the worker would earn as a union member. This is just like $Y_{S,i}$ in the schooling example, except that here S is the dichotomous variable, union status. The effect of union status on an individual worker is $Y_{1i} - Y_{0i}$, but this is never observed directly since only one potential outcome is ever observed for each individual at any one time.[11]

Most analyses of the union problem begin with a constant-coefficients regression model for potential outcomes, where

$$Y_{0i} = X_i'\beta + \varepsilon_i, \qquad Y_{1i} = Y_{0i} + \delta. \tag{10}$$

As in the schooling problem, Y_{0i} has been decomposed into a linear function of observed covariates, $X_i'\beta$, and a residual, ε_i, that is uncorrelated with X_i by construction. Using U_i to indicate union members, this leads to the regression equation,

$$Y_i = X_i'\beta + U_i\delta + \varepsilon_i, \tag{11}$$

which describes the causal relationship of interest.

Many researchers working in this framework have argued that union status is likely to be related to potential non-union wages, Y_{0i}, even after conditioning on covariates, X_i (see,

[10] Another strand of the literature on causal effects of schooling uses sibling data to control for family effects that are shared by siblings; early studies are by Gorseline (1932) and Taubman (1976); see also Griliches' (1979) survey. Here the problem of measurement error is paramount (see Sections 2.2.2 and 4.1).

[11] This notation for counterfactual outcomes was used by Rubin (1974, 1977). Siegfried and Sweeney (1980) and Chamberlain (1980) use a similar notation to discuss the effect of a classroom intervention on test scores.

e.g,, Abowd and Farber, 1982; or Chapters 4 and 5 in Lewis, 1986). This means that U_i is correlated with ε_i, so OLS does not estimate the causal effect, δ. An alternative to OLS uses panel datasets such as matched CPS rotation groups, the Panel Study of Income Dynamics, or the National Longitudinal Surveys, and exploits repeated observations on individuals to control for unobserved individual characteristics that are time-invariant. A well-known study in this genre is Freeman (1984).

The following model, similar to many in the literature on union status, illustrates the fixed-effects approach. Modifying the previous notation to incorporate $t = 1, \ldots, T$ observations on individuals, the fixed-effects solution for this problem begins by writing

$$Y_{0it} = X_{it}'\beta_t + \lambda\alpha_i + \xi_{it}, \tag{12}$$

where α_i is an unobserved variable for person i, that we could, in principle, include as a control if it were observed. Eq. (12) is a regression decomposition with covariates X_{it} and α_i, so ξ_{it} is uncorrelated with X_{it} and α_i by construction (X_{it} can include characteristics from different periods). The causal/regression model for panel data is now

$$Y_{it} = X_{it}'\beta_t + U_{it}\delta_t + \lambda\alpha_i + \xi_{it}, \tag{13}$$

where we have allowed the causal effect of interest to be time-varying. The identifying assumptions are that the coefficient λ does not vary across periods and that

$$\mathrm{E}[U_{it}\xi_{is}] = 0 \qquad \text{for } s = 1, \ldots, T. \tag{14}$$

In other words, whatever the source of correlation is between U_{it} and unobserved earnings potential, it can be described by an additive time-invariant covariate α_i, that has the same coefficient each period. Since differencing eliminates $\lambda\alpha_i$, OLS estimates of the differenced equation

$$Y_{it} - Y_{it-k} = X_{it}'\beta_t - X_{it-k}'\beta_{t-k} + U_{it}\delta_t - U_{it-k}\delta_{t-k} + (\xi_{it} - \xi_{it-k}) \tag{15}$$

are consistent for the parameters of interest.

Any transformation of the data that eliminates the unobserved α_i can be used to estimate the parameters of interest in this model. One of the most popular estimators in this case is the deviations-from-means or the analysis of covariance (ANCOVA) estimator, which is most often used for models where β_t and δ_t are assumed to be fixed. The analysis of covariance estimator is OLS applied to

$$Y_{it} - \bar{y}_i = \beta'(X_{it} - \bar{x}_i) + \delta(U_{it} - \bar{u}_i) + (\xi_{it} - \bar{\xi}_i), \tag{16}$$

where overbars denote person-averages. Analysis of covariance is preferable to differencing on efficiency grounds in some cases; for models with normally distributed homoscedastic errors, ANCOVA is the maximum likelihood estimator. An alternative econometric strategy for the estimation of models with individual effects uses repeated observations on cohort averages instead of repeated data on individuals. For details and examples see Ashenfelter (1984) or Deaton (1985).

Finally, note that while standard fixed-effects estimators can only be used to estimate

the effects of time-varying regressors, Hausman and Taylor (1981) have developed a hybrid panel/IV procedure for models with time-invariant regressors (like schooling). It is also worth noting that even if the causing variable of interest is time-invariant, we can use standard fixed-effects estimators to estimate *changes* in the effect of a time invariant variable. For example, the estimating equation for a model with fixed U_i is

$$Y_{it} - Y_{it-k} = X'_{it}\beta_t - X'_{it-k}\beta_{t-k} + U_i(\delta_t - \delta_{t-k}) + (\xi_{it} - \xi_{it-k}), \tag{17}$$

so $(\delta_t - \delta_{t-k})$ is identified. Angrist (1995b) used this method to estimate changes in schooling coefficients in the West Bank and Gaza Strip even though schooling is approximately time-invariant.

Fixed-effects pitfalls. The use of panel data to eliminate bias from unobserved individual effects raises a number of econometric and statistical issues. Since this material is covered in Chamberlain's (1984) chapter in *The Handbook of Econometrics,* we limit our discussion to an overview of problems that have been of particular concern to labor economists. First, analysis of covariance and differencing estimators are not consistent when the process determining U_{it} involves lagged dependent variables. This issue comes up in the analysis of training programs because participants often experience a pre-program decline in earnings, a fact first noted by Ashenfelter (1978). If past earnings are observed and there are no unobserved individual effects, the simplest strategy is to control for past earnings either by including lagged earnings as a regressor or in matched treatment-control comparisons (see, e.g., Dehejia and Wahba, 1995; Heckman et al., 1997). In fact, the question of whether trainees and a candidate comparison group have similar lagged outcomes is sometimes seen as a litmus test for the legitimacy of the comparison group in the evaluation of training programs (see, e.g., Heckman and Hotz, 1989).

A problem arises in this context, however, when the process determining U_{it} involves past outcomes and an unobserved covariate, α_i. Ashenfelter and Card (1985) discuss an example involving the effect of training on the Social Security-taxable earnings of trainees under the Comprehensive Employment and Training Act (CETA). They propose a model of training status where individuals who enter CETA training in year τ do so because they have low α_i *and* their earnings were unusually low in year $\tau - 1$. Suppose initially we ignore the fact that training status involves past earnings, and estimate an equation like (15). Ignoring other covariates, this amounts to comparing the earnings growth of trainees and controls. But whatever the true program effect is, the growth in the earnings of CETA trainees from year $\tau - 1$ to year $\tau + 1$ will tend to be larger than the earnings growth in a candidate control group simply because of regression-to-the-mean. This generates a spurious positive training effect and the conventional differencing method breaks down.[12]

A natural strategy for dealing with this problem might seem to be to add $Y_{i\tau-1}$ to the list of control variables, and then difference away the fixed effect in a model with $Y_{i\tau-1}$ as regressor. The problem is that now any transformation that eliminates the fixed effect will

[12] Deviations-from-means estimators are also biased in this case.

leave at least one regressor – the lagged dependent variable – correlated with the errors in the transformed equation. Although the lagged dependent variable is not the regressor of interest, the fact that it is correlated with the error term in the transformed equation means that the estimate of the coefficient on $U_{i\tau+1}$ is biased as well. A detailed description of this problem, and the solutions that have been proposed for it, raises technical issues beyond the scope of this chapter. A useful reference is Nickell, 1981, especially pp. 1423–1424. See also Card and Sullivan's (1988) study of the effect of CETA training on the employment rates of trainees, which reports both fixed-effects estimates and matching estimates that control for lagged outcomes.

A second potential problem with fixed-effects estimators is that bias from measurement error is usually aggravated by transformations that eliminate the individual effects (see, e.g., Freeman, 1984; Griliches and Hausman, 1986). This fact may explain why fixed-effects estimates often turn out to be smaller than estimates in levels. Finally, perhaps the most important problem with this approach is that the assumption that omitted variables can be captured by an additive, time-invariant individual effect is arbitrary in the sense that it usually does not come from economic theory or from information about the relevant institutions.[13] On the other hand, the fixed-effects approach has intuitive appeal ("whatever makes us special is timeless") and an identification payoff that is hard to beat. Also, fixed-effects models lend themselves to a variety of specification tests. See, for example, Ashenfelter and Card (1985), Chamberlain (1984), Griliches and Hausman (1986), Angrist and Newey (1991), and Jakubson (1991). Many of these studies also focus on the union example.

The differences-in-differences (DD) model. Differences-in-differences strategies are simple panel-data methods applied to sets of group means in cases when certain groups are exposed to the causing variable of interest and others are not. This approach, which is transparent and often at least superficially plausible, is well-suited to estimating the effect of sharp changes in the economic environment or changes in government policy. The DD method has been used in hundreds of studies in economics, especially in the last two decades, but the basic idea has a long history. An early example in labor economics is Lester (1946), who used the differences-in-differences technique to study employment effects of minimum wages.[14]

The DD approach is explained here using Card's (1990) study of the effect of immigration on the employment of natives as an example. Some observers have argued that immigration is undesirable because low-skilled immigrants may displace low-skilled or less-educated US citizens in the labor market. Anecdotal evidence for this claim includes newspaper accounts of hostility between immigrants and natives in some cities, but the empirical evidence is inconclusive. See Friedberg and Hunt (1995) for a survey of research on this question. As in our earlier examples, the object of research on immigration is to

[13] An exception is the literature on life-cycle labor supply (e.g., MaCurdy, 1981; Altonji, 1986).

[14] The DD method goes by different names in different fields. Psychologist Campbell (1969) calls it the "non-equivalent control-group pretest-posttest design."

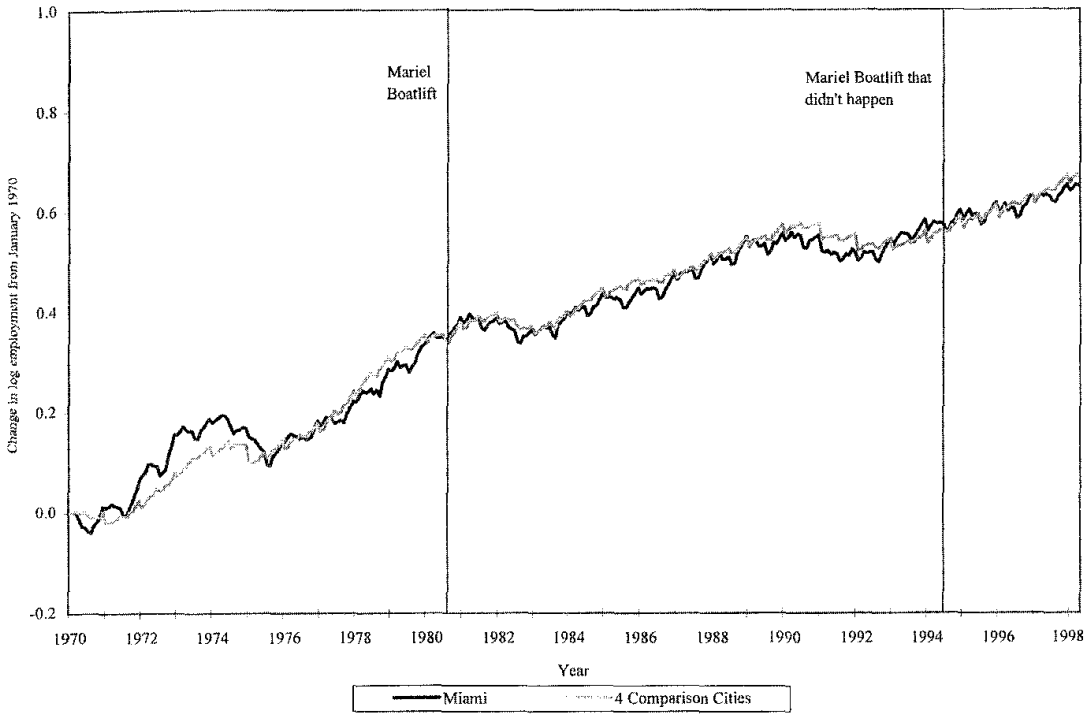

Fig. 1. Changes in employment in Miami and comparison cities. *Source*: authors' calculations from BLS State and Area Employment, Hours, and Earnings Establishment Survey.

find some sort of comparison that provides a compelling answer to "what if" questions about the consequences of immigration.

Card's study used a sudden large-scale migration from Cuba to Miami known as the Mariel Boatlift to make comparisons and answer counterfactual questions about the consequences of immigration. In particular, Card asks whether the Mariel immigration, which increased the Miami labor force by about 7% between May and September of 1980, reduced the employment or wages of non-immigrant groups. An important component of this identification strategy is the selection of comparison cities that can be used to estimate *what would have happened* in the Miami labor market absent the Mariel immigration.

The comparison cities Card used in the Mariel Boatlift study were Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg. These cities were chosen because, like Miami, they have large Black and Hispanic populations and because discussions of the impact of immigrants often focuses on the consequences for minorities. Most importantly, these cities appear to have employment trends similar to those in Miami at least since 1976. This is documented in Fig. 1, which is similar to a figure in Card's (1989) working paper that did not appear in the published version of his study. The figure plots monthly observations on the log of employment in Miami and the four comparison cities from 1970 through 1998. The two series, which are from BLS establishment data, have been normalized by subtracting the 1970 value.

Table 4
Differences-in-differences estimates of the effect of immigration on unemployment[a]

	Group	Year		
		1979 (1)	1981 (2)	1981–1979 (3)
	Whites			
(1)	Miami	5.1 (1.1)	3.9 (0.9)	−1.2 (1.4)
(2)	Comparison cities	4.4 (0.3)	4.3 (0.3)	−0.1 (0.4)
(3)	Miami-Comparison Difference	0.7 (1.1)	−0.4 (0.95)	−1.1 (1.5)
	Blacks			
(4)	Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5)	Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6)	Miami-Comparison Difference	−2.0 (1.9)	−3.0 (2.0)	−1.0 (2.8)

[a] Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

Table 4 illustrates DD estimation of the effect of Boatlift immigrants on unemployment rates, separately for whites and blacks. The first column reports unemployment rates in 1979, the second column reports unemployment rates in 1981, and the third column reports the 1981–1979 difference. The rows give numbers for Miami, the comparison cities, and the difference between them. For example, between 1981 and 1979, the unemployment rate for Blacks in Miami rose by about 1.3%, though this change is not significant. Unemployment rates in the comparisons cities rose even more, by 2.3%. The difference in these two changes, −1.0%, is a DD estimate of the effect of the Mariel immigrants on the unemployment rate of Blacks in Miami. In this case, the estimated effect on the unemployment rate is actually negative, though not significantly different from zero.

The rationale for this double-differencing strategy can be explained in terms of restrictions on the conditional mean function for potential outcomes in the absence of immigration. As in the union example, let Y_{0i} be i's employment status in the absence of immigration and let Y_{1i} be i's employment status if the Mariel immigrants come to i's city. The unemployment rate in city c in year t is $\mathrm{E}[Y_{0i} \mid c, t]$, with no immigration wave, and $\mathrm{E}[Y_{1i} \mid c, t]$ if there is an immigration wave. In practice, we know that the Mariel immigration happened in Miami in 1980, so that the only values of $\mathrm{E}[Y_{1i} \mid c, t]$ we get to see are for c = Miami and $t > 1980$. The Mariel Boatlift study uses the comparison cities to estimate the counterfactual average, $\mathrm{E}[Y_{0i} \mid c = \text{Miami}, t > 1980]$, i.e., what the unemployment rate in Miami would have been if the Mariel immigrants had not come.

The DD method identifies causal effects by restricting the conditional mean function $\mathrm{E}[Y_{0i} \mid c, t]$ in a particular way. Specifically, suppose that

$$\mathrm{E}[Y_{0i} \mid c, t] = \beta_t + \gamma_c, \tag{18}$$

that is, in the absence of immigration, unemployment rates can be written as the sum of a year effect that is common to cities and a city effect that is fixed over time. The additive model pertains to $E[Y_{0i} \mid c, t]$ instead of Y_{0i} directly because the latter is a zero/one variable. Suppose also that the effect of the Mariel immigration is simply to add a constant to $E[Y_{0i} \mid c, t]$, so that

$$E[Y_{1i} \mid c, t] = E[Y_{0i} \mid c, t] + \delta. \tag{19}$$

This means the employment status of individuals living in Miami and the comparison cities in 1979 and 1981 can be written as

$$Y_i = \beta_t + \gamma_c + \delta M_i + \varepsilon_i, \tag{20}$$

where $E[\varepsilon_i \mid c, t] = 0$ and M_i is a dummy variable that equals 1 if i was exposed to the Mariel immigration by living in Miami after 1980. Differencing unemployment rates across cities and years gives

$$\{E[Y_i \mid c = \text{Miami}, t = 1981] - E[Y_i \mid c = \text{Comparison}, t = 1981]\}$$

$$-\{E[Y_i \mid c = \text{Miami}, t = 1979] - E[Y_i \mid c = \text{Comparison}, t = 1979]\} = \delta. \tag{21}$$

Note that M_i in Eq. (20) is an interaction term equal to the product of a dummy indicating observations after 1980 and a dummy indicating residence in Miami. The DD estimate can therefore also be computed in a regression of stacked micro data for cities and years. The regressors consist of dummies for years, dummies for cities, and M_i. Similarly, a regression-adjusted version of the DD estimator adds a vector of individual characteristics, X_i to Eq. (20):

$$Y_i = X_i'\beta_0 + \beta_t + \gamma_c + \delta M_i + \varepsilon_i,$$

where β_0 is now a vector of coefficients that includes a constant. Controlling for X_i changes the estimate of δ only if M_i are X_i are correlated, conditional on city and year main-effects. (In practice, δ might be allowed to differ for different post-treatment years.)

DD pitfalls. Like any other identification strategy, DD is not guaranteed to identify the causal effect of interest. Meyer (1995) and Campbell (1969) outline a range of threats to the causal interpretation of DD estimates. The key identifying assumption is clearly that interaction terms are zero in the absence of the intervention. In fact, it is easy to imagine that unemployment rates evolve differently across cities regardless of shocks like the Mariel immigration. One way to test this is to compare trends in outcomes before or after the event of interest. As noted above, the comparison cities in this case were chosen partly on the basis of Fig. 1, which shows that the comparison cities exhibited a pattern of economic growth similar to that in Miami. Identification of causal effects using city/year comparisons clearly turns on the assumption that the two sets of cities would have had the same employment trends had the boatlift not occurred. We introduce some new evidence on this question in Section 2.4.

2.2.3. Instrumental variables

Identification strategies based on instrumental variables can be thought of as a scheme for using exogenous field variation to approximate randomized trials. Again, we illustrate with an example where there is an underlying causal relationship, in this case the effect of Vietnam-era military service on the earnings of veterans later in life. In the 1960s and early 1970s, young men were at risk of being drafted for military service. Policy makers, veterans groups, and economists have long been interested in what the consequences of this military service were for the men involved. A belief that military service is a burden helped to mobilize support for a range of veterans' programs and for ending the draft in 1973 (see, e.g., Taussig, 1974). Concerns about fairness also led to the institution of a draft lottery in 1970 that was used to determine priority for conscription in cohorts of 19-year-olds. This lottery was used by Hearst et al. (1986) to estimate the effects of military service on civilian mortality and by Angrist (1990) to construct IV estimates of the effects of military service on civilian earnings.

As in the union problem, the causal relationship of interest is based on the notion that there are two potential outcomes, Y_{0i}, denoting what someone from the Vietnam-era cohort would earn if they did not serve in the military and Y_{1i}, denoting earnings as a veteran. Again, using a constant-effects model for potential outcomes, we can write

$$Y_{0i} = \beta_0 + \eta_i, \qquad Y_{1i} = Y_{0i} + \delta, \tag{22}$$

where $\beta_0 \equiv \mathrm{E}[Y_{0i}]$. The constant effect δ is the parameter of interest. IV estimates have a causal interpretation under weaker assumptions than this, but we postpone a discussion of this point until Section 2.3. As in the union and schooling problems, η_i is the random part of potential outcomes, but at this point there are no observed covariates in the model for Y_{0i}. Using D_i to indicate veteran status, the causal relationship between veteran status and earnings can be written

$$Y_i = \beta_0 + D_i\delta + \eta_i. \tag{23}$$

Also as in the union and schooling problems, there is a concern that since D_i is not randomly assigned, a comparison of all veterans to all non-veterans would not identify δ. Suppose, for example, that individuals with low civilian earnings potential are more likely to serve in the military, either because they want to or because they are less adept at obtaining deferments. Then the regression coefficient in (23), which is also the difference in means by veteran status, is biased downwards:

$$\mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[Y_i \mid D_i = 0] = \delta + \{\mathrm{E}[\eta_i \mid D_i = 1] - \mathrm{E}[\eta_i \mid D_i = 0]\} < \delta. \tag{24}$$

IV methods can eliminate this sort of bias if the researcher has access to an instrumental variable Z_i, that is correlated with D_i, but otherwise independent of potential outcomes. A natural instrument is draft-eligibility status, since this was determined by a lottery over birthdays. In particular, in each year from 1970 to 1972, random sequence numbers (RSNs) were randomly assigned to each birth date in cohorts of 19-year-olds. Men with lottery numbers below an eligibility ceiling were eligible for the draft, while men with

Table 5
IV estimates of the effects of military service on white men[a]

Earnings year	Earnings		Veteran status		Wald estimate of veteran effect
	Mean (1)	Eligibility effect (2)	Mean (3)	Eligibility effect (4)	(5)
A. Men born 1950					
1981	16461	−435.8 (210.5)	0.267	0.159 (0.040)	−2741 (1324)
1970	2758	−233.8 (39.7)			−1470 (250)
1969	2299	−2.0 (34.5)			
B. Men born 1951					
1981	16049	−358.3 (203.6)	0.197	0.136 (0.043)	−2635 (1497)
1971	2947	−298.2 (41.7)			−2193 (307)
1970	2379	−44.8 (36.7)			
C. Men born 1953 (no one drafted)					
1981	14762	34.3 (199.0)	0.130	0.043 (0.037)	No first stage
1972	3989	−56.5 (54.8)			
1971	2803	2.1 (42.9)			

[a] Note: Adapted from Angrist (1990, Tables 2 and 3), and unpublished author tabulations. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 observations with earnings in each cohort.

numbers above the ceiling could not be drafted. In practice, many draft-eligible men were still exempted from service for health or other reasons, while many men who were draft-exempt nevertheless volunteered for service. So veteran status was not completely determined by randomized draft-eligibility; eligibility and veteran status are merely correlated.

For white men who were at risk of being drafted in the 1970–1971 draft lotteries, draft-eligibility is clearly associated with lower earnings in years after the lottery. This can be seen in Table 5, which reports the effect of randomized draft-eligibility status on Social Security earnings in column (2). Column (1) shows average annual earnings for purposes of comparison. These data are the FICA-taxable earnings of men with earnings covered by OASDI (for details see the appendix to Angrist (1990)). For men born in 1950, there are significant negative effects of eligibility status on earnings in 1970, when these men were being drafted, and in 1981, 10 years later. In contrast, there is no evidence of an association between eligibility status and earnings in 1969, the year the lottery drawing for men born in 1950 was held but before anyone born in 1950 was actually drafted. Similarly, for men born in 1951, there are large negative eligibility effects in 1971 and 1981, but no evidence of an effect in 1970, before anyone born in 1951 was actually drafted. The timing of these effects suggests that the negative association between draft-eligibility status and earnings is caused by the military service of draft-eligible men.

Because eligibility status was randomly assigned, the claim that the estimates in column

(2) represent the effect of *draft-eligibility* on earnings seems uncontroversial. How do we go from the effect of draft-eligibility to the effect of veteran status? The identifying assumption in this case is that Z_i is independent of potential earnings, which in this case means that Z_i is uncorrelated with η_i. It follows immediately that $\delta = C(Y_i, Z_i)/C(D_i, Z_i)$. The intuition here is that only part of the variation in D_i – the part that is associated with Z_i – is used to identify the parameter of interest (δ). Because Z_i is a binary variable, we also have

$$\delta = \{\mathrm{E}[Y_i \mid Z_i = 1] - \mathrm{E}[Y_i \mid Z_i = 0]\}/\{\mathrm{E}[D \mid Z_i = 1] - \mathrm{E}[D \mid Z_i = 0]\}. \tag{25}$$

The sample analog of (25) is the Wald (1940) estimator that was originally applied to measurement error problems.[15] Note that we could have arrived at (25) directly, i.e., without reference to the $C(Y_i, Z_i)/C(D_i, Z_i)$ formula, because the independence of Z_i and potential outcomes implies $\mathrm{E}[\eta_i \mid Z_i] = 0$. In this case, the Wald estimator is simply the difference in mean earnings between draft-eligible and ineligible men, divided by the difference in the probability of serving in the military between draft-eligible and ineligible men.

The only information required to go from draft-eligibility effects to veteran-status effects is the denominator of the Wald estimator, which is the effect of draft-eligibility on the probability of serving in the military. This information, which comes from the Survey of Income and Program Participation (SIPP), appears in column (4) of Table 5.[16] For earnings in 1981, long after most Vietnam-era servicemen were discharged from the military, the Wald estimates of the effect of military service amount to about 16% of earnings. Effects for men while in the service are much larger (in percentage terms), which is not surprising since military pay during the conscription era was extremely low.

An important feature of the Wald/IV estimator is that the identifying assumptions are easy to assess and interpret. The basic claim justifying a causal interpretation of the estimator is that the only reason why $\mathrm{E}[Y_i \mid Z_i]$ varies with Z_i is because $\mathrm{E}[D_i \mid Z_i]$ varies with Z_i. A simple way to check this is to look for an association between Z_i and personal characteristics that should not be affected by D_i, such as age, race, sex, or any other characteristic that was determined before D_i was determined. Another useful check is to look for an association between the instrument and outcomes in samples where there is no reason for such a relationship. If it really is true that the only reason why draft-eligibility affects earnings is veteran status, then in samples where eligibility status is unrelated to veteran status, draft-eligibility effects on earnings should be zero. This idea is illustrated in section C of Table 5, which reports estimates for men born in 1953. Although there was a lottery drawing which assigned RSNs to the 1953 cohort in February of 1972, no one born in 1953 was actually drafted (the draft officially ended in July 1973). This is reflected in

[15] The relationship between IV with binary instruments and Wald estimators was first noted by Durbin (1954).

[16] In this case, the denominator of the Wald estimates does not come from the same data set as the numerator since the Social Security administration has no information on veteran status. As long as the information used to estimate the numerator and denominator are representative of the same population, the resulting two-sample estimate will be consistent. The econometrics behind this two-sample approach to IV are discussed briefly in Section 3.4.

the insignificant first-stage relationship between veteran status and draft-eligibility for men born in 1953 (defined using the 1952 RSN cutoff of 95). In fact, there is no significant relationship between Y_i and Z_i for this cohort as well. Evidence of a relationship between Z_i and Y_i would cast doubt on the claim that the only reason for draft-eligibility effects is the military service of the men who were draft-eligible. We discuss other specification checks of this type in Section 2.4.

So far the discussion of IV has allowed for only three variables: the outcome, the endogenous regressor, and the instrument. In many cases, the assumption that $E[Z_i\eta_i] = 0$ is more plausible after controlling for a vector of covariates, X_i. Decomposing the random part of potential outcomes in (22) into a linear function of k control variables and an error term so that $\eta_i = X_i'\beta + \varepsilon_i$ as before, the resulting estimating equation is

$$Y_i = X_i'\beta + D_i\delta + \varepsilon_i. \tag{26}$$

Note that since ε_i is defined as the residual from a regression of η_i on X_i, it is uncorrelated with X_i by construction. In contrast with δ, which has a causal interpretation., the coefficient vector β is not meant to capture the causal effect of the X-variables. As in the discussion of regression, we find it useful to distinguish between control variables and causing variables when using instrumental variables.

Equations like (26) are typically estimated using 2SLS, i.e., by substituting the fitted values from a first-stage regression of D_i on X_i and Z_i. In some applications, more than one instrument is available to estimate the single causal effect, δ. 2SLS accommodates this situation by including all the instruments in the first-stage equation. The combination of multiple instruments to produce a single estimate makes the most sense in a constant-coefficients framework. The assumptions of instrument validity and constant coefficients can also be tested in this case (see, e.g., Hansen, 1982; Newey, 1985). In a more general setting with heterogeneous potential outcomes, different instruments estimate different weighted averages of the difference $Y_{1i} - Y_{0i}$ (Imbens and Angrist, 1994). We return to this point in Section 2.3.

IV pitfalls. The most important IV pitfall is the validity of instruments, i.e., the possibility that η_i and Z_i are correlated. Suppose, for example, that Z_i is related to the vector of control variables, X_i, and we do not account for this in the estimation. The Wald/IV estimator in that case has probability limit

$$\delta + \beta'\{E[X_i \mid Z_i = 1] - E[X_i \mid Z_i = 0]\}/\{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]\}.$$

This is a version of the omitted-variables bias formula for IV. The formula captures the fact that "a little omitted variables bias can go a long way" in an IV setting, because the association between X_i and Z_i gets multiplied by $\{E[D \mid Z = 1] - E[D \mid Z = 0]\}^{-1}$. In the draft lottery case, for example, any draft-eligibility effects on omitted variables get multiplied by about $1/0.15 \approx 6.7$.

A second important point about bias in instrumental variables estimates is that random assignment alone does not guarantee a valid instrument. Suppose, for example, that in

addition to being more likely to serve in the military, men with low draft-lottery numbers were more likely to stay in college so as to extend a draft deferment. This fact will create a relationship between potential earnings and Z_i even for non-veterans, in which case IV yields biased estimates of the causal effect of veteran status. Random assignment of Z_i does not rule out this sort of bias since draft-eligibility can in principle have consequences in addition to influencing the probability of being a veteran. In other words, while the randomization of Z_i ensures that the reduced-form relationship between Y_i and Z_i represents the causal effect of draft eligibility on earnings, it does not guarantee that the only reason for this relationship is D_i. The distinction between the assumed random assignment of an instrument and the assumption that a single causal mechanism explains effects on outcomes is discussed in greater detail by Angrist et al. (1996).

Finally, the use of 2SLS to combine many different instruments can lead to finite-sample bias. The standard inference framework for 2SLS uses asymptotic theory, i.e., inference is based on approximations that are increasingly accurate as sample sizes grow. Typically, inferences about OLS coefficient estimates also use asymptotic theory since the relevant finite-sample theory assumes normally distributed errors. A key difference between IV and OLS estimators, however, is that even without normality OLS provides an unbiased estimate of population regression coefficients (provided the regression function is linear; see, e.g., Goldberger, 1991, Chapter 13). In contrast, IV estimators are consistent but not unbiased. This means that under repeated sampling with a fixed sample size, IV estimates may systematically deviate from the corresponding population parameter.[17] Moreover, this bias tends to pull IV estimates towards the corresponding OLS estimates, giving a misleading impression of similarity between the two sets of estimates (see, e.g., Sawa, 1969).

How bad is the finite-sample bias of an IV estimate likely to be? In practice, this largely turns on the number of instruments relative to the sample size, and the strength of the first-stage relationship. Other things equal, more instruments, smaller samples, and weaker instruments each mean more bias (see, e.g., Buse, 1992). The fact that IV estimates can be noticeably biased even with very large datasets was highlighted by Bound et al. (1995), which focuses on Angrist and Krueger's (1991) compulsory schooling study. This study uses hundreds of thousands of observations from Census data to implement an instrumental variables strategy for estimating the returns to schooling. The instruments are quarter-of-birth dummies since children born earlier in the year enter school at an older age and are therefore allowed to drop out of school (typically on their 16th birthday) after having completed less schooling. Some of the 2SLS estimates in Angrist and Krueger (1991) use many quarter-of-birth/state-of-birth interaction terms in addition to quarter-of-birth main effects as instruments. Since the underlying first-stage relationship in these models is not very strong, there is potential for substantial bias towards the OLS estimates in these specifications.

[17] A similar problem arises with Generalized Method of Moments estimation of models for covariance structures (see Altonji and Segal, 1996).

Bound et al. (1995) discuss the question of how strong a first-stage relationship has to be in order to minimize the potential for bias. They suggest using the F-statistic for the joint significance of the excluded instruments in the first-stage equation as a diagnostic. This is clearly sensible, since, if the instruments are so weak that the relationship between instruments and endogenous regressors cannot be detected with a reasonably high level of confidence, then the instruments should probably be abandoned. On the other hand, Hall et al. (1996) point out that this sort of selection procedure also has the potential to induce a bias from pre-testing.

A simple alternative (or complement) to screening on the first-stage F is to use estimators that are approximately unbiased. One such estimator is Limited Information Likelihood (LIML), which has no integral moments but is nevertheless median-unbiased. This means that the sampling distribution is centered at the population parameter.[18] In fact, any just-identified 2SLS estimator is also median-unbiased since 2SLS and LIML are identical for just-identified models. The class of median-unbiased instrumental variables estimators therefore includes the Wald estimator discussed in the previous section. Other approximately unbiased estimators are based on procedures that estimate the first-stage and second-stage relationship in separate datasets. This includes Two-Sample and Split-Sample IV (Angrist and Krueger, 1992, 1995), and an IV estimator that uses a set of leave-one-out first-stage estimates called Jackknife Instrumental Variables (Angrist et al., 1998).[19] An earlier literature discussed combination estimators that are approximately unbiased (see, e.g., Sawa, 1973). Recently, Chamberlain and Imbens (1996) introduced a Bayesian IV estimator that also avoids bias.

A final and related point is that the reduced-form OLS regression of the dependent variable on exogenous covariates and instruments is unbiased in a sample of any size, regardless of the power of the instrument (assuming the reduced form is linear). This is important because the reduced form effects of the instrument on the dependent variable are proportional to the coefficient on the endogenous regressor in the equation of interest. The existence of a causal relationship between the endogenous regressor and dependent variable can therefore be gauged through the reduced form without fear of finite-sample bias even if the instruments are weak.

2.2.4. Regression-discontinuity designs

The Latin motto Marshall placed on the title page of his *Principles of Economics* (Marshall, 1890) is, *"Natura non facit saltum,"* which means: "Nature does not make

[18] Anderson et al. (1982, p. 1026) report this in a Monte Carlo study: "To summarize, the most important conclusion from the study of LIML and 2SLS estimators is that the 2SLS estimator can be badly biased and in that sense its use is risky. The LIML estimator, on the other hand, has a little more variability with a slight chance of extreme values, but its distribution is centered at the parameter value." Similar Monte Carlo results and a variety of analytic justifications for the approximate unbiasedness of LIML appear in Bekker (1994), Donald and Newey (1997), Staiger and Stock (1997), and Angrist et al. (1998).

[19] A SAS program that computes Split-Sample and Jackknife IV is available at http://www.wws.princeton.edu/faculty/krueger.html.

jumps." Marshall argues that most economic behavior evolves gradually enough to be modeled or explained. The notion that human behavior is typically orderly or smooth is at the heart of a research strategy called the regression-discontinuity (RD) design. RD methods use some sort of parametric or semi-parametric model to control for smooth or gradually evolving trends, inferring causality when the variable of interest changes abruptly for non-behavioral or arbitrary reasons. There are a number of ways to implement this idea in practice. We focus here on an approach that can viewed as a hybrid regression-control/IV identification strategy. This is distinct from conventional IV strategies because the instruments are derived explicitly from non-linearities or discontinuities in the relationship between the regressor of interest and a control variable. Recent applications of the RD idea include van der Klauuw's (1996) study of financial aid awards; Angrist and Lavy's (1998) study of class size; and Hahn et al.'s (1998) study of anti-discrimination laws.

The RD idea originated with Campbell (1969), who discussed the (theoretical) problem of how to identify the causal effect of a treatment that is assigned as a deterministic function of an observed covariate which is also related to the outcomes of interest. Campbell used the example of estimating the effect of National Merit scholarships on applicants' later academic achievement. He argued that if there is a threshold value of past achievement that determines whether an award is made, then one can control for any smooth function of past achievement and still estimate the effect of the award at the point of discontinuity. This is done by matching discontinuities or non-linearities in the relationship between outcomes and past achievement to discontinuities or non-linearities in the relationship between awards and past achievement.[20] van der Klauuw (1996) pointed out the link between Campbell's suggestion and IV, and used this idea to estimate the effect of financial aid awards on college enrollment.[21]

Angrist and Lavy (1998) used RD to estimate the effects of class size on pupil test scores in Israeli public schools, where class size is officially capped at 40. They refer to the cap of 40 as "Maimonides' Rule," after the 12th Century Talmudic scholar Maimonides, who first proposed it. According to Maimonides' Rule, class size increases one-for-one with enrollment until 40 pupils are enrolled, but when 41 students are enrolled, there will be a sharp drop in class size, to an average of 20.5 pupils. Similarly, when 80 pupils are enrolled, the average class size will again be 40, but when 81 pupils are enrolled the average class size drops to 27. Thus, Maimonides' Rule generates discontinuities in the relationship between grade enrollment and average class size at integer multiples of 40.

The class size function derived from Maimonides' Rule can be stated formally as

[20] Goldberger (1972) discusses a similar idea in the context of compensatory education programs.

[21] Campbell's (1969) discussion of RD focused mostly on what he called a "sharp design", where the regressor of interest is a discontinuous but deterministic function of another variable. In the sharp design there is no need to instrument – the regressor of interest is entered directly. This is in contrast with what Campbell called a "fuzzy design", where the function is not deterministic. Campbell did not propose an estimator for the fuzzy design, though his student Trochim (1984) developed an IV-like procedure for that case. The discussion here covers the fuzzy design only since the sharp design can be viewed as a special case.

follows. Let b_s denote beginning-of-the-year enrollment in school s in a given grade, and let z_s denote the size assigned to classes in school s, as predicted by applying Maimonides' Rule to that grade. Assuming cohorts are divided into classes of equal size, the predicted class size for all classes in the grade is

$$z_s = b_s/(\text{int}((b_s - 1)/40) + 1).$$

This function is plotted in Fig. 2A for the population of Israeli fifth graders in 1991, along with actual fifth grade class sizes. The x-axis shows September enrollment and the y-axis shows either predicted class size or the average actual class size in all schools with that enrollment. Maimonides' Rule does not predict actual class size perfectly because other factors affect class size as well, but average class sizes clearly display a sawtooth pattern induced by the Rule.

In addition to exhibiting a strong association with average class size, Maimonides' Rule is also correlated with average test scores. This is shown in Fig. 2B, which plots average reading test scores and average values of z_s by enrollment size, in enrollment intervals of 10. The figure shows that test scores are generally higher in schools with larger enrollments and, therefore, larger predicted class sizes. Most importantly, however, average scores by enrollment size exhibit a sawtooth pattern that is, at least in part, the mirror image of the class size function. This is especially clear in Fig. 2C, which plots average scores by enrollment after running auxiliary regressions to remove a linear trend in enrollment and the effects of pupils' socioeconomic background.[22] The up and down pattern in the conditional expectation of test scores given enrollment probably reflects the causal effect of changes in class size that are induced by exogenous changes in enrollment. This interpretation is plausible because Maimonides' Rule is known to have this pattern, while it seems likely that other mechanisms linking enrollment and test scores will be smoother.

Fig. 2B makes it clear that Maimonides' Rule is not a valid instrument for class size without controlling for enrollment because predicted class size increases with enrollment and test scores increase with enrollment. The RD idea is to use the discontinuities (jumps) in predicted class size to estimate the effect of interest while controlling for smooth enrollment effects. Angrist and Lavy implement this by using z_s as an instrument while controlling for smooth effects of enrollment using parametric enrollment trends. Consider a causal model that links the score of pupil i in school s with class size and school characteristics:

$$y_{is} = X_s'\beta + n_{is}\delta + \varepsilon_{is}, \tag{27}$$

where n_{is} is the size of i's class, and X_s is a vector of school characteristics, including functions of grade enrollment, b_s. As before, we imagine that this function tells us what test

[22] The figure plots the residuals from regressions of y_{is} and z_s on b_s and the proportion of low-income pupils in the school.

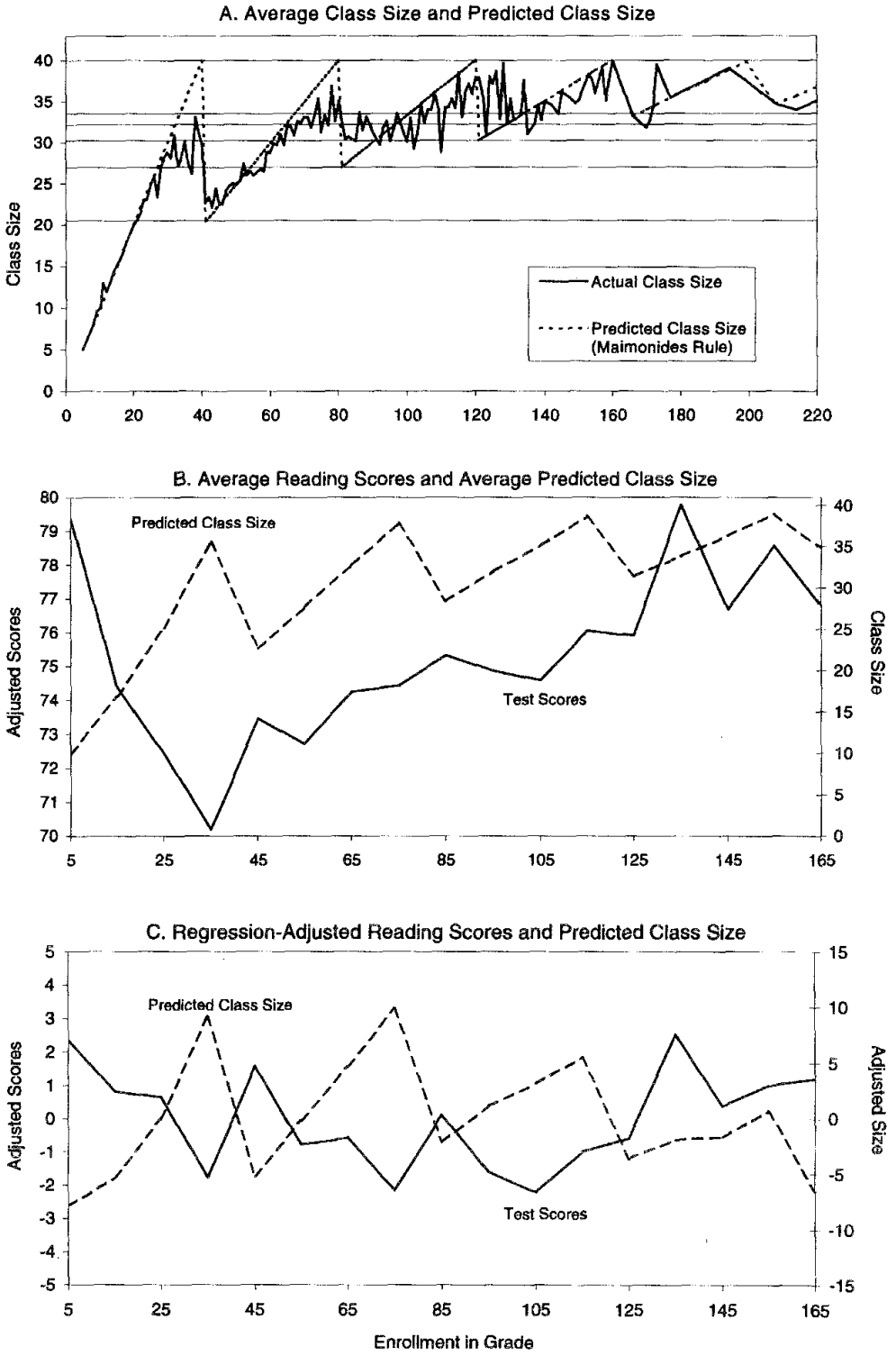

Fig. 2. Illustration of regression-discontinuity method for estimating the effect of class size on pupil's test scores. Data are from Angrist and Lavy (1998).

scores would be if class size were manipulated to be other than the observed size, n_{is}. The first-stage equation for 2SLS estimation of (27) is

$$n_{is} = X_s'\pi_0 + z_s\pi_1 + v_{is}. \tag{28}$$

A simple example is a model that includes b_s linearly to control for enrollment effects not attributable to changing class size, along with a regressor measuring the proportion of low-income students in the school.[23] The resulting 2SLS estimate of δ in standard deviation units is -0.037 (with a standard error of 0.009), meaning just over a one-third standard deviation decline in test scores for a 10 pupil increase in class size.

Since RD is an IV estimator, we do not have a separate section for pitfalls. As before, the most important issue is instrument validity and the choice of control variables. The choice of controls is even more important in RD than conventional IV, however, since the instrument is actually a function of one of the control variables. In the Angrist and Lavy application, for example, identification of δ clearly turns on the ability to distinguish z_s from X_s since z_s does not vary within schools. This suggests that RD depends more on functional form assumptions than other IV procedures, although Hahn et al. (1998) consider ways to weaken this dependence.

2.3. *Consequences of heterogeneity and non-linearity*

The discussion so far involves a highly stylized description of the world, wherein causal effects are the same for everyone, and, if the causing variable takes on more than two values, the effects are linear. Although some economic models can be used to justify these assumptions, there is no reason to believe they are true in general. On the other hand, these strong assumptions provide a useful starting place because they may provide a good approximation of reality, and because they focus attention on basic causality issues.

The cost of these simplifying assumptions is that they gloss over the fact that even when a set of estimates has a causal interpretation, they are generated by variation for a particular group of individuals over a limited range of variation in the causing variable. There is a tradition in Psychology of distinguishing between the question of *internal validity*, i.e., whether an empirical relationship has a causal interpretation in the setting where it is observed, and the question of *external validity*, i.e., whether a set of internally valid estimates has predictive value for groups or values of the response variable other than those observed in a given study.[24] Constant-coefficient and linear models make it harder to discuss the two types of validity separately, since external validity is automatic in a constant-coefficients-linear setting. For example, the constant-effects model says that the economic consequences of military service are the same for high-school dropouts and college graduates. Similarly, the linear model says the economic value of a year of

[23] In practice, Angrist and Lavy estimated (27) and (28) using class-level averages and not micro data.

[24] See, e.g., Campbell and Stanley (1963) and Meyer (1995).

schooling is the same whether the year is second grade or the last year of college. We therefore discuss the interpretation of traditional estimators when constant-effects and linearity assumptions are relaxed.

2.3.1. Regression and the conditional expectation function

Returning to the schooling example of Section 2.2.1, the causal relationship of interest is $f_i(S)$, which describes the effect of schooling on earnings. In the absence of any further assumptions, the average causal response function is $E[f_i(S)]$, with average derivative $E[f_i'(S)]$. Earlier, we assumed $f_i'(S)$ is equal to a constant, ρ, in which case averaging is not needed. In practice, however, the derivative may be heterogeneous; that is, it may vary with i or with i's characteristics, X_i. In economics, models for heterogenous treatment effects are commonly called "random coefficient" models (see, e.g., Björklund and Moffitt, 1987 or Heckman and Robb, 1985 for discussions of such models). The derivative also might be non-constant (i.e., vary with S). In either case, it makes sense to focus on the average response function or its average derivative. The principal statistical tool for doing this is the Conditional Expectation Function (CEF) of Y_i given S_i, i.e., $E[Y_i \mid S_i = S]$ or $E[Y_i \mid X_i, S_i = S]$, viewed as a function of S.

To see the connection between the CEF and the average causal response, consider first the difference in average earnings between people with S years of schooling and people with $S - 1$ years of schooling:

$$E[Y_i \mid S_i = S] - E[Y_i \mid S_i = S - 1] = E[f_i(S) - f_i(S - 1) \mid S_i = S]$$

$$+ \{E[f_i(S - 1) \mid S_i = S] - E[f_i(S - 1) \mid S_i = S - 1]\}.$$

The first term in this decomposition is the average causal effect of going from $S - 1$ to S years of schooling for those who actually have S years of education. The counterfactual average $E[f_i(S - 1) \mid S_i = S]$ is never observed, however. The second term reflects the fact that the average earnings of those with $S - 1$ years of schooling do not necessarily provide a good answer to the "what if" question for those with S years of schooling. This term is the counterpart of regression-style "omitted variables bias" for this more general model.

In this setting, the selection-on-observables assumption asserts that conditioning on a set of observed characteristics, X_i, serves to eliminate the omitted variables bias in naive comparisons. That is,

$$E[f_i(S - 1) \mid X_i, S_i = S] = E[f_i(S - 1) \mid X_i, S_i = S - 1] \qquad \text{for all } S, \tag{29}$$

so that conditional on X, the CEF and average causal response function are the same:

$$E[Y_i \mid X_i, S_i = S] = E[f_i(S) \mid X_i].$$

In this case, the conditional-on-X comparison does estimate the causal effect of schooling:

$$E[Y_i \mid X_i, S_i = S] - E[Y_i \mid X_i, S_i = S - 1] = E[f_i(S) - f_i(S - 1) \mid X_i].$$

This is analogous to the notion that adding X_i to a regression eliminates omitted variables bias in OLS estimates of the returns to schooling.

The preceding discussion provides sufficient conditions for the CEF to have a causal interpretation. We next consider the relationship between regression parameters and the CEF. One interpretation of regression is that the population OLS slope vector provides a minimum mean squared error (MMSE) linear approximation to the CEF. This feature of regression is discussed in Goldberger's (1991) econometrics text (see especially Section 5.5).[25] A related property is the fact that regression coefficients have an "average derivative" interpretation. In multivariate regression models, however, this interpretation is complicated by the fact that the OLS slope vector is actually matrix-weighted average of the gradient of the CEF. Matrix-weighted averages are difficult to interpret except in special cases (see Chamberlain and Leamer, 1976).[26]

One interesting special case where the OLS slope vector can be readily interpreted is when S_i is the single regressor of interest and the CEF of this regressor given all other regressors is linear, so that

$$\mathrm{E}[S_i \mid X_i] = X_i'\pi, \tag{30}$$

where π is a conformable vector of coefficients. This assumption is satisfied in the schooling regression, for example, in a model where all X-variables are discrete and the parameterization allows a separate effect for each possible value of X_i. This is not unrealistic in applications with large datasets; see, for example, Angrist and Krueger (1991) and Angrist (1998). In this case, the population regression coefficient from a regression of Y_i on X_i and S_i can be written

$$\begin{aligned}\rho_{\mathrm{r}} &= \mathrm{E}[(S_i - \mathrm{E}[S_i \mid X_i])Y_i]/\mathrm{E}[(S_i - \mathrm{E}[S_i \mid X_i])S_i] \\ &= \mathrm{E}[(S_i - \mathrm{E}[S_i \mid X_i])\mathrm{E}[Y \mid X_i, S_i]]/\mathrm{E}[(S_i - \mathrm{E}[S_i \mid X_i)S_i],\end{aligned} \tag{31}$$

which is derived by iterating expectations over X_i and S_i.

Maintaining assumption (30), i.e., that $\mathrm{E}[S_i \mid X_i]$ is linear, first consider the case where $\mathrm{E}[Y_i \mid X_i, S_i]$ is linear in S_i but not X_i. Then we can write

$$\rho_X \equiv \mathrm{E}[Y_i \mid X_i, S_i = S] - \mathrm{E}[Y_i \mid X_i, S_i = S - 1],$$

for all S, which means

[25] Proof that OLS gives a MMSE linear approximation to the CEF: The vector of population regression coefficients for regressor vector W_i solves $\min_b \mathrm{E}(Y_i - W_i'b)^2$. But $(Y_i - W_i'b)^2 = [(Y_i - \mathrm{E}[Y_i \mid W_i]) + (\mathrm{E}[Y_i \mid W_i] - W_i'b)]^2$ and $\mathrm{E}[(Y_i - \mathrm{E}[Y_i \mid W_i])\,(\mathrm{E}[Y_i \mid W_i] - W_i'b)] = 0$, so $\min_b \mathrm{E}([Y_i \mid W_i] - W_i'b)]^2$ has the same solution.

[26] The population slope vector is $\mathrm{E}[W_iW_i']^{-1}\mathrm{E}[W_iY_i] = \mathrm{E}[W_iW_i']^{-1}\mathrm{E}[W_i\mathrm{E}(Y_i \mid W_i)]$. Assume $\mathrm{E}(W_i) = 0$ so these are the non-intercept coefficients. Linearizing the CEF, we have $\mathrm{E}(Y_i \mid W_i) = \mathrm{E}(Y_i \mid W_i = 0) + W_i'\nabla\mathrm{E}(Y_i \mid \tilde{w}_i)$, where $\nabla\mathrm{E}(Y_i \mid \tilde{w}_i)$ is the gradient of the conditional expectation function, and $\tilde{w}_i$ is a random vector that lies between W_i and zero. So the slope vector is $\mathrm{E}[W_iW_i']^{-1}\mathrm{E}[(W_iW_i')\nabla\mathrm{E}(Y_i \mid \tilde{w}_i)]$, which is a matrix-weighted average of the gradient with weights (W_iW_i').

$$E[Y_i \mid X_i, S_i] = E[Y_i \mid X_i, S_i = 0] + S_i\rho_X. \tag{32}$$

In other words, the CEF is linear in schooling, but the schooling coefficient is not constant and depends on X_i.

Substituting (32) into (31), we have

$$\rho_r = E[(S_i - E[S_i \mid X_i])^2\rho_X]/E[(S_i - E[S_i \mid X_i])^2] = E[\sigma_S^2(X_i)\rho_X]/E[\sigma_S^2(X_i)], \tag{33}$$

where $\sigma_S^2(X_i) \equiv E[S_i - E[S_i|X_i])^2|X_i]$ is the variance of S_i given X_i. So in this case, regression provides a variance-weighted average of the slope at each X_i. Values of X_i that get the most weight are those where the conditional variance of schooling is largest.

What if the CEF of Y_i varies with both X_i and S_i? Let

$$\rho_{SX} \equiv E[Y_i \mid X_i, S_i = S] - E[Y_i \mid X_i, S_i = S - 1],$$

where the ρ_{SX} notation reflects variation with both S and X_i. Then the coefficient on S_i in a regression of Y_i on X_i and S_i can be written

$$\rho_r = E\left[\sum_{S=1}^{\tilde{S}} \rho_{SX}\mu_{SX}\right]E\left[\sum_{S=1}^{\tilde{S}} \mu_{SX}\right]^{-1}, \tag{34}$$

where

$$\mu_{SX} = (E[S_i \mid X_i, S_i \geq S] - E[S_i \mid X_i, S_i < S])P[S_i \geq S \mid X_i](1 - P[S_i \geq S \mid X_i]) \geq 0.$$

and S takes on values in the set $\{0, 1, \ldots, \tilde{s}\}$. This result, which is proved in Appendix A, is a generalization of the formula for bivariate regression coefficients given by Yitzhaki (1996).[27]

The weighting formula in (34) has a sum and an expectation. The sum averages ρ_{SX} for all schooling increments, given a particular value of X_i (this averaging matters if the CEF is non-linear). The expectation then averages this sum in the distribution of X_i (this averaging matters if the response function is heterogeneous). The formula for the weights, μ_{SX}, can be used to characterize the OLS slope vector. First, for any particular X_i, weight is given to ρ_{SX} for each S in proportion to the change in the conditional mean of S_i, as S_i falls above or below S. More weight is also given to points in the domain of $f_i(S)$ that are close to the conditional median of S_i given X_i since this is where $P[S_i \geq S \mid X_i](1 - P[S_i \geq S \mid X_i])$ is maximized. Second, as in the linear case discussed above, weight is also given in proportion to conditional variance of S_i given X_i, except now this variance is defined separately for each S using dummies for the event that $S_i \geq S$. Note also that the OLS estimate contains no information about the returns to schooling for values of X_i where

[27] Yitzhaki gives examples and describes the OLS weighting function for a model with a single continuously distributed regressor in detail. For Normally distributed regressors, the weighting function is the Normal density function, so that OLS provides a density-weighted average of the sort discussed by Powell et al. (1989). For an alternative non-parametric interpretation of OLS coefficients see Stoker (1986).

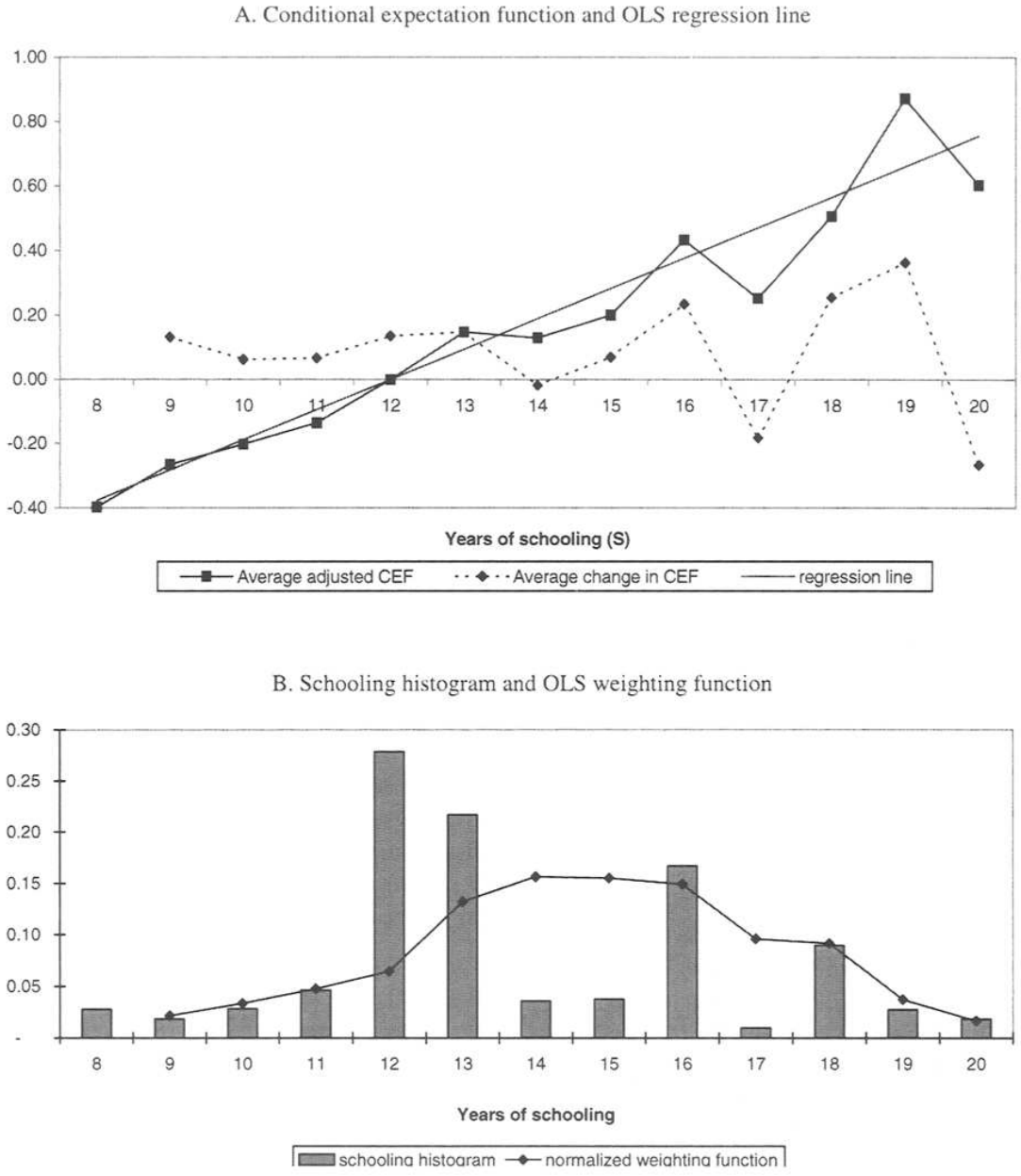

Fig. 3. (A) The conditional expectation function (CEF) of log weekly earnings given schooling, adjusted for covariates as described in the text. Also plotted is the average change in the CEF and the OLS regression line. (B) The schooling histogram and OLS weighting function. Data are for men aged 40–49 in the 1990 Census.

$P[S \geq S \mid X_i]$ equals 0 or 1. This includes values of X_i where S_i does not vary across observations, because $P[S \geq S \mid X_i] = 1$ if $P[S_i = S \mid X_i] = 1$.

The weighting function is illustrated in Fig. 3 using data from the 1990 Census. The top panel plots an estimate of the earnings-schooling CEF, i.e., average log weekly wages against years of schooling for men with 8–20 years of schooling, adjusted for covariates. In other words, the plot shows $E\{E[Y_i \mid X_i, S_i = S]\}$, plotted against S. Years of schooling

are not recorded in the 1990 Census and were therefore imputed from categorical schooling variables as described in the appendix. The X-variables are race (white, non-white), age (40–49), and state of birth. The covariates in this case are similar to those used in some of the specifications in the Angrist and Krueger (1991) study of the returns to schooling, although the data underlying this figure are more recent.

The dotted line in the figure plots the change in $\mathrm{E}\{\mathrm{E}[Y_i \mid X_i, S_i = S]\}$ with S. This is the covariate-adjusted difference in average log weekly wages at each schooling increment,

$$\rho_S \equiv \mathrm{E}\{\mathrm{E}[Y_i \mid X_i, S_i = S] - \mathrm{E}[Y_i \mid X_i, S_i = S - 1]\} = \sum_X \rho_{SX} \mathrm{P}(X_i = X).$$

For example, the first point on the dotted line is an estimate of ρ_9–ρ_8, which is the average difference in earnings between those with 9 years of schooling and those with 8 years of schooling, adjusting for differences in the distribution of X_i between the two schooling groups.[28] The returns measured in this way are remarkably stable until 13 years of schooling, but quite variable after that and sometimes even negative.

The straight line in the figure is the OLS regression line obtained from fitting Eq. (1) with a saturated model for X_i (in other words, the model includes a full set of dummies d_{iX}, which equal one when $X_i = X$ for every value X; the OLS estimate of ρ in this case is 0.094). This parameterization satisfies assumption (30), i.e., $\mathrm{E}[S_i \mid X_i]$ is linear. The figure illustrates the sense in which OLS captures the average return. The OLS weighting function for each value of S_i is plotted in the lower panel, along with the histogram of schooling.[29] Like the distribution of schooling itself, the OLS weighting scheme puts the most weight on values between 12 and 16. It is interesting to note, however, that while the histogram of schooling is bimodal, the weighting function is smoother and unimodal. Moreover, the population average of ρ_S, i.e., the weighted average of the covariate-adjusted return using the schooling histogram, $\sum_S \rho_S \mathrm{P}(S_i = S)$, is 0.144, which is considerably larger than the OLS estimate. This is because about half of the sample has 12–13 years of schooling, where the returns are 0.136 and 0.148. The OLS weighting function gives more weight than the histogram to other schooling values, like 14, 15, and 17, where the returns are small and even negative.

2.3.2. Matching instead of regression

The previous section shows how regression produces a weighted average of covariate-specific effects for each value of the causing variable. The empirical consequences of the OLS weighting scheme in any particular application depend on the distribution of regressors and the amount of heterogeneity in the causal effect of interest. Matching methods provide an alternative estimation strategy that affords more control over the weighting scheme used to produce average causal effects. Matching methods also have the advantage

[28] The unadjusted difference in average wages is $\{\mathrm{E}[Y_i \mid S_i = S] - \mathrm{E}[Y_i \mid S_i = S - 1]\}$, which equals $\mathrm{E}[\mathrm{E}(Y_i \mid X_i, S_i = S) \mid S_i = S] - \mathrm{E}[\mathrm{E}(Y_i \mid X_i, S_i = \mathrm{S} - 1) \mid S_i = S - 1]$.

[29] Since the regression model has covariates, the weights vary with X_i as well as for each schooling increment. The average weighting function plotted in the figure is $\sum_X \mu_{SX} \mathrm{P}(X_i = X)$.

of making the comparisons that are used for statistical identification transparent. Matching is most practical in cases where the causing variable takes on two values, as in the union status and military service examples discussed previously.

Again, we use the example of estimating the effect of military service to illustrate this technique. Angrist (1998) reported matching and regression estimates of the effects of voluntary military service on civilian earnings. As in the Vietnam study, the potential outcomes are Y_{i0}, denoting what someone would earn if they did not serve in the military, and Y_{1i} denoting earnings as a veteran. Since $Y_{1i} - Y_{0i}$ is not constant, and we never observe both potential outcomes for any one person, it makes sense to focus on average effects. One possibility is the "average treatment effect," $E[Y_{1i} - Y_{0i}]$, but this is not usually the first choice in studies of this kind since people who serve in the military tend to have personal characteristics that differ, on average, from those of people who did not serve. The manpower policy innovations that are typically contemplated affect those individuals who either now serve or who might be expected to serve in the future. For example, between 1989 and 1992, the size of the military declined sharply because of increasing enlistment standards. Policy makers would like to know whether the people who would have served under the old rules but are unable to enlist under the new rules were hurt by the lost opportunity for service. This sort of reasoning leads researchers to try to estimate the "effect of treatment on the treated," which is $E[Y_{1i} - Y_{0i} \mid D_i = 1]$ in our notation.[30]

As in the study of Vietnam veterans, simply comparing the earnings of veterans and non-veterans is unlikely to provide a good estimate of the effect of military service on veterans. The comparison by veteran status is

$$E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]$$

$$= E[Y_{1i} - Y_{0i} \mid D_i = 1] + \{E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]\}.$$

This is the average causal effect of military service on veterans, $E[Y_1 - Y_0 \mid D = 1]$, plus a bias term attributable to the fact that the earnings of non-veterans are not necessarily representative of what veterans would have earned had they not served in the military. For example, veterans may have higher earnings simply because they must have higher test scores and be high school graduates to meet military screening rules.

The bias term in naive comparisons goes away if D_i is randomly assigned because then D_i will then be independent of Y_{0i} and Y_{1i}. Since voluntary military service is not randomly assigned (and there is no longer a draft lottery), Angrist (1998) used matching and regression techniques to control for observed differences between veterans and non-veterans who applied to get into the all-volunteer forces between 1979 and 1982. The motivation for a control strategy in this case is the fact that the military screens applicants to the armed forces primarily on the basis of age, schooling, and test scores, characteristics that are

[30] Heckman and Robb (1985) discuss the rationale for estimating effects on the treated when evaluating subsidized training programs.

observed in the Angrist (1998) data. Identification in this case is based on the claim that after conditioning on all of the observed characteristics that are known to affect veteran status, veterans and non-veterans are comparable in the sense that

$$\mathrm{E}[Y_{0i} \mid X_i, D_i = 1] = \mathrm{E}[Y_{0i} \mid X_i, D_i = 0]. \tag{35}$$

This assumption seems plausible for two reasons. First, the non-veterans who provide observations on Y_{0i} did in fact apply to get into the military. Second, selection for military service from the pool of applicants is based almost entirely on variables that are observed and included in the X-variables. Variation in veteran status conditional on X_i comes solely from the fact that some qualified applicants nevertheless fail to enlist at the last minute. Of course, the considerations that lead a qualified applicant to "drop out" of the enlistment process could be related to earnings potential, so assumption (35) is clearly not guaranteed.

Given assumption (35), the effect of treatment on the treated can be constructed as follows:

$$\begin{aligned}\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1] &= \mathrm{E}\{\mathrm{E}[Y_{1i} \mid X_i, D_i = 1] - \mathrm{E}[Y_{0i} \mid X_i, D_i = 1] \mid D_i = 1\} \\ &= \mathrm{E}\{\mathrm{E}[Y_{1i} \mid X_i, D_i = 1] - \mathrm{E}[Y_{0i} \mid X_i, D_i = 0] \mid D_i = 1\} = \mathrm{E}[\delta_X \mid D_i = 1],\end{aligned} \tag{36}$$

where

$$\delta_X \equiv \mathrm{E}[Y_i \mid X_i, D_i = 1] - \mathrm{E}[Y_i \mid X_i, D_i = 0].$$

Here δ_X is a random variable that represents the set of differences in mean earnings by veteran status corresponding to each value taken on by X_i. This is analogous to the random coefficient ρ_X that was defined for the schooling problem. Note, however, that since D_i is binary, the response function in this case is automatically linear in D_i.

The matching estimator in Angrist (1998) uses the fact that X_i is discrete to construct (36), which can also be written

$$\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1] = \sum_X \delta_X \mathrm{P}(X_i = X \mid D_i = 1), \tag{37}$$

where $\mathrm{P}(X_i = X \mid D = 1)$ is the probability mass function for X_i given $D_i = 1$ and the summation is over the values of X_i.[31] In this case, X_i, takes on values determined by all possible combinations of year of birth, AFQT test-score group,[32] year of application to the military, and educational attainment at the time of application.

Naive comparisons clearly overestimate the benefit of military service. This can be seen in Table 6, which reports differences-in-means, matching, and regression estimates of the effect of voluntary military service on the 1988–1991 Social Security-taxable earnings of men who applied to join the military between 1979 and 1982. The matching estimates were constructed from the sample analog of (37), i.e., from covariate-value-specific differ-

[31] This matching estimator is discussed by Rubin (1977) and used by Card and Sullivan (1988) to estimate the effect of subsidized training on employment.

[32] This is the Armed Forces Qualification Test, used by the military to screen applicants.

Table 6
Matching and regression estimates of the effects of voluntary military service[a]

Race	Average earnings in 1988–1991 (1)	Differences in means by veteran status (2)	Matching estimates (3)	Regression estimates (4)	Regression minus matching (5)
Whites	14537	1233.4	−197.2	−88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
Non-whites	11664	2449.1	839.7	1074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

[a] Notes: Adapted from Angrist (1998, Tables II and V). Standard errors are reported in parentheses. The tables shows estimates of the effect of voluntary military service on the 1988–1991 Social Security-taxable earnings of men who applied to enter the armed forces between 1979 and 1982. The matching and regression estimates control for applicants' year of birth, education at the time of application, and AFQT score. There are 128,968 whites and 175,262 non-whites in the sample.

ences in earnings, δ_X, weighted to form a single estimate using the distribution of covariates among veterans. Although white veterans earn \$1233 more than non-veterans, this difference becomes negative once the adjustment for differences in covariates is made. Similarly, while non-white veterans earn \$2449 more than non-veterans, controlling for covariates reduces this to \$840.

Table 6 also reports regression estimates of the effect of voluntary service, controlling for exactly the same covariates used in the matching estimates. These are estimates of δ_r in the equation

$$Y_i = \sum_X d_{iX}\beta_X + \delta_r D_i + e_i, \tag{38}$$

where β_X is a regression-effect for $X_i = X$ and δ_r is the regression treatment effect. This corresponds to a saturated model for X_i. Despite the fact that the matching and regression estimates control for the same variables, the regression estimates are significantly larger than the matching estimates for both whites and non-whites.[33] The reason the regression estimates are larger than the matching estimates is that the two estimation strategies use different weighting schemes. While the matching estimator combines covariate-value-specific estimates, δ_X, to produces an estimate of the effect of treatment on the treated, regression produces a variance-weighted average of these effects. To see this, note that since D_i is binary and $\mathrm{E}[D_i \mid X_i]$ is linear, formula (33) from the previous section implies

$$\delta_r = \mathrm{E}[(D_i - \mathrm{E}[D_i \mid X_i])^2 \delta_X]/\mathrm{E}[(D_i - \mathrm{E}[D_i \mid X_i])^2] = \mathrm{E}[\sigma_D^2(X_i)\delta_X]/\mathrm{E}[\sigma_D^2(X_i)],$$

But in this case, $\sigma_D{}^2(X_i) = \mathrm{P}(D_i = 1 \mid X_i)(1 - \mathrm{P}(D_i = 1 \mid X_i))$, so

[33] The formula for the covariance of regression and matching estimates is derived in Angrist (1998, p. 274).

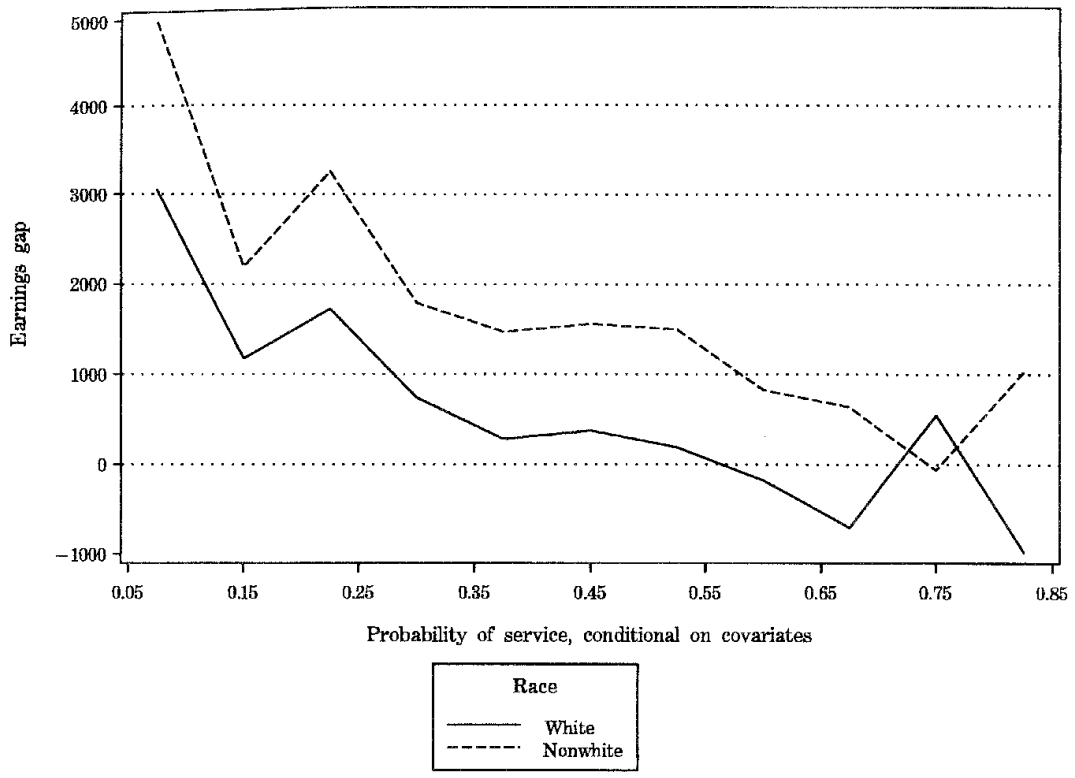

Fig. 4. Effects of voluntary military service on earnings in 1988–1991, plotted by race and probability of service, conditional on covariates. The earnings data are from Social Security administrative records.

$$\delta_r = \frac{\sum_X \delta_X[\mathrm{P}(D_i = 1 \mid X_i = X)(1 - \mathrm{P}(D_i = 1 \mid X_i = X))]\mathrm{P}(X_i = X)}{\sum_X [\mathrm{P}(D_i = 1 \mid X_i = X)(1 - \mathrm{P}(D_i = 1 \mid X_i = X))]\mathrm{P}(X_i = X)}.$$

In other words, regression weights each covariate-specific treatment effect by $\mathrm{P}(X_i = X \mid D_i = 1)(1 - \mathrm{P}(X_i = X \mid D_i = 1))$. In contrast, the matching estimator, (37), can be written

$$\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1] = \frac{\sum_X \delta_X \mathrm{P}(D_i = 1 \mid X_i = X)\mathrm{P}(X_i = X)}{\sum_X \mathrm{P}(D_i = 1 \mid X_i = X)\mathrm{P}((X_i = X)}.$$

because $\mathrm{P}(X_i = X \mid D_i = 1) = \mathrm{P}(D_i = 1 \mid X_i = X)\mathrm{P}(X_i = X)/\mathrm{P}(D_i)$.

The weights underlying $\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$ are proportional to the probability of veteran status at each value of the covariates. So the men most likely to serve get the most weight in estimates of the effect of treatment on the treated. In contrast, regression estimation weights each of the underlying treatment effects by the conditional variance of treatment status, which in this case is maximized when $\mathrm{P}(D_i = 1 \mid X_i = X) = 1/2$. Of course, the difference in weighting schemes is of no importance if the effect of interest

does not vary with X_i. But Fig. 4, which plots X-specific estimates (δ_X) of the effect of veteran status on average 1988–1991 earnings against $P[D_i = 1 \mid X_i = X]$, shows that the men who were most likely to serve in the military benefit least from their service. This fact leads matching estimates of the effect of military service to be smaller than regression estimates based on the same vector of controls.

2.3.3. Matching using the propensity score

It is easy to construct a matching estimator based on (37) when, as in Angrist (1998), the conditioning variables are discrete and the sample has many observations at almost every value taken on by the vector of explanatory variables. What about situations where X_i is continuous, so that exact matching is not practical? Problems involving more finely distributed X-variables are often solved by aggregating values to make coarser groupings or by pairing observations that have similar, though not necessarily identical, values. See Cochran (1965), Rubin (1973), or Rosenbaum (1995, Chapter 3) for discussions of this approach. More recently, Deaton and Paxson (1998) used non-parametric methods to accommodate continuous-valued control variables in a matching estimator.

The problem of how to aggregate the X-variables also motivates a matching method first developed in a series of papers by Rosenbaum and Rubin (1983, 1984, 1985). These papers show that full control for covariates can be obtained by controlling solely for a function of X_i called the propensity score, which is simply the conditional probability of treatment, $p(X_i) \equiv P(D_i = 1 \mid X_i)$. The formal result underlying this approach says that if conditioning on X_i eliminates selection bias,

$$E[Y_{0i} \mid X_i, D_i = 1] = E[Y_{0i} \mid X_i, D_i = 0],$$

then it must also be true that conditioning on $p(X_i)$ eliminates selection bias:

$$E[Y_{0i} \mid p(X_i,), D_i = 1] = E[Y_{0i} \mid e(X_i), D_i = 0].$$

This leads to the following modification of (36):

$$E[Y_{1i} - Y_{0i} \mid D_i = 1] = E\{E[Y_{1i} \mid X_i, D_i = 1] - E[Y_{0i} \mid X_i, D_i = 1] \mid D_i = 1\}$$

$$= E\{E[Y_{1i} \mid p(X_i), D_i = 1] - E[Y_{0i} \mid p(X_i), D_i = 0] \mid D_i = 1\}.$$

Of course, to make this expression into an estimator, the propensity score $p(X_i)$ must first be estimated. The practical value of this result is that in some cases, it may be easier to estimate $p(X_i)$ and then condition on the estimates of $p(X_i)$ than to condition on X_i directly. For example, even if X_i is continuous, $p(X_i)$ may have some "flat spots", or we may have some prior information about $p(X_i)$. The propensity score approach is also conceptually appealing because it focuses attention on variables that are related to the regressor of interest. Although Y_i may vary with X_i in complicated ways, this is only of concern for values of X_i where $p(X_i)$ varies as well.

An example using the propensity score in labor economics is Dehejia and Wahba's (1995) reanalysis of the National Supported Work (NSW) training program studied by

Lalonde (1986). The NSW provided training to different groups of "hard-to-employ" men and women in a randomized demonstration project. Lalonde's study uses observational control groups from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID) to look at whether econometric methods are likely to generate conclusions similar to those found in the experimental study. One hurdle facing the non-experimental investigator attempting to construct a control group for trainees is how to control for lagged earnings. As we noted earlier, controlling for lagged earnings is important since participants in government training programs are often observed to experience a decline in earnings before entering the program (see, e.g., Ashenfelter and Card, 1985, and the chapter on training by Heckman, Lalonde, and Smith in this volume).

Lalonde (1986) found that non-experimental methods based on regression models, including models with fixed effects and control for lagged earnings, fail to replicate the NSW experimental findings. Using the same observational control groups as Lalonde (1986), Dehejia and Wahba (1995) control for lagged earnings and other covariates by first estimating a logit model that relates participation in the program to the covariates and two lags of earnings. Following an example by Rosenbaum and Rubin (1984), they then divide the sample into quintiles on the basis of fitted values from this logit, i.e., based on estimates of the propensity score. The overall estimate of the effect of treatment on the treated is the difference between average trainee and average control earnings in each quintile, weighted by the number of trainees in the quintile and summed across quintiles. The estimates produced using this method are similar to those based on the experimental random assignment (and apparently more reliable than regression estimates). It should be clear, however, that use of propensity score methods requires a number of decisions about how to model and control for the score. There is little in the way of formal statistical theory to guide this process, and the question of whether propensity score methods are better than other methods remains open. See Heckman et al. (1997) for further empirical evidence, and Hahn (1998) for recent theoretical results on efficiency considerations in these models.

2.3.4. Interpreting instrumental variables estimates

The discussion of IV in Section 2.2.3 used the example of veteran status, with two potential outcomes and a constant causal effect, $Y_{1i} - Y_{0i} = \delta$. What is the interpretation of an IV estimate when the constant-effects assumption is relaxed? We begin with a model where the causing variable is binary, as in the veteran status example, turning afterwards to a more general model. As before, the discussion is initially limited to the Wald estimator since this is an important and easily-analyzed IV estimator.

Without the constant-effects assumption, we can write the observed outcome, Y_i, in terms of potential outcomes as

$$Y_i = Y_{i0} + (Y_{1i} - Y_{0i})D_i = \beta_0 + \delta_i D_i + \eta_i, \tag{39}$$

where $\beta_0 \equiv \mathrm{E}[Y_{i0}]$ and $\delta_i \equiv Y_{1i} - Y_{0i}$ is the heterogeneous causal effect. The expression after the second equals sign is a "random-coefficients" version of the causal model in Section 2.3.3 (see Eq. (23)). To facilitate the discussion of IV, we also introduce some

notation for the first-stage relationship between the causing variable, D_i, and the binary instrument, Z_i. To allow for as much heterogeneity as possible, the first stage equation is written in a manner similar to (39):

$$D_i = D_{i0} + (D_{1i} - D_{0i})Z_i = \pi_0 + \pi_{1i}Z_i + \nu_i, \tag{40}$$

where $\pi_0 \equiv \mathrm{E}[D_{i0}]$ and $\pi_{1i} \equiv (D_{1i} - D_{0i})$ is the causal effect of the *instrument* on D_i. In the draft lottery example, D_{0i} tells us whether i would serve in the military if not draft-eligible and D_{1i} tells us whether i would serve when draft-eligible. The effect of draft-eligibility on D_i is the difference between these two potential treatment assignments.

The principle identifying assumption in this setup is that the vector of potential outcomes and potential treatment assignments is jointly independent of the instrument. Formally,

$$\{Y_{1i}, Y_{0i}, D_{1i}, D_{0i}\} \amalg Z_i,$$

where $\amalg$ is notation for statistical independence (see, e.g., Dawid, 1979, or Rosenbaum and Rubin, 1983).[34] In the lottery example, Z_i is clearly independent of $\{D_{0i}, D_{1i}\}$ since Z_i was randomly assigned. As noted in Section 2.2.3, however, independence of $\{Y_{0i}, Y_{1i}\}$ and Z_i is not guaranteed by randomization since Y_{0i} and Y_{1i} refer to potential outcomes under alternative assignments of *veteran status* and not Z_i itself. Even though Z_i was randomly assigned, so the relationship between Z_i and Y_i is causal, in principle there might be reasons other than veteran status for an effect of draft-eligibility on earnings. The independence assumption, which is similar to the assumption that Z_i and η_i are uncorrelated in the constant-effects model, rules this possibility out.

A second assumption that is useful here, and one that does not arise in a constant-effects setting, is that either $\pi_{1i} \geq 0$ for all i or $\pi_{1i} \leq 0$ for all i. This *monotonicity* assumption, introduced by Imbens and Angrist (1994), means that while the instrument may have no effect on some people, it must be the case that the instrument acts in only one direction, either $D_{1i} \geq D_{0i}$ or $D_{1i} \leq D_{0i}$ for all i. In what follows, we assume $D_{1i} \geq D_{0i}$ for all i. In the draft-lottery example, this means that although draft-eligibility may have had no effect on the probability of military service for some men, there is no one who was actually kept out of the military by being draft-eligible. Without monotonicity, instrumental variables estimators are not guaranteed to estimate a weighted average of the underlying causal effects, $Y_{1i} - Y_{0i}$.

Given independence and monotonicity, the Wald estimator in this example can be interpreted as the effect of veteran status on those whose treatment status was changed by the instrument. This parameter is called the local average treatment effect (LATE; Imbens and Angrist, 1994), and can be written as follows:

$$\frac{\mathrm{E}[Y_i \mid Z_i = 1] - \mathrm{E}[Y_i \mid Z_i = 0]}{\mathrm{E}[D_i \mid Z_i = 1] - \mathrm{E}[D_i \mid Z_i = 0]} = \mathrm{E}[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}] = \mathrm{E}[\delta_i \mid \pi_{1i} > 0].$$

[34] The independence assumption using random-coefficients notation is $\{\delta_i, \eta_i, \pi_{1i}, \nu_i\} \amalg Z_i$.

Thus, IV estimates of effects of military service using the draft lottery estimate the effect of military service on men who served because they were draft-eligible, but would not otherwise have served.[35] This obviously excludes volunteers and men who were exempted from military service for medical reasons, but it includes men for whom the draft policy was binding. Much of the debate over compulsory military service focused on draftees, so LATE is clearly a parameter of policy interest in the Vietnam context.

The LATE parameter can be linked to the parameters in traditional econometric models for causal effects. One commonly used specification for dummy endogenous regressors like veteran status is a latent-index model (see, e.g., Heckman, 1978), where

$$D_i = 1 \quad \text{if } \gamma_0 + \gamma_1 Z_i > \nu_i \quad \text{and 0 otherwise,}$$

and ν_i is a random factor assumed to be independent of Z_i. This specification can be motivated by comparisons of utilities and costs under alternative choices. In the notation of Eq. (40), the latent-index model characterizes potential treatment assignments as

$$D_{0i} = 1 \text{ if } [\gamma_0 > \nu_i] \quad \text{and} \quad D_{1i} = 1 \text{ if } [\gamma_0 + \gamma_1 > \nu_i].$$

Note that in this model, monotonicity is automatically satisfied since γ_1 is a constant. Assuming $\gamma_1 > 0$,

$$\mathrm{E}[Y_{1i} - Y_{0i} \mid D_1 > D_{0i}] = \mathrm{E}[Y_{1i} - Y_{0i} \mid \gamma_0 + \gamma_1 > \nu_i > \gamma_0],$$

which is a function of the structural first-stage parameters, γ_0 and γ_1. The LATE parameter is representative of a larger group the larger is the first-stage parameter, γ_1.

LATE can also be compared with the effect of treatment on the treated for this problem, which depends on the same first-stage parameters and the marginal distribution of Z_i. Note that in the latent-index specification, $D_i = 1$ in one of two ways: either $\gamma_0 > \nu_i$, in which case the instrument does not matter, or $\gamma_0 + \gamma_1 > \nu_i > \gamma_0$ and $Z_i = 1$. Since these two possibilities partition the group with $D_i = 1$, we can write

$$\begin{aligned}
\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1] &= \mathrm{P}(D_i = 1)^{-1} \\
&\times \{\mathrm{E}[Y_{1i} - Y_{0i} \mid \gamma_0 + \gamma_1 > \nu_i > \gamma_0, Z_i = 1]\mathrm{P}(\gamma_0 + \gamma_1 > \nu_i > \gamma_0, Z_i = 1) \\
&+ \mathrm{E}[Y_{1i} - Y_{0i} \mid \gamma_0 > \nu_i]\mathrm{P}(\gamma_0 > \nu_i)\} \\
&= \mathrm{P}(D_i = 1)^{-1} \times \{\mathrm{E}[Y_{1i} - Y_{0i} \mid \gamma_0 + \gamma_1 > \nu_i > \gamma_0]\mathrm{P}(\gamma_0 + \gamma_1 > \nu_i > \gamma_0)\mathrm{P}(Z_i = 1) \\
&+ \mathrm{E}[Y_{1i} - Y_{0i} \mid \gamma_0 > \nu_i]\mathrm{P}(\gamma_0 > \nu_i)\}.
\end{aligned}$$

[35] Proof of the LATE result: $\mathrm{E}[Y_i \mid Z_i = 1] = \mathrm{E}[Y_{i0} + (Y_{1i} - Y_{0i})D_i \mid Z_i = 1]$, which equals $\mathrm{E}[Y_{i0} + (Y_{1i} - Y_{0i})D_{1i}]$ by independence. Likewise $\mathrm{E}[Y_i \mid Z_i = 0] = \mathrm{E}[Y_{i0} + (Y_{1i} - Y_{0i})D_{0i}]$, so the numerator of the Wald estimator is $\mathrm{E}[(Y_{1i} - Y_{0i})(D_{1i} - D_{0i})]$. Monotonicity means $D_{1i} - D_{0i}$ equals one or zero, so $\mathrm{E}[(Y_{1i} - Y_{0i})(D_{1i} - D_{0i})] = \mathrm{E}[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]\mathrm{P}[D_{1i} > D_{0i}]$. A similar argument shows $\mathrm{E}[D_i \mid Z_i = 1] - \mathrm{E}[D_i \mid Z_i = 0] = \mathrm{E}[D_{1i} - D_{0i}] = \mathrm{P}[D_{1i} > D_{0i}]$.

This shows that the effect on the treated is a weighted average of LATE and the effect on men whose treatment status is unaffected by the instrument.[36] Note, however, that although LATE equals the Wald estimator, the effect on the treated is not identified in this case without additional assumptions (see, e.g., Angrist and Imbens, 1991).

Interpreting IV estimates with cardinal variables. So far the discussion of IV has focused on models with a binary regressor. What does the Wald estimator estimate when the regressor takes on more than two values, like schooling? As in the discussion of regression in Section 2.2.1, suppose the causal relationship of interest is characterized by a function that describes exactly what a given individual would earn if they obtained different levels of education. This relationship is person-specific, so we write $f_i(S)$ to denote the earnings or wage that i would receive after obtaining S years of education. The observed earnings level is $Y_i = f_i(S_i)$.

Again, it is useful to have a general notation for the first-stage relationship between S_i and Z_i:

$$S_i = S_{0i} + (S_{1i} - S_{0i})Z_i = \kappa_0 + \kappa_{1i}Z_i + \nu_i, \tag{41}$$

where S_{0i} is the schooling i would get if $Z_i = 0$, S_{1i} is the schooling i would get if $Z_i = 1$, and $\kappa_0 \equiv \mathrm{E}[S_{0i}]$. In random-coefficients notation, the causal effect of Z_i on S_i is $\kappa_{1i} \equiv S_{1i} - S_{0i}$. To make this concrete, suppose the instrument is a dummy for being born in the second, third, or fourth quarter of the year, as for the Wald estimate in Angrist and Krueger (1991, Table 3). Since compulsory attendance laws allow people to drop out of school on their birthday (typically the 16th) and most children enter school in September of the year they turn 6, pupils born later in the year are kept in school longer than those born earlier. In this example, S_{0i} is the schooling i would get if born in the first quarter and S_{1i} is the schooling i would get if born in a later quarter.

Now the independence assumption is $\{f_i(S), S_{1i}, S_{0i}\} \perp\!\!\!\perp Z$, and the monotonicity assumption is $S_{1i} \geq S_{0i}$. This means the instrument is independent of what an individual *could* earn with schooling level S, and independent of the random elements in the first stage.[37] Using the independence assumption and Eq. (41) to substitute for S_i, the Wald estimator can be written

$$\frac{\mathrm{E}[f_i(S_i) \mid Z_i = 1] - \mathrm{E}[f_i(S_i) \mid Z_i = 0]}{\mathrm{E}[S_i \mid Z_i = 1] - \mathrm{E}[S_i \mid Z_i = 0]} = \frac{\mathrm{E}[f_i(S_{1i}) - f_i(S_{0i})]}{\mathrm{E}[S_{1i} - S_{0i}]}$$

$$= \mathrm{E}\{\omega_i[(f_i(S_{1i}) - f_i(S_{0i}))/(S_{1i} - S_{0i})]\}, \tag{42}$$

where $\omega_i \equiv (S_{1i} - S_{0i})/\mathrm{E}[S_{1i} - S_{0i}]$. This is a weighted average arc-slope of $f_i(S)$ on the interval $[S_{0i}, S_{1i}]$. We can simplify further using the fact that $f_i(S_{1i}) =$

[36] Note that $\mathrm{P}[\gamma_0 + \gamma_1 > \nu_i > \gamma_0]\mathrm{P}[Z_i = 1] + \mathrm{P}[\gamma_0 > \nu_i] = (\mathrm{E}[D_i \mid Z_i = 1] - \mathrm{E}[D_i \mid Z_i = 0])\mathrm{P}(Z_i = 1) + \mathrm{E}[D_i \mid Z_i = 0] = \mathrm{P}[D_i = 1]$, so the weights sum to one. In the special case where $\mathrm{P}[\gamma_0 > \nu_i] = 0$ for everyone, LATE and the effect of treatment on the treated are the same.

[37] For example, if $f_i(S) = \beta_0 + \rho_i S + \eta_i$, then we assume $\{\rho_i, \eta_i, \kappa_{1i}, \nu_i\}$ are independent of Z_i.

$f_i(S_{0i}) + f_i'(S_i^*)(S_{1i} - S_{0i})$, for some S_i^* in the interval $[S_{0i}, S_{1i}]$.[38] Now we can write the Wald estimator as an average derivative:

$$\frac{\mathrm{E}[f_i(S_{1i}) - f_i(S_{0i})]}{\mathrm{E}[S_{1i} - S_{0i}]} = \frac{\mathrm{E}[(S_{1i} - S_{0i})f_i'(S_i^*)]}{\mathrm{E}[S_{1i} - S_{0i}]} = \mathrm{E}[\omega_i f_i'(S_i^*)]. \tag{43}$$

Given the monotonicity assumption, ω_i is positive for everyone, so the Wald estimator is a weighted average of individual-specific slopes at a point in the interval $[S_{0i}, S_{1i}]$. The weight each person gets is proportional to the size of the causal effect of the instrument on him or her. The range of variation in $f_i(S)$ summarized by this average is always between S_{0i} and S_{1i}.

Angrist et al. (1995) note that the Wald estimator can be characterized more precisely in a number of important special cases. First, suppose that the effect of the instrument is the same for everybody, i.e., κ_{1i} is constant. Then we obtain the average derivative $\mathrm{E}[f_i'(S_i^*)]$, and no weighting is involved. If $f_i(S)$ is linear in S, as in Section 2.2.1, but with a random coefficient, ρ_i then the Wald estimator is a weighted average of the random coefficient: $\mathrm{E}[(S_{1i} - S_{0i})\rho_i]/\mathrm{E}[S_{1i} - S_{0i}]$. If κ_{1i} is constant *and* $f_i(S)$ is linear, then the Wald estimator is the population average slope, $\mathrm{E}[\rho_i]$.

Another interesting special case is when $f_i(S)$ is a quadratic function of S, as in Lang (1993) and Card's (1995) parameterization of a structural human-capital earnings function. The quadratic function captures the notion that returns to schooling decline as schooling increases. Note that for a quadratic function, the point of linearization is always $S_i^* = (S_{1i} + S_{0i})/2$. The Wald estimator is therefore

$$\mathrm{E}[\omega_i f_i'((S_{1i} + S_{0i})/2)],$$

i.e., a weighted average of individual slopes at the midpoint of the interval $[S_{0i}, S_{1i}]$ for each person. The fact that the weights are proportional to $S_{1i} - S_{0i}$ sometimes has economic significance. In the Card and Lang models, for example, the first-stage effect, $S_{1i} - S_{0i}$, is assumed to be proportional to individual discount rates. Since people with higher discount rates get less schooling and the schooling-earnings relationship has been assumed to be concave, this tends to make the Wald estimate higher than the population average return. Lang (1993) called this phenomenon "discount rate bias".

In some applications, it is interesting to characterize the range of variation captured by the Wald estimator further. Returning to (42), which describes the estimator as a weighted average of slopes in the interval $[S_{0i}, S_{1i}]$, it seems natural to ask which values are most likely to be covered by this interval. For example, does $[S_{0i}, S_{1i}]$ usually cover 12 years of education, or is it more likely to cover 16 years? The probability $S \in [S_{0i}, S_{1i}]$ is $\mathrm{P}[S_{1i} \geq S \geq S_{0i}]$. Because S_i is discrete, it is easier to work with $\mathrm{P}[S_{1i} > S \geq S_{0i}]$, since this can be expressed as

[38] Here we assume that $f_i(S)$ is continuously differentiable with domain equal to a subset of the real line.

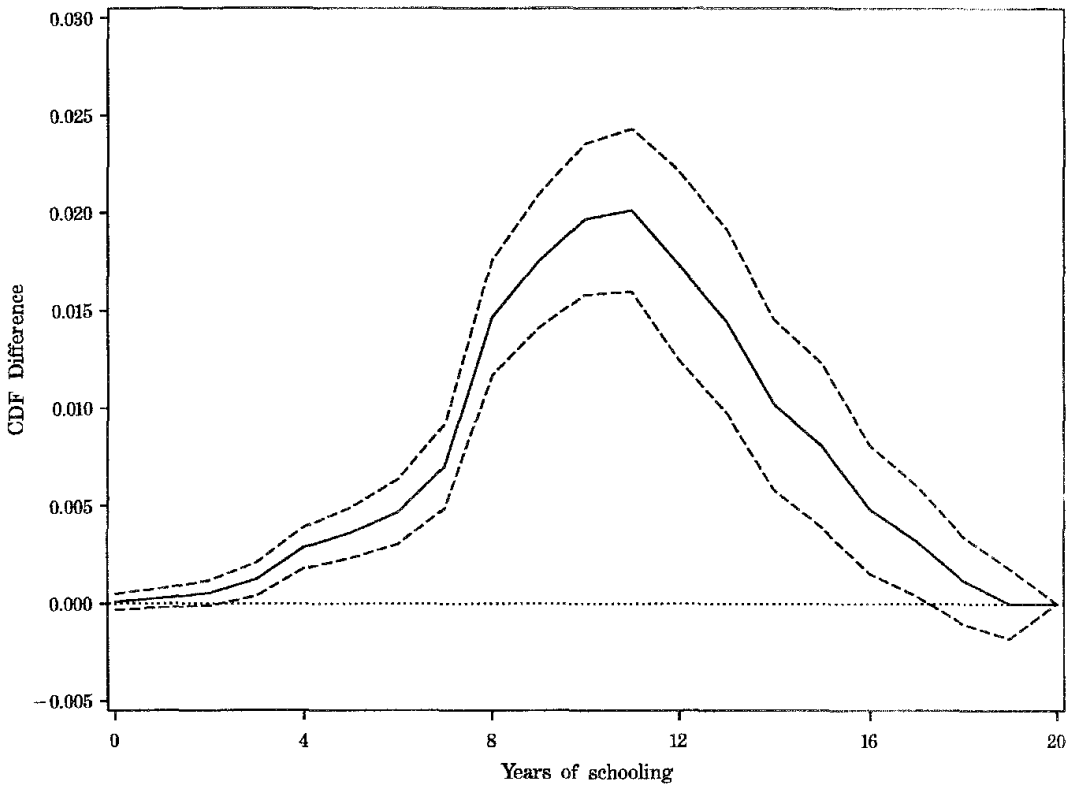

Fig. 5. First quarter–fourth quarter difference in schooling CDFs, for men born 1930–1939 in the 1980 Census. The dotted lines are 95% confidence intervals.

$$\mathrm{P}[S_{1i} > S \geq S_{0i}] = \mathrm{P}[S_{1i} > S] - \mathrm{P}[S_{0i} > S] = \mathrm{P}[S_i \leq S \mid Z_i = 0] - \mathrm{P}[S_i \leq S \mid Z_i = 1]. \tag{44}$$

This is the difference in the cumulative distribution function (CDF) of schooling with the instrument switched off and on. The schooling values where the CDF-gap is largest are those most likely to be covered by the interval $[S_{0i}, S_{1i}]$, and therefore most often represented in the Wald/weighted average.

Angrist and Imbens (1995) used Eq. (44) to interpret the Wald estimates of the returns to schooling reported by Angrist and Krueger (1991).[39] They report a Wald estimate based on first quarter/fourth quarter differences in log weekly wages and years of schooling using data on men born 1930–1939 in the 1980 Census. Their Wald estimate is 0.089, and the corresponding OLS estimate is 0.07. The first quarter/fourth quarter difference in CDFs is plotted in Fig. 5. The difference is largest in the 8–14 years-of-schooling range. This is not surprising since compulsory attendance laws mainly affect high school students, i.e., those with 8–12 years of education. The CDF gap for men with more than 12 years of schooling

[39] See Kling (1998) for a similar analysis of instrumental variables estimates using distance to college as an instrument for schooling.

may be caused by men who were compelled to complete high school but then attended college later.

Finally, we note that the discussion of IV in heterogeneous and non-linear models so far has ignored covariates. 2SLS estimates in heterogeneous-effects models with covariates can be interpreted in much the same way as regression estimates of models with covariates were interpreted in Section 2.3.1. That is, IV estimates in models with covariates can be thought of as producing a weighted average of covariate-specific Wald estimates as long as the model for covariates is saturated and $E[S_i \mid X_i, Z_i]$ is used as an instrument. In other cases it seems reasonable to assume that some sort of approximate weighted average is being generated, but we are unaware of a precise causal interpretation that fits all cases.[40]

2.4. *Refutability*

Causality can never be proved by associations in non-experimental data. But sometimes the lack of association between variables for a particular group, or the occurrence of an association between the "causing variable" and outcome variable for a group thought to be unaffected by the treatment, can cast doubt on, or even refute, a causal interpretation. R.A. Fisher (quoted in Cochran, 1965) argued that the case for causality is stronger when the causal model has many implications that appear to hold. For this reason, he suggested that scientific theories be made "complicated," in the sense that they yield many testable implications.

A research design is more likely to be successful at assessing causality if possibilities for checking collateral implications of causal processes are "built in." At one level, this involves estimating less restrictive models. A good example is Freeman's (1984) panel data study of union status, which looks separately at workers who join unions and leave unions. If unions truly raise wages of their members, then workers who move from non-union to union jobs should experience a raise, and workers who move from union to non-union jobs should experience a pay cut. Although a less restrictive model may yield imprecise estimates or be subject to different biases which render the results difficult to interpret (e.g., different unobserved variables may cause workers to join and exit union jobs), a causal story is strengthened if the results of estimating a less restrictive model are consistent with the story.

In addition to these considerations of robustness, a causal model will often yield testable predictions for sub-populations in which the "treatment effect" should *not* be observed, either because the sub-population is thought to be immune to the treatment or did not receive the treatment. Perhaps the best-known example of this type of analysis is Bound's (1989) study of the effect of Disability Insurance (DI) benefits on the labor force participation rates of older men. Earlier studies (e.g., Parsons, 1980) established an inverse

[40] A recent effort in this direction is Abadie (1998), who presents conditions under which 2SLS estimates can be interpreted as the best linear predictor for an underlying causal relationship. He also introduces a new IV estimator that always has this property for models with a single binary instrument.

relationship between the participation rate and the DI benefit-wage replacement ratio. But because the replacement ratio is a decreasing function of a worker's past earnings, Bound argued that this association may reflect pre-existing patterns of labor force participation rather than a causal response to DI benefits.[41]

To test the causal interpretation of earlier work, Bound performed two types of analyses. First, he estimated essentially the same econometric model of the relationship between employment and potential DI benefits that had been estimated previously, *except* he estimated the model for a sub-sample of older men who had never applied for DI. Because one would not expect DI benefits to provide a strong work disincentive for this sub-sample, there should be a much weaker relationship, or no relationship at all, if the causal interpretation of DI benefit coefficients is correct. Instead, he found that DI benefits had about the same effect in this sample as in a sample that included men who actually applied for and received DI benefits, suggesting that a causal interpretation of the effect of DI benefits was not warranted. Second, Bound examined the labor force behavior of men who applied for DI but were turned down. He reasoned that because men in this sub-sample were less severely disabled than men who received DI, the labor force participation rate of this sub-sample provided a "natural 'control' group" (p. 482) for predicting the upper bound of the labor force participation rate of DI recipients *had they been* denied DI benefits. Because half of the presumably healthier rejected DI applicants did not work even without receiving benefits, Bound concluded that most DI recipients did not work because they were disabled, not because DI benefits induced them to leave the labor force.

Notions of "refutability" also carry over to IV models. In Angrist and Krueger (1991) we were concerned that quarter of birth, which was the instrument for schooling, might have influenced educational attainment through some mechanism other than the interaction of school start age and compulsory schooling laws. To test this threat to a causal interpretation of the IV estimates, we examined whether quarter of birth influenced schooling or earnings for college graduates, who presumably were unaffected by compulsory schooling laws. Although quarter of birth had an effect on these outcomes for college graduates, the effect was weak and had a different pattern than that found for the less-than-college group, suggesting that compulsory schooling was responsible for the effects of quarter of birth in the less-than-college sample.

Tests of refutability may have flaws. It is possible, for example, that a subpopulation that is believed to be unaffected by the intervention is indirectly affected by it. For example, Parsons (1991) argues that rejected DI applicants are a misleading control group because they may exit the labor force to strengthen a possible appeal of their rejected application or a future re-application for DI benefits.[42] Likewise, some students who complete high school because of compulsory schooling may be induced to go on to college as a result, invalidating our 1991 test of refutability. An understanding of the institutions underlying the intervention being evaluated is necessary to assess tests of

[41] Welch (1977) provides a closely related criticism of work on Unemployment Insurance benefits.

[42] Bound (1989) considered and rejected these threats to his control group. See also Bound's (1991) response to Parsons (1991).

refutability, as well as to identify subpopulations that are immune from the intervention according to the causal story but still subject to possible confounding effects.

Lastly, there has been much recent interest in evaluating entire research designs, as in Lalonde's (1986) landmark study comparing experimental and non-experimental research methods. Only rarely, however, have experiments been conducted that can be used to validate non-experimental research strategies. Nonetheless, non-experimental research designs can still be assessed by comparing "pre-treatment" trends for the treatment and comparison group (e.g., Ashenfelter and Card, 1985; Heckman and Hotz, 1989) or by looking for effects where there should be none (e.g., Bound, 1989). We provide another illustration of this point with some new evidence on the differences-in-differences approach used in Card's (1990) immigration study.

In the summer of 1994, tens of thousands of Cubans boarded boats destined for Miami in an attempt to emigrate to the United States in a second Mariel Boatlift that promised to be almost as large as the first one, which occurred in the summer of 1980. Wishing to avoid the political fallout that accompanied the earlier boatlift, the Clinton Administration interceded and ordered the Navy to divert the would-be immigrants to a base in Guantanamo Bay. Only a small fraction of the Cuban emigres ever reached the shores of Miami. Hence, we call this event, "The Mariel Boatlift That Did not Happen."

Had the migrants been allowed to reach the United States, there is little doubt that researchers would have used this "natural experiment" to extend Card's (1990) influential study of the earlier influx of Cuban immigrants. Nonetheless, we can use this "non-event" to explore Card's research design. In particular, we can ask whether Miami's and the comparison cities' experiences were in fact similar absent the large wave of immigrants to Miami. Fig. 1, which we referred to earlier in the discussion of Card's paper, shows that non-agricultural employment growth in Miami tracks that of the four comparison cities rather well in the year before and few years after the summer of 1994. (A vertical bar indicates the date of the thwarted boatlift.) To provide a more detailed analysis by ethnic group, we followed Card and calculated unemployment rates for Whites, Blacks and Hispanics in Miami and the four comparison cities using data from the CPS Outgoing Rotation Groups. These results are reported in Table 7.

The Miami unemployment data are imprecise and variable, but still indicate a large increase in unemployment in 1994, the year the potential immigrants were diverted to Guantanamo Bay. On the other hand, 1994 was the first year the CPS redesign was implemented (see Section 3.1). We therefore take 1993 as the "pre" period and 1995 as the "post" period for a difference-in-differences comparison. For Whites and Hispanics, the unemployment rate fell in Miami and fell even more in the comparison cities between the pre and post periods, though the difference between these two changes is not significant. This is consistent with a causal interpretation of Card's (1990) results, which attributes the difference-in-differences to the effect of immigration. For blacks, however, the unemployment rate rose by 3.6 percentage points in Miami between 1993 and 1995, while it fell by 2.7 points in the comparison cities. The 6.3 point difference-in-differences estimate is on the margin of statistical significance ($t = 1.70$), and would have made it

Table 7
Unemployment rates of individuals age 16–61 in Miami and four comparison cities, 1988–1996[a]

	1988	1989	1990	1991	1992	1993	1994	1995	1996
Miami									
Whites	2.8	3.6	3.3	5.7	4.2	4.9	6.2	3.9	4.4
	(0.8)	(0.9)	(0.9)	(1.2)	(1.1)	(1.3)	(1.4)	(1.4)	(1.2)
Blacks	10.0	11.8	11.9	8.8	10.1	10.1	15.1	13.7	11.1
	(1.7)	(1.8)	(1.9)	(1.9)	(2.0)	(2.1)	(2.4)	(2.8)	(2.4)
Hispanics	5.5	7.6	7.2	9.1	10.3	8.5	9.4	8.4	8.9
	(1.4)	(1.5)	(1.4)	(1.6)	(1.7)	(1.6)	(1.8)	(1.8)	(1.6)
Comparison cities									
Whites	4.2	3.5	3.8	4.9	5.1	5.4	5.0	4.1	4.1
	(0.3)	(0.2)	(0.2)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)
Blacks	11.3	8.4	9.6	9.6	13.6	11.5	10.9	8.8	9.3
	(0.9)	(0.8)	(0.8)	(0.9)	(1.0)	(0.9)	(0.9)	(0.8)	(0.8)
Hispanics	7.2	7.5	5.8	9.1	10.9	11.3	11.0	10.0	9.4
	(0.7)	(0.6)	(0.4)	(0.5)	(0.6)	(0.6)	(0.6)	(0.7)	(0.6)

[a] Note: Standard errors are in parentheses. The four comparison cities (Atlanta, Houston, Los Angeles, and Tampa-St. Petersburg), are the same comparison cities used by Card (1990). The reported unemployment rates are from the authors' tabulations of CPS Outgoing Rotation Groups.

look like the immigrant flow had a negative impact on Blacks in Miami in a DD study. Since there was no immigration shock in 1994, this illustrates that different labor market trends can generate spurious findings in research of this type.

3. Data collection strategies

Table 1 documents that labor economists use many different types of datasets. The renewed emphasis on quasi-experiments in empirical research places a premium on finding datasets for a particular population and time period containing certain key variables. Often this type of analysis requires large samples, because only part of the variation in the variables of interest is used in the estimation. Familiarity with datasets is as necessary for modern labor economics as is familiarity with economic theory or econometrics. Knowledge of the populations covered by the main surveys, the design of the surveys, the response rate, the variables collected, the size of the samples, the frequency of the surveys, and any changes in the surveys over time is essential for successfully implementing an empirical strategy and for evaluating others' empirical research. This section provides an overview of the most commonly used datasets and data collection strategies in labor economics.

Table 8
Commonly-used micro datasets in labor economics

Dataset	Sampling unit	Strengths	Weaknesses	Key variables
Current Population Survey (CPS), esp. March, May, and Outgoing Rotation Groups (ORG)	Household	Large samples; many years of data; many questions and supplements; basic labor force data every month; can link one survey month to another	Some questions change over time; earnings truncated and truncation point varies over time; about 80–85% non-response to income questions; mismatches in linked files	March: annual earnings and work experience, migration, cash and non-cash income, pension and health insurance coverage; May: earnings, multiple job holding, and premium pay; ORG: Labor force questions and earnings; available after 1979
Panel Study of Income Dynamics (PSID)	Household and household spin-offs	Long panel; many labor questions; low non-response rate after first few years; useful for intergenerational issues; oversamples poor families	Small for some purposes; wage rate not always available	Income sources and amounts, employment, family composition changes, demographic events, housing and food expenditures, wealth, housework time, and health status

National Longitudinal Surveys (NLS)	Individual	Concentrates on cohorts by specific age ranges; designed to be longitudinal data	Not all cohorts covered; alternate questions asked in different years; hours data are inconsistent	Labor force experience, education and training, local labor market variables, and cohort-specific questions
Census Data, esp. 1940, 1950, 1960, 1970, 1980, and 1990	Household	Gigantic samples; precise information on census tract	Wage data are noisy; Collects data on fewer variables than CPS; non-response potentially a problem	Place of birth, annual earnings, labor force status, weeks worked, month of birth, and housing variables
Survey of Income and Program Participation (SIPP)	Household	Large sample; emphasis on income and government programs	Short panel dimension; unwieldy dataset; long survey; non-response high; underreporting of program participation	Employment, education and training, program participation, assets and liabilities, migration, fertility, work schedules, child care, pension, property, and time spent out of work

3.1. Secondary datasets

The most commonly used secondary datasets in labor economics are the National Longitudinal Surveys (NLS), the Current Population Survey (CPS), the Panel Study of Income Dynamics (PSID), and the Decennial Censuses. Table 8 summarizes several features of the main secondary datasets used by labor economists. In this section we provide a more detailed discussion of the "big three" micro datasets in labor economics: the NLS, CPS and PSID. We also discuss historical comparability in the CPS and the census.

Perhaps because of its easy-to-use CD-ROM format and the breadth of its questionnaire, the National Longitudinal Surveys are popular in applied work. The NLS actually consists of six age-by-gender datasets: a cohort of 5020 "older men" (age 45–59 in 1966); a cohort of 5083 mature women (age 30–44 in 1967), a cohort of 5225 young men (age 14–24 in 1966); a cohort of 5159 young women (age 14–24 in 1968) in 1968); a cohort of 12,686 "youth" known as the NLSY (age 14–22 in 1979); and a cohort of 7035 children of respondents in the NLSY (age 0–20 in 1986).[43] Sampled individuals are interviewed annually. All but the older-men and young-men surveys continue today.

The CPS is an ongoing survey of more than 50,000 households that is conducted each month by the Census Bureau for the Bureau of Labor Statistics (BLS).[44] Sampled households are included in the survey for four consecutive months, out of the sample for 8 months, and then included for a final four consecutive months. Thus, the survey has a "rotation group" design, with new rotation groups joining or exiting the sample each month. The resulting data are used by the Bureau of Labor Statistics to calculate the unemployment rate and other labor force statistics. The CPS has a hierarchical household-family-person record structure which enables household-level and family-level analyses, as well as individual-level analyses. The design of the CPS has been copied by statistical agencies in several other countries and is similarly used to calculate labor force statistics.

In the US, regular and one-time supplements are included in the survey to collect information on worker displacement, contingent work, school enrollment, smoking, voting, and other important behaviors. In addition, annual income data from several sources are collected each month. A great strength of the CPS is that the survey began in the 1940s, so a long time-series of data are available; on the other hand, there have been several changes that affect the comparability of the data over time, and micro data are only available to researchers for years since 1964. In addition, because of its rotation group design, continuing households can be linked from one month to the next, or between years; however, individuals who move out of sampled households are not tracked, and it is possible that individuals who move into a sampled household may be mis-matched to other individuals' earlier records. High attrition rates are a particular problem in the linked CPS for young workers. Unless a very large sample size is required, it is often preferable to

[43] See NLS Users' Guide (NLS Handbook, 1995) for further information.

[44] See Polivka (1996) for an analysis of recent changes in the CPS, and for a list of supplements.

use a dataset that was designed to track respondents longitudinally, instead of a linked CPS.

The PSID is a national probability sample that originally consisted of 5000 families in 1968.[45] The original families, and new households that have grown out of those in the original sample, have been followed each year since. Consequently, the PSID provides a unique dataset for studying family-related issues. The number of individuals covered by the PSID increased from 18,000 in 1968 to a cumulative total exceeding 40,000 in 1996, and the number of families increased to nearly 8000. Brown et al. (1996) note that the "central focus of the data is economic and demographic, with substantial detail on income sources and amounts, employment, family composition changes and residential location." The PSID is also one of the few datasets that contains information on consumption and wealth. A recent paper by Fitzgerald et al. (1998) finds that, despite attrition of nearly half the sample since 1968, the PSID remained roughly representative through 1989.[46]

The accessibility of secondary datasets is changing rapidly. The ICPSR remains a major collector and distributor of datasets and codebooks. In addition, CPS data can be obtained directly from the Bureau of Labor Statistics. Increasingly, data collection agencies are making their data directly available to researchers via the internet. In 1996, for example, the Census Bureau made the recent March Current Population Surveys, which include supplemental information on annual income and demographic characteristics, available over the internet. Because the March CPS contains annual income data, many researchers have matched these data from one year to the next.

Because secondary datasets are typically collected for a broad range of purposes or for a purpose other than that intended by the researcher, they often lack information required for a particular project. For example, the PSID would be ideal for a longitudinal study of the impact of personal computers on pay, except it lacks information on the use of personal computers. In other situations, the data collector may omit survey items from public-use files to preserve respondent confidentiality. Nonetheless, several large public-use surveys enable researchers to add questions, or will provide customized extracts with variables that are not on the public-use file. For example, Vroman (1991) added supplemental questions to the CPS on the utilization of unemployment insurance benefits. The cost of adding 7 questions was $100,000.[47] From time to time, survey organizations also solicit researchers' advice on new questions or new modules to add to on-going surveys. Since 1993, for example, the PSID sponsors have held an open competition among researchers to add supplemental questions to the survey.

[45] This paragraph is based on Brown et al. (1996).

[46] See also Becketti et al. (1988) for evidence on the representativeness of the PSID.

[47] Because of concern that the additional questions might affect future responses, the supplement was only asked of individuals who were in their final rotation in the sample. The supplement was added to the survey in the months of May, August, November 1989 and February 1990. The sample size was 2859 eligible unemployed individuals.

3.1.1. Historical comparability in the CPS and Census

Statistical agencies are often faced with a tradeoff between adjusting questions to make them more relevant for the modern economy and maintaining historical comparability. Often it seems that statistical agencies place insufficient weight on historical consistency. For example, after 50 years of measuring education by the highest grade of school individuals attended and completed, the Census Bureau switched to measuring educational attainment by the highest degree attained in the 1990 Census. The CPS followed suit in 1992. This is a subtle change in the education data, but one that could potentially affect estimates of the economic return to education (see Park, 1994; Jaeger, 1993). Because many statistics are most informative in comparison to their values in earlier years, it is important that statistical agencies place weight on historical comparability even though the concepts being measured may have changed.

Fortunately, the Bureau of Labor Statistics and the Census Bureau typically introduce a major change in a questionnaire after studying the likely effects of the change on the survey results. Because some changes have a major impact on certain variables (or on certain populations), it is important that analysts be aware of changes in on-going surveys, and of their likely effects. For example, a major redesign of the CPS was introduced in January 1994, after 8 years of study. The redesigned CPS illustrates the importance of questionnaire changes, as well as the difficulty of estimating the likely impact of such changes.

The redesigned CPS is conducted with computer-assisted interviewing technology, which facilitates more complicated skip patterns, more narrowly tailored questions, and dependent interviewing (in which respondents' answers to an earlier month's question are integrated into the current month's question). In addition, the redesign changed the way key labor force variables were collected in the basic, i.e., non-supplemental, CPS. Most importantly, individuals who are not working are now probed more thoroughly for actions taken to search for work. In the older survey, interviewers were instructed to ask a respondent who "appears to be a homemaker" whether she was keeping house most of last week or doing something else. The new question is gender neutral. Another major change concerns the earnings questions. Prior to the redesign, the CPS asked respondents for their usual weekly wage and usual weekly hours.[48] The ratio of these two variables gives the implied hourly wage. The redesigned CPS first asks respondents for the easiest way they could report their total earnings on their main job (e.g., hourly, weekly, annually, or on some other basis), and then collects usual earnings on that basis.

To gauge the impact of the survey redesign on responses in 1992 and 1993, the BLS and Census Bureau conducted an overlap survey in which a separate sample of households was interviewed using the redesigned CPS, while the regular sample was still given the old CPS questionnaire. Then, for the first 5 months of 1994, this overlap sample was given the old CPS, while the regular sample was given the new one. Overlap samples can be extremely informative, but they are also difficult to implement properly. In this instance,

[48] The old CPS also collected hourly earnings for workers who indicated they were paid hourly.

the overlap sample was drawn with different procedures than the regular CPS sample, and there appear to be systematic differences between the two samples which complicate comparisons. Taking account of these difficulties, Polivka (1996) and Polivka and Miller (1995) estimate that the redesign had an insignificant effect on the unemployment rate, although it appears to have raised the employment-to-population ratio of women by 1.6%, raised the proportion of self-employed women by 20%, increased the proportion of all workers who are classified as part-time by 10%, and decreased the fraction of discouraged workers (i.e., those out of the labor force who have given up searching for work because they believe no jobs are available for them) by 50%. Polivka (1997) addresses the effect of the redesign on the derived hourly wage rate. She finds that the redesign causes about a 5% increase in the average earnings of college graduates relative to those who failed to complete high school, and about a 2% increase in the male-female gap. The potential changes in measurement brought about by the redesigned CPS could lead researchers to incorrectly attribute shifts in employment or wages to economic forces rather than to changes in the questionnaire and survey technology.

Three other changes in the CPS are especially noteworthy. First, beginning in 1980 the Annual Demographic Supplement of the March CPS was expanded to ask a more probing set of income questions. The impact of these changes can be estimated because the 1979 March CPS administered the old (pre-1980) questionnaire to five of the eight rotation groups in the sample, and administered the new, more detailed questionnaire to the other three rotation groups.[49] Second, as noted above, the education question (which is on the "control card" rather than the basic monthly questionnaire) was switched from the number of years of school completed to the highest degree attained in 1992 (see Park, 1994; Jaeger, 1993). Third, the "top code" for the income and earnings questions – that is, the highest level of income reported in the public-use file – has changed over time, which obviously may have implications for studies of income inequality.

3.2. Primary data collection and survey methods

It is increasingly common for labor economists to be involved in collecting their own data. Labor economists' involvement in the design and collection of original datasets takes many forms. First, it should be noted that labor economists have long played a major role in the design and collection of some of the major public-use data files, including the PSID and NLS.

Second, researchers have turned to collecting smaller, customized data to estimate specific quantities or describe certain economic phenomenon. Some of Richard Freeman's research illustrates this approach. Freeman and Hall (1986) conducted a survey to estimate the number of homeless people in the US, which came very close to the official Census

[49] See Krueger (1990a) for an analysis of the change in the questionnaire on responses to the question on workers' compensation benefits. The new questionnaire seems to have detected 20% more workers' compensation recipients. See Coder and Scoon-Rogers (1996) for a comparison of CPS and SIPP income measures.

Bureau estimate in 1990. Borjas et al. (1991) conducted a survey of border crossing behavior of illegal aliens to estimate the number of illegal aliens in the US. Freeman (1990) surveyed inner-city youths in Boston, as part of a follow-up to the survey by Freeman and Holzer (1986). Often, data collected in these surveys are combined with secondary data files to derive national estimates.

Third, some surveys have been conducted to probe the sensitivity of results in large-scale secondary datasets, or to probe the sensitivity of responses to question wording or order. For example, Farber and Krueger (1993) surveyed 102 households in which non-union respondents were asked two different questions concerning their likelihood of joining a union, with the order of the questions randomly interchanged. The two questions, which are listed below, were included in earlier surveys conducted by the Canadian Federation of Labor (CFL) and the American Federation of Labor-Congress of Industrial Organizations (AFL-CIO), and had been analyzed by Riddell (1992). Based on comparing responses to these questions, Riddell concluded that American workers have a higher "frustrated demand" for unions than Canadians:

> CFL Q.: Thinking about your own needs, and your current employment situation and expectations, would you say that it is very likely, somewhat likely, not very likely, or not likely at all that you would consider joining or associating yourself with a union or a professional association in the future?
> AFL Q.: If an election were held tomorrow to decide whether your workplace would be unionized or not, do you think you would definitely vote for a union, probably vote for a union, probably vote against a union, or definitely vote against a union?

In their small-scale survey, Farber and Krueger (1993) found that the responses to the CFL question were extremely sensitive to the questions that preceded them. If the AFL question was asked first, 55% of non-union members answered the CFL question affirmatively, but if the CFL question was asked first, 26% of non-union members answered affirmatively to the CFL question.[50] Thus, the Farber and Krueger results suggest a good deal of caution is warranted when interpreting the CFL-style question, especially across countries.

Finally, and of most interest for our purposes, researchers have conducted special-purpose surveys to evaluate natural experiments or exploit unusual circumstances. Probably the best known example of this type of survey is Card and Krueger's (1994) survey of fast food restaurants in New Jersey and Pennsylvania. Other examples include: Ashenfelter and Krueger's (1994) survey of twins; Behrman et al.'s (1996) survey of twins; Mincer and Higuchi's (1988) survey of turnover at Japanese plants in the US and their self-identified competitors; and Freeman and Kleiner's (1990) survey of companies undergoing a union drive and their competitors.

Several excellent volumes have been written on the design and implementation of

[50] The t-ratio for the difference between the proportions is 3.3.

surveys, and a detailed overview of this material is beyond the scope of this paper.[51] But a few points that may be of special interest to labor economists are outlined below.

Customized surveys seem especially appropriate for rare populations, which are likely to be under-represented or not easily identified in public-use datasets. Examples include identical twins, illegal aliens, homeless people, and disabled people.

To conduct a survey, one must obviously have a questionnaire. Preparing a questionnaire can be a time- consuming and difficult endeavor. Survey researchers often find that answers to questions – even factual economic questions – are sensitive to the wording and ordering of questions. Fortunately, one does not have to begin writing a questionnaire from scratch. Survey questionnaires typically are not copyright protected. Because many economists are familiar with existing questionnaires used in the major secondary datasets (e.g., the CPS), and because a great deal of effort typically goes into designing and testing these questionnaires, it is often advisable to copy as many questions as possible verbatim from existing questionnaires when formulating a new questionnaire. Aside from the credibility gained by replicating questions from well known surveys, another advantage of duplicating others' questions is that the results from the sampled population can be compared directly to the population as a whole with the secondary survey. Furthermore, if data from a customized survey are to be pooled with data from a secondary survey, it is essential that the questions be comparable.

One promising recent development in questionnaire design involves "follow-up brackets" (also known as "unfolding" brackets). This technique offers bracketed categories to respondents who initially refuse or are unable to provide an exact value to an open ended question. Juster and Smith (1997) find that follow-up brackets reduced non-response to wealth questions in the Health and Retirement Survey (HRS) and Asset and Health Dynamics among the Oldest Old Survey (AHEAD). See Hurd, et al. (1998) for experimental evidence of "anchoring effects" in responses based on the sequence of unfolding brackets for consumption and savings data in the AHEAD survey. Follow-up brackets have also been used to measure wealth in the PSID. Follow-up brackets seem particularly useful for hard-to-measure quantities, such as income, wealth, saving and consumption.

Lastly, power calculations should guide the determination of sample size prior to the start of a survey. For example, suppose the goal of the survey is to estimate a 95% confidence interval for a mean. With random sampling, the expected sample size (n) required to obtain a confidence interval of width $2W$ is $n = 4\sigma^2/W^2$, where σ^2 is the population variance of the variable in question. Although the variance generally will not be known prior to conducting the survey, an estimate from other surveys can be used for the power calculation. Also notice that in the case of a binary variable (i.e., if the goal is to estimate a proportion, p), the variance is $p(1-p)$, so in the worse-case scenario the variance is $0.25 = 0.5 \times 0.5$. It should also be noted that in complex sample designs involving clustering and stratification, more observations are usually needed than in simple random samples to attain a given level of precision.

[51] See, e.g., Groves (1989), Sudman and Bradburn (1991), and Singer and Presser (1989).

3.3. Administrative data and record linkage

Administrative data, i.e., data produced as a by-product of some administrative function, often provide inexpensive large samples. The proliferation of computerized record keeping in the last decade should increase the number of administrative datasets available in the future. Examples of widely used administrative data bases include social security earnings records (Ashenfelter and Card, 1985; Vroman, 1990; Angrist, 1990), unemployment insurance payroll and benefit records (Anderson, 1993; Katz and Meyer, 1990; Jacobson et al., 1994; Card and Krueger, 1998), workers' compensation insurance records (Meyer et al., 1995; Krueger, 1990b), company personnel records (Medoff and Abraham, 1980; Lazear, 1992; Baker et al., 1994), and college records (Bowen and Bok, 1998). An advantage of administrative data is that they often contain enormous samples or even an entire population. Another advantage is that administrative data often contain the actual information used to make economic decisions. Thus, administrative data may be particularly useful for identifying causal effects from discrete thresholds in administrative decision making, or for implementing strategies that control for selection on observed characteristics.

A frequent limitation of administrative data, however, is that they may not provide a representative sample of the relevant population. For example, companies that are willing to make their personnel records available are probably not representative of all companies. In some cases administrative data have even been obtained as a by-product of court cases or collected by parties with a vested interest in the outcome of the research, in which case there is additional reason to be concerned about the representativeness of the samples.

Another common limitation of administrative data is that they are not generated with research purposes in mind, so they may lack key variables used in economic analyses. For example, social security earnings records lack data on individuals' education. As a consequence, it is common for researchers to link survey data to administrative data, or to link across administrative datasets. Often these links are based on social security numbers or individuals' names. Examples of linked datasets include: the Continuous Longitudinal Manpower Survey (CLMS) survey, which is a link between social security records and the 1976 CPS; the 1973 Exact Match file which contains CPS, IRS, and social security data; and the Longitudinal Employer-Employee Data Set (LEEDS). All of these linked datasets are now dated, but they can still be used for some important historical studies (e.g., Chay, 1996). More recently, the Census Bureau has been engaged in a project to link Census data to the Survey of Manufacturers.

It is also possible to petition government agencies to release administrative data. Although the Internal Revenue Service severely limits disclosure of federal administrative records collected for tax purposes, State data is often accessible and even federal data can still be linked and released under some circumstances. For example, Angrist (1998) linked military personnel records to Social Security Administration (SSA) data. The HRS has also linked SSA data to survey-based data. Some new Social Security-Census linked datasets are available on a restricted basis through the Census Regional Data Centers. Furthermore, many states provide fairly free access to UI payroll tax data

to researchers for the purpose of linking data.[52] There is also a literature on data release schemes for administrative records that preserve confidentiality and meet legal requirements (see, e.g., Duncan and Pearson, 1991).

3.4. Combining samples

Although in some cases individual records can be linked across different data sources, an alternative linkage strategy exploits the fact that many of the estimators used in empirical research can be constructed from separate sets of first and second moments. So, in principle, individual records with a full complement of variables are not always needed to carry out a multivariate analysis. It is sometimes enough to have all the moments required, even though these moments may be drawn from more than one sample. In practice, this makes it possible to undertake empirical projects even if the required data are not available in any single source.

Recent versions of the multiple-sample approach to empirical work include the two-sample instrumental variables estimators developed by Arellano and Meghir (1992) and Angrist and Krueger (1992, 1995), and used by Lusardi (1996), Japelli et al. (1998), and Kling (1998). The use of two samples to estimate regression coefficients dates back at least to Durbin (1953), who discussed the problem of how to update OLS estimates with information from a new sample. Maddala (1971) discussed a similar problem using a maximum likelihood framework. This idea was recently revived by Imbens and Lancaster (1994), who address the problem of how to use macroeconomic data in micro-econometric models. Deaton (1985) focuses on estimating panel data models with aggregate data on cohorts.

4. Measurement issues

In his classic volume on the accuracy of economic measurement, Morgenstern (1950) quotes the famed mathematician Norbert Wiener as remarking, "Economics is a one or two digit science." The fact that the focus of most empirical research has moved from aggregate time-series data to micro-level cross-sectional and longitudinal survey data in recent years only magnifies the importance of measurement error, because (random) errors tend to average out in aggregate data. Consequently, a good deal of attention has been paid to the extent and impact of "noisy" data in the last decade, and much has been learned.

Measurement error can arise for several reasons. In survey data, a common source of measurement error is that respondents give faulty answers to the questions posed to them.[53] For example, some respondents may intentionally exaggerate their income or

[52] An example is Krueger and Kruse (1996), which links New Jersey unemployment insurance payroll tax data to a dataset the authors collected in a survey of disabled individuals.

[53] Even well-trained economists can make errors of this sort. Harvard's Dean of Faculty Henry Rosovsky (1990, p. 40) gives the following account of a meeting he had with an enraged economics professor who complained about his salary: "After a quick calculation, this quantitatively oriented economist concluded that his raise was all of 1%: an insult and an outrage. I had the malicious pleasure of correcting his mistaken calculation. The raise was 6%: he did not know his own salary and had used the wrong base."

educational attainment to impress the interviewer, or they may shield some of their income from the interviewer because they are concerned the data may somehow fall into the hands of the IRS, or they may unintentionally forget to report some income, or they may misinterpret the question, and so on. Even in surveys like the SIPP, which is specifically designed to measure participation in public programs like UI and AFDC, respondents appear to under-report program participation by 20–40% (see Marquis et al., 1996). It should also be stressed that in many situations, even if all respondents correctly answer the interviewers' questions, the observed data need not correspond to the concept that researchers would like to measure. For example, in principle, human capital should be measured by individuals' acquired knowledge or skills; in practice it is measured by years of schooling.[54]

For these reasons, it is probably best to think of data as routinely being mismeasured. Although few economists consider measurement error the most exciting research topic in economics, it can be of much greater practical significance than several hot issues. Topel (1991), for example, provides evidence that failure to correct for measurement error greatly affects the estimated return to job tenure in panel data models. Fortunately, the direction of biases caused by measurement error can often be predicted. Moreover, in many situations the extent of measurement error can be estimated, and the parameters of interest can be corrected for biases caused by measurement error.

4.1. Measurement error models

4.1.1. The classical model

Suppose we have data on variables denoted X_i and Y_i for a sample of individuals. For example, X_i could be years of schooling and Y_i log earnings. The variables X_i and Y_i may or may not equal the correctly-measured variables the researcher would like to have data on, which we denote X_i^* and Y_i^*. The error in measuring the variables is simply the deviation between the observed variable and the correctly-measured variable: for example, $e_i = X_i - X_i^*$, where e_i is the measurement error in X_i. Considerations of measurement error usually start with the assumption of "classical" measurement errors.[55] Under the classical assumptions, e_i is assumed to have the properties $C(e_i, X_i^*) = E(e_i) = 0$. That is, the measurement error is just mean-zero "white noise". Classical measurement error is not a necessary feature of measurement error; rather, these assumptions are best viewed as a convenient starting point.

What are the implications of classical measurement error? First, consider a situation in which the dependent variable is measured with error. Specifically, suppose that $Y_i = Y_i^* + u_i$, where Y_i is the observed dependent variable, Y_i^* is the correctly-measured,

[54] Measurement error arising from the mismatch between theory and practice also occurs in administrative data. In fact, this may be a more severe problem in administrative data than in survey data.

[55] References for the effect of measurement error include Duncan and Hill (1985), Griliches (1986), Fuller (1987), and Bound and Krueger (1991).

desired, or "true" value of the dependent variable, and u_i is classical measurement error. If Y_i is regressed on one or more correctly-measured explanatory variables, the expected value of the coefficient estimates is not affected by the presence of the measurement error. Classical measurement error in the dependent variable leads to less precise estimates – because the errors will inflate the standard error of the regression – but does not bias the coefficient estimates.[56]

Now consider the more interesting case of measurement error in an explanatory variable. For simplicity, we focus on a bivariate regression, with mean zero variables so we can suppress the intercept. Suppose Y_i^* is regressed on the observed variable X_i, instead of on the correctly-measured variable X_i^*. The population regression of Y_i^* on X_i^* is

$$Y_i^* = X_i^*\beta + \varepsilon_i, \tag{45}$$

while if we make the additional assumption that the measurement error (e_i) and the equation error (ε_i) are uncorrelated, the population regression of Y_i^* on X_i is

$$Y_i^* = X_i\lambda\beta + \tilde{\varepsilon}_i, \tag{46}$$

where $\lambda = C(X^*, X)/V(X)$. If X_i is measured with classical measurement error, then $C(X^*, X) = V(X^*)$ and $V(X) = V(X^*) + V(e)$, so the regression coefficient is necessarily *attenuated*, with the proportional "attenuation bias" equal to $(1 - \lambda) < 1$.[57] The quantity λ is often called the "reliability ratio". If data on both X_i^* and X_i were available, the reliability ratio could be estimated from a regression of X_i^* on X_i. A higher reliability ratio implies that the observed variability in X_i contains less noise.

Although classical measurement error models provide a convenient starting place, in some important situations classical measurement error is impossible. If X_i is a binary variable, for example, then it *must* be the case that measurement errors in X_i are dependent on the values of X_i^*. This is because a dummy variable can only be misclassified in one of two ways (a true 1 can be classified as a 0, and a true 0 can be classified as a 1), so only two values of the error are possible and the error automatically depends on the true value of the variable. An analogous situation arises with variables whose range is limited. Aigner (1973) shows that random misclassification of a binary variable still biases a bivariate regression coefficient toward 0 even though the resulting measurement error is not classical. But, in general, if measurement error in X_i is not classical, the bias factor could be greater than or less than one, depending on the correlation between the measurement error and the true variable. Note, however, that regardless of whether or not the classical

[56] If the measurement error in the dependent variable is not classical, then the regression coefficients will be biased. The bias will equal the coefficients from a hypothetical regression of the measurement error on the explanatory variables.

[57] Notice these are descriptions of population regressions. The estimated regression coefficient is asymptotically biased by a factor $(1 - \lambda)$, although the bias may differ in a finite sample. If the conditional expectation of Y is linear in X, such as in the case of normal errors, the expected value of the bias is $(1 - \lambda)$ in a finite sample.

measurement error assumptions are met, the proportional bias $(1-\lambda)$ is still given by one minus the regression coefficient from a regression of X_i^* on X_i.[58]

Another important special case of non-classical measurement error occurs when a group average is used as a "proxy-variable" for an individual-level variable in micro data. For example, average wages in an industry or county might be substituted for individual wage rates on the right-hand side of an equation if micro wage data are missing. Although this leads to measurement error, since the proxy-variable replaces a desired regressor, asymptotically there is no measurement-error bias in a bivariate regression in this case. One way to see this is to note that the coefficient from a regression of, say, X_i on $\mathrm{E}[X_i \mid \text{industry } j]$ has a probability limit of 1.

So far the discussion has considered the case of a bivariate regression with just one explanatory variable. As noted in Section 2, adding additional regressors will typically exacerbate the impact of measurement error on the coefficient of the mismeasured variable because the inclusion of additional independent variables absorbs some of the signal in X_i, and thereby reduces the residual signal-to-noise ratio. Assuming that the other explanatory variables are measured without error, the reliability ratio conditional on other explanatory variables becomes $\lambda' = (\lambda - R^2)/(1 - R^2)$ where R^2 is the coefficient of determination from a regression of the mismeasured X_i on the other explanatory variables. If the measurement error is classical, then $\lambda' \leq \lambda$. And even if the measurement error is not classical, it still remains true that when there are covariates in Eq. (45), the proportional bias is given by the coefficient on X_i in a regression of X_i^* on X_i and the covariates. Note, however, that in models with covariates, the use of aggregate proxy variables may generate asymptotic bias.

An additional feature of measurement error important for applied work is that, for reasons similar to those raised in the discussion of models with covariates, attenuation bias due to classical measurement error is generally exacerbated in panel data models. In particular, if the independent variable is expressed in first differences and if we assume that X_i^* and e_i are covariance stationary, the reliability ratio is

$$\lambda = V(X_i^*)/\{V(X_i^*) + V(e_i)[(1 - \tau)/(1 - r)]\}, \tag{47}$$

where r is the coefficient of first-order serial correlation in X_i^* and τ is the first-order serial correlation in the measurement error. If the (positive) serial correlation in X_i^* exceeds the (positive) serial correlation in the measurement error, attenuation bias is greater in first-differenced data than in cross-sectional data (Griliches and Hausman, 1986). Classical measurement errors are usually assumed to be serially uncorrelated ($\tau = 0$), in which case the attenuation bias is greater in a first-differenced regression than in a levels regression.

[58] This result requires the previously mentioned assumption that e_i and ε_i be uncorrelated. It may also be the case that the measurement error is not mean zero. Statistical agencies often refer to such phenomenon as "non-sampling error" (see, e.g., McCarthy, 1979). Such non-sampling errors may arise if the questionnaire used to solicit information does not pertain to the economic concept of interest, or if respondents systematically under or over report their answers even if the questions do accurately reflect the relevant economic concepts. An important implication of non-sampling error is that aggregate totals will be biased.

The intuition for this is that some of the signal in X_i cancels out in the first-difference regression because of serial correlation in X_i^*, while the effect of independent measurement errors is amplified because errors can occur in the first or second period. A similar situation arises if differences are taken over dimensions of the data other than time, such as between twins or siblings.

Finally, note that if an explanatory variable is a function of a mismeasured dependent variable, the measurement errors in the dependent and independent variables are automatically correlated. Borjas (1980) notes that this situation often arises in labor supply equations where the dependent variable is hours worked and the independent variable is average hourly earnings, derived by dividing weekly or annual earnings by hours worked. In this situation, measurement error in Y_i will induce a negative bias when $(Y_i^* + u_i)$ is regressed on $X_i^*/(Y_i^* + u_i)$. In other situations, both the dependent and independent variables may have the same noisy measure in the denominator, such as when the variables are scaled to be per capita (common in the economic growth literature). If the true regression parameter were 0, this would bias the estimated coefficient toward 1. The extent of bias in these situations is naturally related to the extent of the measurement error in the variable that appears on both the right-hand and left-hand side of the equation.

4.1.2. Instrumental variables and measurement error

One of the earliest uses of IV was as a technique to overcome errors-in-variables problems. For example, in his classic work on the permanent income hypothesis, Friedman (1957) argued that annual income is a noisy measure of permanent income. The grouped estimator he used to overcome measurement errors in permanent income can be thought of as IV. It is now well known that IV yields consistent parameter estimates even if the endogenous regressor is measured with classical error, assuming that a valid instrument exists. Indeed, one explanation why IV estimates of the return to schooling frequently exceed OLS estimates is that measurement error attenuates the OLS estimates (e.g., Griliches, 1977).

In a recent paper, Kane et al. (1997) emphasize that IV can yield inconsistent parameter estimates if the endogenous regressor is measured with non-classical measurement error.[59] Specifically, they show that if the mismeasured endogenous regressor, X_i, is a dummy variable, the measurement error will be correlated with the instrument, and typically bias the magnitude of IV coefficients upward.[60] The probability limit of the IV estimate in this case is

$$\frac{\beta}{1 - \mathrm{P}(X_i = 0 \mid X_i^* = 1) - \mathrm{P}(X_i = 1 \mid X_i^* = 0)}. \tag{48}$$

Intuitively, the parameter of interest is inflated by one minus the sum of the probabilities of

[59] A similar point has been made by James Heckman in an unpublished comment on Ashenfelter and Krueger (1994).

[60] The exception is if X_i is so poorly measured that it is negatively correlated with X_i^*.

the two types of errors that can be made in measuring X_i (observations that are 1's can be classified as 0's, and observations that are 0's can be classified as 1's). The reason IV tends to overestimate the parameter of interest is that if X_i is a binary variable, the value of the measurement error is automatically dependent on the true value of X_i^*, and therefore must be correlated with the instrumental variable because the instrumental variable is correlated with X_i^*. Combining this result with the earlier discussion of attenuation bias, it should be clear that if the regressor is a binary variable (in a bivariate regression), the probability limit of the OLS and IV estimators bound the coefficient of interest, assuming the specifications are otherwise appropriate. In the more general case of non-classical measurement error in a continuous explanatory variable, IV estimates can be attenuated or inflated, as in the case of OLS.

4.2. The extent of measurement error in labor data

Mellow and Sider (1983) provide one of the first systematic studies of the properties of measurement error in survey data. They examined two sources of data: (1) employee-reported data from the January 1977 CPS linked to employer-reported data on the same variables for sampled employees; (2) an exact match between employees and employers in the 1980 Employment Opportunity Pilot Project (EOPP). Mellow and Sider focus on the extent of agreement between employer and employee reported data, rather than the reliability of the CPS data per se. For example, they find that 92.3% of employers and employees reported the same one-digit industry, while at the three-digit-industry level, the rate of agreement fell to 71.1%. For wages, they find that the employer-reported data exceeded the employee-reported data by about 5%. The mean unionization rate was slightly higher in the employer-reported data than in the employee-reported data. They also found that estimates of micro-level human capital regressions yielded qualitatively similar results whether employee-reported or employer-reported data are used. This similarity could result from the occurrence of roughly equal amounts of noise in the employer- and employee-reported data.

Several other studies have estimated reliability ratios for key variables of interest to labor economists. Two approaches to estimating reliability ratios have typically been used. First, if the researcher is willing to call one source of data the truth, then λ can be estimated directly as the ratio of the variances: $V(X_i^*)/V(X_i)$. Second, if two measures of the same variable are available (denoted X_{1i} and X_{2i}), and if the errors in these variables are uncorrelated with each other and uncorrelated with the true value, then the covariance between X_{1i} and X_{2i} provides an estimate of $V(X_i^*)$. The reliability ratio λ can then be estimated by using the variance of either measure as the denominator or by using the geometric average of the two variances as the denominator. The former can be calculated as the slope coefficient from a regression of one measure on the other, and the latter can be calculated as the correlation coefficient between the two measures. If a regression approach is used, the variable that corresponds most closely to the data source that is usually used in analysis

should be the explanatory variable (because the two sources may have different error variances).

An example of two mismeasured reports on a single variable are respondents' reports of their parents' education in Ashenfelter and Krueger's (1994) twins study. Each adult twin was asked to report the highest grade of education attained by his or her mother and father. Because each member of a pair of twins has the same parents, the responses should be the same, and there is no reason to prefer one twin's response over the other's. Differences between the two responses for the same pair of twins represent measurement error on the part of at least one twin. The correlation between the twins' reports of their father's education is 0.86, and the correlation between reports of their mother's education is 0.84. These figures probably overestimate the reliability of the parental education data because the reporting errors are likely to be positively correlated; if a parent mis-represented his education to one twin, he is likely to have similarly mis-represented his education to the other twin as well.

Table 9 summarizes selected estimates of the reliability ratio for self-reported log earnings, hours worked, and years of schooling, three of the most commonly studied variables in labor economics. These estimates provide an indication of the extent of attenuation bias when these variables appear as explanatory variables. All of the estimates of the reliability of earnings data in the table are derived by comparing employees' reported earnings data with their employers' personnel records or tax reports. The estimates from the PSID validation study are based on data from a single plant, which probably reduces the variance of correctly-measured variables compared to their variance in the population. This in turn reduces the estimated reliability ratio if reporting errors have the same distribution in the plant as in the population.

Estimates of λ for cross-sectional earnings range from 0.70 to 0.80 for men; λ is somewhat higher for women. The estimated reliability falls to about 0.60 when the earnings data are expressed as year-to-year changes. The decline in the reliability of the earnings data is not as great if 4-year changes are used instead of annual changes, reflecting the fact that there is greater variance in the signal in earnings over longer time periods. Interestingly, the PSID validation study also suggests that hours data are considerably less reliable than earnings data.

The reliability of self-reported education has been estimated by comparing the same individual's reports of his own education at different points in time, or by comparing different siblings' reports of the same individual's education. The estimates of the reliability of education are in the neighborhood of 0.90. Because education is often an explanatory variable of interest in a cross-sectional wage equation, measurement error can be expected to reduce the return to a year of education by about 10% (assuming there are no other covariates). The table also indicates that if differences in educational attainment between pairs of twins or siblings are used to estimate the return to schooling (e.g., Taubman, 1976; Behrman et al., 1980; Ashenfelter and Krueger, 1994; Ashenfelter and Zimmerman, 1997), then the effect of measurement error is greatly exacerbated. This is because schooling levels are highly correlated between twins, while measurement error is

Table 9
Precision of selected variables

	Study	Variable	Dataset	Reliability ratio
1.	Duncan and Hill (1985)	Log earnings 1982	PSID-Validation Study	0.76
		Log earnings 1981		0.71
		Δ Log annual earnings		0.61
2.	Bound and Krueger (1991)	Log annual earnings men	CPS-SER	0.82
		Δ Log annual earnings men		0.65
		Log annual earnings women		0.92
		Δ Log annual earnings women		0.81
3.	Bound et al. (1994)	Log 1986 annual earnings	PSID-Validation Study	0.70
		Log 1982 annual earnings		0.85
		4-year Δ log annual earnings		0.71
		Log 1986 annual hours		0.63
		Log 1982 annual hours		0.72
4.	Siegel and Hodge (1968)	Education	1960 Census Post Enumeration Survey	0.93
5.	Bielby et al. (1977)	Education	1970 Census Post Enumeration Survey	0.80
6.	Ashenfelter and Krueger (1994)	Education	Twinsburg Twins Study	0.90
7.	Behrman et al. (1994)	Education	NAS-NRC Twins Sample	0.92
8.	Behrman et al. (1996)	Education	Minnesota Twin Registry	0.94

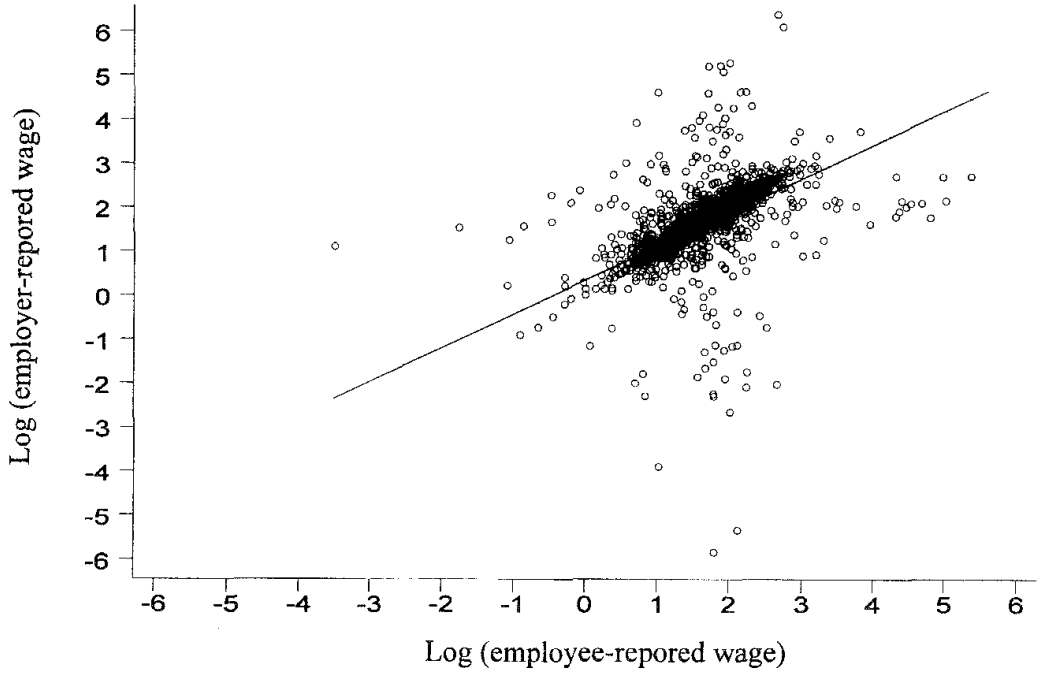

Fig. 6. Scatter plot of employer versus employee-reported log wages, with regression line. Data are from Mellow and Sider (1983).

magnified because reporting errors appear to be uncorrelated between twins. This situation is analogous to the effect of measurement error in panel data models discussed above.

To further explore the extent of measurement error in labor data, we re-analyzed the CPS data originally used by Mellow and Sider (1983). Fig. 6 presents a scatter diagram of the employer-reported log hourly wage against the employee-reported log hourly wage.[61] Although most points cluster around the 45 degree line, there are clearly some outliers. Some of the large outliers probably result from random coding errors, such as a misplaced decimal point.

Researchers have employed a variety of "trimming" techniques to try to minimize the effects of observations that may have been misreported. An interesting study of historical data by Stigler (1977) asks whether statistical methods that downweight outliers would have reduced the bias in estimates of physical constants in 20 early scientific datasets. These constants, such as the speed of light or parallax of the sun, have since been determined with certainty. Of the 11 estimators that he evaluated, Stigler found that the unadjusted sample mean, or a 10% "winsorized mean," provided estimates that were closest to the correct parameters. The 10% winsorized mean sets the values of observations in the

[61] Earnings in the data analyzed by Mellow and Sider were calculated in a manner similar to that used in the redesigned CPS. First, households and firms were asked for the basis on which the employee was paid, and then earnings were collected on that basis. Usual weekly hours were also collected. The household data may have been reported by the worker or by a proxy respondent.

Table 10
Alternative treatment of outliers in Mellow and Sider's matched employee-employer CPS sample[a]

	Mean employee minus employer	r	β	Employee variance	Employer variance
A. Unadjusted data					
ln wage	0.017	0.65	0.77	0.305	0.427
ln hours	−0.043	0.78	0.87	0.147	0.181
B. Employee data winsorized or truncated					
1% winsorized sample					
ln wage	0.021	0.68	0.88	0.258	0.427
ln hours	−0.044	0.77	0.91	0.131	0.181
10% winsorized sample					
ln wage	0.034	0.68	1.04	0.183	0.427
ln hours	−0.069	0.72	1.28	0.057	0.181
1% truncated sample					
ln wage	0.023	0.68	0.91	0.232	0.413
ln hours	−0.041	0.75	0.87	0.117	0.155
10% truncated sample					
ln wage	0.021	0.60	0.94	0.126	0.307
ln hours	−0.030	0.62	0.96	0.031	0.072
C. Both employee and employer data winsorized or truncated					
1% winsorized sample					
ln wage	0.025	0.8	0.86	0.258	0.303
ln hours	−0.04	0.78	0.85	0.131	0.153
10% winsorized sample					
ln wage	0.028	0.88	0.92	0.183	0.198
ln hours	−0.024	0.84	0.85	0.057	0.059
1% truncated sample					
ln wage	0.032	0.88	0.92	0.230	0.250
ln hours	− 0.036	0.76	0.81	0.109	0.125
10% truncated sample					
ln wage	0.024	0.91	0.94	0.119	0.125
ln hours	− 0.012	0.8	0.83	0.027	0.028

[a] Notes: r is the correlation coefficient between the employee- and employer-reported values. β is the slope coefficient from a regression of the employer-reported value on the employee-reported value. Sample size is 3856 for unadjusted wage data and 3974 for unadjusted hours data. In the 1% winsorized sample, the bottom and top 1% of observations were rolled back to the value corresponding to the 1st or 99th percentile; in the truncated sample these observations were deleted from the sample.

bottom or top decile equal to the value of the observation at the 10th or 90th percentile, and simply calculates the mean for this "adjusted" sample.

In a similar vein, we used Mellow and Sider's linked employer-employee CPS data to explore the effect of various methods for trimming outliers. The analysis here is less clear cut than in Stigler's paper because the true values are not known (i.e., we are not sure the employer-reported data are the "true" data), but we can still compare the reliability of the employee and employer reported data using various trimming methods. The first column of Table 10 reports the difference in mean earnings between the employee and employer responses for the wage and hours data. The differences are small and statistically insignificant. Column 2 reports the correlation between the employee report and the employer report, while column 3 reports the slope coefficient from a bivariate regression of the employer report on the employee report. The regression coefficient in column 3 probably provides the most robust measure of the reliability of the data. Columns 4 and 5 report the variances of the employee and employer data. Results in Panel A are based on the full sample without any trimming. Panel B presents results for a 1% and a 10% "winsorized" sample. We also report results for a 1% and 10% truncated sample. Whereas the winsorized sample rolls back extreme values (defined as the bottom or top X%) but retains them in the sample, the truncated sample simply drops the extreme observations from the sample.[62] In Panel B only the employee-reported data have been trimmed, because that is all that researchers typically observe. In Panel C, we trim both the employee- and employer-reported data.

For hours, the unadjusted data have reliability ratios around 0.80. Interestingly, the reliability of the hours data is considerably higher in Mellow and Sider's data than in the PSID validation study. This may result because the PSID validation study was confined to one plant (which restricted true hours variability compared to the entire workforce), or because there is a difference between the reliability of log weekly hours and annual hours.

The reliability ratio is lower for the wage data than the hours data in the CPS sample. For hours and wages, the correlation coefficients change little when the samples are adjusted (either by winsorizing or truncating the sample), but the slope coefficients are considerably larger in the adjusted data and exceed 1.0 in the 10% winsorized samples. When both the employer and employee data are trimmed, the reliability of the wage data improves considerably, while the reliability of the hours data is not much affected. These results suggest that extreme wage values are likely to be mistakes. Overall, this brief exploration suggests that a small amount of trimming could be beneficial. In a study of the effect of UI benefits on consumption, Gruber (1997) recommends winsorizing the extreme 1% of observations on the dependent variable (consumption), to reduce residual variability. A similar practice seems justifiable for earnings as well.

[62] Loosely speaking, winsorizing the data is desirable if the extreme values are exaggerated versions of the true values, but the true values still lie in the tails. Truncating the sample is more desirable if the extremes are mistakes that bear no resemblance to the true values.

Table 11
Estimates of reliability ratios from Mellow and Sider's CPS dataset[a]

Variable	r	Bivariate β	Multivariate β
ln wage unadjusted	0.65	0.77	0.66
ln wage 1% truncated[b]	0.68	0.91	0.85
ln wage 1% winsorized[b]	0.68	0.88	0.79
ln hours unadjusted	0.78	0.87	0.86
ln hours 1% truncated[b]	0.75	0.87	0.85
ln hours 1% winsorized[b]	0.77	0.91	0.90
Union	0.84	0.84	0.84
2-digit industry premium	0.93	0.93	0.92
1-digit industry premium	0.91	0.92	0.90
1-digit occupation premium	0.84	0.84	0.75

[a] Notes: r is the correlation coefficient between the employee- and employer-reported values. β is the coefficient from a regression of the employer-reported value on the employee-reported value. In the multiple regression, covariates include: highest grade of school completed, high school diploma; college diploma dummy, marital status, non-white, female, potential work experience, potential work experience squared, and veteran status. Sample size varies from 3806 (for industry) to 4087 (for occupation).

[b] Only the employee data were truncated or winsorized.

The estimates in Table 9 or 10 could be used to "inflate" regression coefficients for the effect of measurement error bias, provided that there are no covariates in the equation. Typically, however, regressions include covariates. Consequently, in Table 11 we use Mellow and Sider's CPS sample to regress the employer-reported data on the employee-reported data *and* several commonly used covariates (education, marital status, race, sex, experience and veteran status). For comparison, the first two columns present the correlation coefficient and the slope coefficient from a bivariate regression of the employer on the employee data. The third column reports the coefficient on the employee-reported variable from a multiple regression which specifies the employer-reported variable as the dependent variable, and the corresponding employee-reported variable as an explanatory variable along with other commonly used explanatory variables; this column provides the appropriate estimates of attenuation bias for a multiple regression which includes the same set of explanatory variables as included in the table. Notice that the reliability of the wage data falls from 0.77 to 0.66 once standard human capital controls are included. By contrast, the reliability of the hours data is not very much affected by the presence of control variables, because hours are only weakly correlated with the controls.

Table 11 also reports estimates of the reliability of reported union coverage status, industry and occupation. Assuming the employer-reported data are correct, the bivariate

regression suggests that union status has a reliability ratio of 0.84.[63] Interestingly, this is unchanged when covariates are included. To convert the industry and occupation dummy variables into a one-dimensional variable, we assigned each industry and occupation the wage premium associated with employment in that sector based on Krueger and Summers (1987). The occupation data seem especially noisy, with an estimated reliability ratio of.75 conditional on the covariates.

Earlier we mentioned that classical measurement error has a greater effect if variables are expressed as changes. Although we cannot examine longitudinal changes with Mellow and Sider's data, a dramatic illustration of the effect of measurement error on industry and occupation changes is provided by the 1994 CPS redesign. The redesigned CPS prompts respondents who were interviewed the previous month with the name of the employer that they reported working for the previous month, and then asks whether they still work for that employer. If respondents answer "no," they are asked an independent set of industry and occupation questions. If they answer "yes," they are asked if the usual activities and duties on their job changed since last month. If they report that their activities and duties were unchanged, they are then asked to verify the previous month's description of their occupation and activities. Lastly, if they answer that their activities and duties changed, they are asked an independent set of questions on occupation, activities, and class of worker. Based on pre-tests of the redesigned CPS in 1991, Rothgeb and Cohany (1992) find that the proportion of workers who appear to change three-digit occupations from one month to the next falls from 39% in the old version of the CPS to 7% in the redesigned version.[64] The proportion who change three-digit industry between adjacent months falls from 23% to 5%. These large changes in the gross industry and occupation flows obviously change one's impression of the labor market.[65]

[63] Union status is a dummy variable, so measurement errors will be correlated with true union status. But if union status is correctly reported by employers, the regression coefficient in Table 11 nonetheless provides a consistent estimate of the attenuation bias. Additionally, note that the reliability of data on union status depends on the true fraction of workers who are covered by a union contract. Since union coverage as a fraction of the workforce has declined over time, the reliability ratio might be even lower today. As an extreme example, note that even if the true union coverage rate falls to zero, the measured rate will exceed zero because some (probably around 3%) non-union workers will be erroneously classified as covered by a union. See Freeman (1984), Jakubson (1986) and Card (1996) for analyses of the effect of measurement error in union status in longitudinal data.

[64] It is also possible that dependent interviewing reduces occupational changes because some respondents find it easier to complete the interview by reporting that they did not change employers even if they did. Although this is possible, Rothgeb and Cohany point out that asking independent occupation and industry questions of individuals who report changing employers could result in spurious industry and occupation changes. In addition, the large number of mismatches between employer and employee reported occupation and industry data in Mellow and Sider's dataset are consistent with a finding of grossly overestimated industry and occupation flows.

[65] See also Poterba and Summers (1986), who estimate the measurement error in employment-status transitions.

4.3. Weighting and allocated values

Many datasets use complicated sampling designs and come with sampling weights that reflect the design. Researchers are often confronted with the question of whether to employ sample weights in their statistical analyses to adjust for non-random sampling. For example, if the sampling design uses stratified sampling by state, with smaller states sampled at a higher rate than larger states, then observations from small states should get less weight if national statistics are to be representative. In addition to providing sample weights for this purpose, the Census Bureau also "allocates" answers for individuals who do not respond to a question in one of their surveys. Missing data are allocated by inserting information for a randomly chosen person who is matched to the person with missing data on the basis of major demographic characteristics. Consequently, there are no "missing values" on Census Bureau micro data files. But researchers may decide to include or exclude observations with allocated responses since information that has been allocated is identified with "allocation flags." Unfortunately, although there is a large literature on weighting and survey non-response, this literature has not produced any easy answers that apply to all datasets and research questions (see, e.g., Rubin, 1983; Dickens, 1985; Lillard et al., 1986; Deaton, 1995, 1997; Groves, 1998).[66]

Two datasets where both weighting and allocation issues come up are the CPS and the 1990 Census Public Use Micro Sample (PUMS), neither of which is a simple random sample. The CPS uses a complicated multi-stage probability sample that over-samples some states, and recently oversamples Hispanics in the March survey (see, e.g., US Bureau of the Census, 1992). The 1990 PUMS also deviates from random sampling because of over-sampling of small areas and Native Americans (US Bureau of the Census, 1996).[67] And even random samples may fail to be representative by chance, or because some sampled households are not actually interviewed. The sampling weights including with CPS and PUMS micro data are meant to correct for these features of the sample design, as well as deviations from random sampling due to chance or non-response that affect the age, Sex, Hispanic origin, or race make-up of the sample. Missing data for respondents in these datasets are also allocated. And in the CPS, if someone fails to answer a monthly supplement (e.g., the March income supplement), then entire record is allocated by drawing a randomly matched "donor record" from someone who did respond.

To assess the consequences of weighting and allocation for one important area of research, we estimated a standard human capital earnings function with data from the 1990 March CPS and 1990 5% PUMS for the four permutations of weighting or not weighting, and including or excluding observations with allocated responses. The samples

[66] But see DuMouchel and Duncan (1983), who note that if the object of regression is a MMSE linear approximation to the CEF then estimates from non-random samples should be weighted.

[67] The 1980 PUMS are simple random samples. The CPS was stratified but self-weighting (i.e. all observations were equally likely to be sampled) until January 1978.

Table 12
Weighting and allocation in the Census and CPS[a]

Covariate	1990 Census				March 1990 CPS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log wages mean	6.405	6.415	6.425	6.437	6.34	6.348	6.351	6.357
Standard deviation	0.746	0.747	0.723	0.721	0.732	0.734	0.717	0.723
Education	0.10932	0.10828	0.10920	0.10813	0.10839	0.11139	0.10950	0.11314
	(0.00047)	(0.00047)	(0.00049)	(0.00049)	(0.00442)	(0.00438)	(0.00459)	(0.00459)
White	0.208	0.213	0.199	0.202	0.194	0.219	0.196	0.211
	(0.003)	(0.003)	(0.004)	(0.003)	(0.030)	(0.027)	(0.031)	(0.029)
Married	0.386	0.387	0.381	0.382	0.386	0.387	0.343	0.362
	(0.004)	(0.003)	(0.004)	(0.004)	(0.031)	(0.029)	(0.032)	(0.031)
Widowed	0.181	0.165	0.190	0.171	0.110	0.200	0.077	0.075
	(0.013)	(0.013)	(0.014)	(0.014)	(0.108)	(0.105)	(0.117)	(0.115)
Divorced or separated	0.193	0.187	0.202	0.196	0.167	0.135	0.141	0.123
	(0.004)	(0.004)	(0.005)	(0.004)	(0.037)	(0.035)	(0.039)	(0.037)
Hispanic	−0.142	−0.151	−0.138	−0.145	−0.125	−0.179	−0.107	−0.155
	(0.005)	(0.005)	(0.005)	(0.005)	(0.040)	(0.048)	(0.041)	(0.049)
Veteran	−0.012	−0.014	−0.018	−0.021	−0.0001	−0.012	−0.002	−0.015
	(0.002)	(0.002)	(0.002)	(0.002)	(0.016)	(0.016)	(0.017)	(0.017)
Potential experience	0.040	0.041	0.041	0.042	0.0005	−0.002	0.013	0.013
	(0.002)	(0.002)	(0.002)	(0.002)	(0.021)	(0.022)	(0.022)	(0.023)
Pot. experience squared*100	−0.055	−0.055	−0.057	−0.057	0.024	0.035	0.003	0.008
	(0.004)	(0.004)	(0.005)	(0.005)	(0.043)	(0.043)	(0.045)	(0.045)
Allocated	Yes	Yes	No	No	Yes	Yes	No	No
Weighted	No	Yes	No	Yes	No	Yes	No	Yes
N	603763	603731	527095	527071	7134	7134	6361	6361

[a] Notes: The table reports OLS estimates of wage equations with the indicated covariates. Standard errors are reported in parentheses. The samples include black and white men aged 40–49 with at least 8 years of schooling. The Census sample excludes active-duty military personnel and the CPS sample excludes military personnel and the Hispanic over-sample. The CPS schooling variable is highest year completed while the census variable is imputed as described in the appendix.

consist of white and black men age 40–49 with at least 8 years of education.[68] Regression results and mean log weekly earnings are summarized in Table 12. In both datasets, the estimated regression coefficients are remarkably similar regardless of whether the equation is estimated by OLS or weighted least squares to adjust for sample weights, and regardless of whether the observations with allocated values are excluded or included in the sample. Moreover, except for potential experience, the regression coefficients are quite similar if they are estimated with either the Census or CPS sample. One notable difference between the datasets, however, is that mean log earnings are about 6 points higher in the Census than the CPS for this age group.

The results in Table 12 suggest that estimates of a human capital earnings function using CPS and Census data are largely insensitive to whether or not the sample is weighted to account for the sample design, and whether or not observations with allocated values are included in the sample. At least for this application, non-random sampling and the allocation of missing values are not very important.[69] It should be noted, however, that the Census Bureau surveys analyzed here are relatively close to random samples, and that the sample strata involve covariates that are included in the regression models. Some of the datasets discussed earlier, most notably the NLSY and the PSID, include large non-random sub-samples that more extensively select or over-sample certain groups using a wider range of characteristics, including racial minorities, low-income respondents, or military personnel. When working with these data is it important to check whether the use of a non-representative sample affects empirical results. Moreover, since researchers often compare results across samples, weighting may be desirable to reduce the likelihood that differences in sample design generate different results.

5. Summary

This chapter attempts to provide an overview of the empirical strategies used in modern labor economics. The first step is to specify a causal question, which we think of as comparing actual and counterfactual states. The next step is to devise a strategy that can, in principle, answer the question. A critical issue in this context is how the causal effect of interest is identified by the statistical analysis. In particular, why does the explanatory variable of interest vary when other variables are held constant? Who is implicitly being compared to whom? Does the source of variation used to identify the key parameters provide plausible "counterfactuals"? And can the identification strategy be tested in a situation in which the causal variable is not expected to have an effect? Finally, imple-

[68] In addition, to make the samples comparable, the Census sample excludes men who were on active duty in the military, and the CPS sample excludes the Hispanic oversample and men in the armed forces. The education variable in both datasets was converted to linear years of schooling based on highest degree attained.

[69] Of course, the standard errors of the estimates should reflect the sample design and account for changes in variability due to allocation. But for samples of this size, the standard errors are extraordinarily small, so adjusting them for these features of the data is probably of second-order importance.

mentation of the empirical strategy requires appropriate data, and careful attention to the many measurement problems that are likely to arise along the way.

Appendix A

A.1. Derivation of Eq. (9) in the text

The model is

$$Y_i = \beta_0 + \rho S_i + \eta_i, \qquad \mathrm{E}[S_i \eta_i] = 0,$$

$$A_i = \gamma_0 + \gamma_1 S_i + \eta_{1i}, \qquad \mathrm{E}[S_i \eta_{1i}] = 0.$$

The coefficient on S_i in a regression of Y_i on S_i and A_i is $C(Y_i, S_{\cdot Ai})/V(S_{\cdot Ai})$ where

$$S_{\cdot Ai} = S_i - \pi_0 - \pi_1 A_i \qquad \text{and} \qquad \pi_1 = \gamma_1 V(S_i)/V(A_i).$$

Also

$$V(S_{\cdot Ai}) = V(S_i) - \pi_1^2 V(A_i) = [V(S_i)/V(A_i)][V(A_i) - \gamma_1^2 V(S_i)] = [V(S_i)/V(A_i)]V(\eta_{1i}).$$

So

$$C(Y_i, S_{\cdot Ai})/V(S_{\cdot Ai}) = \rho + C(\eta_i, S_i - \pi_0 - \pi_1 A_i)/V(S_{\cdot Ai}) = \rho - \pi_1 C(\eta_i, A_i)/V(S_{\cdot Ai})$$

$$= \rho - \pi_1 C(\eta_i, \eta_{1i})/V(S_{\cdot Ai}) = \rho - \gamma_1 \varphi_{01}.$$

A.2. Derivation of Eq. (34) in the text

To economize on notation, we use $\mathrm{E}[Y \mid X, j]$ as shorthand for $\mathrm{E}[Y_i \mid X_i, S_i = j]$. Repeating Eq. (31) in the text without "i" subscripts:

$$\rho_{\mathrm{r}} = \mathrm{E}[Y(S - \mathrm{E}[S \mid X])]/\mathrm{E}[S(S - \mathrm{E}[S \mid X])]$$

$$= \mathrm{E}[E(Y \mid S, X)(S - \mathrm{E}[S \mid X])]/\mathrm{E}[S(S - \mathrm{E}[S \mid X])]. \tag{A.1}$$

Now write

$$\mathrm{E}[Y \mid X, S] = \mathrm{E}[Y \mid X, 0] + \sum_{j=1}^{S} \{\mathrm{E}[Y \mid X, j] - \mathrm{E}[Y \mid X, j-1]\} = \mathrm{E}[Y \mid X, S = 0] + \sum_{j=1}^{S} \rho_{jx}, \tag{A.2}$$

where

$$\rho_{jx} \equiv \mathrm{E}[Y \mid X, j] - \mathrm{E}[Y \mid X, j-1].$$

We first simplify the numerator of ρ_{r}. Substituting (A.2) into (A.1):

$$\mathrm{E}[E(Y \mid X, S)(S - \mathrm{E}[S \mid X])] = \mathrm{E}\left\{\left(\sum_{j=1}^{S} \rho_{jx}\right)(S - \mathrm{E}[S \mid X])\right\}$$

$$= \mathrm{E}\left\{\mathrm{E}\left[\sum_{j=1}^{S} \rho_{jx}(S - \mathrm{E}[S \mid X]) \mid X\right]\right\}.$$

Working with the inner expectation,

$$E\left[\sum_{j=1}^{S} \rho_{jx}(S - \mathrm{E}[S \mid X]) \mid X\right] = \sum_{s=1}^{\bar{s}} \sum_{j=1}^{s} \rho_{jx}(s - \mathrm{E}[S \mid X])P_{sx},$$

where

$$P_{sx} = P(S = s \mid X).$$

Reversing the order of summation, this equals

$$\sum_{j=1}^{\bar{s}} \rho_{jx}\left[\sum_{s=j}^{\bar{s}} (s - \mathrm{E}[S \mid X])P_{sx}\right] = \sum_{j=1}^{\bar{s}} \rho_{jx}\mu_{jx},$$

where

$$\mu_{jx} = \sum_{s=j}^{\bar{s}} (s - \mathrm{E}[S \mid X])P_{sx}.$$

Now, simplifying,

$$\mu_{jx} = \sum_{j=1}^{\bar{s}} sP_{sx} - \sum_{s=j}^{\bar{s}} \mathrm{E}[S \mid X])P_{sx} = (\mathrm{E}[S \mid X, S \geq j] - \mathrm{E}[S \mid X])P(S \geq j \mid X),$$

Since

$$\mathrm{E}[S \mid X] = \mathrm{E}[S \mid X, S \geq j]P(S \geq j \mid X) + \mathrm{E}[S \mid X, S < j](1 - P(S \geq j \mid X)),$$

$$\mu_{jx} = (\mathrm{E}[S \mid S \geq j, X] - \mathrm{E}[S \mid S < j, X])P(S \geq j \mid X)(1 - P(S \geq j \mid X)).$$

So we have shown

$$\mathrm{E}[Y_i(S_i - \mathrm{E}[S_i \mid X_i])] = E\left[\sum_{j=1}^{\bar{s}} \rho_{jx}\mu_{jx}\right].$$

A similar argument for the denominator shows

$$\mathrm{E}[S_i(S_i - \mathrm{E}[S_i \mid X_i])] = E\left[\sum_{j=1}^{\bar{s}} \mu_{jx}\right].$$

Substitute S for j in the summations to get Eq. (34) using the notation in the text.

A.3. Schooling in the 1990 Census

Years of schooling was coded from the 1990 Census categorical schooling variables as follows:

Years of schooling	Educational attainment
8	5th, 6th, 7th, or 8th grade
9	9th grade
10	10th grade
11	11th grade or 12th grade, no diploma
12	High school graduate, diploma or GED
13	Some college, but no degree
14	Completed associate degree in college, occupational program
15	Completed associate degree in college, academic program
16	Completed bachelor's degree, not attending school
17	Completed bachelor's degree, but now enrolled
18	Completed master's degree
19	Completed professional degree
20	Completed doctorate

References

Abadie, Alberto (1998), "Semiparametric estimation of instrumental variable models for causal effects", Mimeo. (Department of Economics, MIT).

Abowd, John M. and Henry S. Farber (1982), "Job queues and the union status of workers", Industrial and Labor Relations Review 35: 354–367.

Aigner, Dennis J. (1973), "Regression with a binary independent variable subject to errors of observation", Journal of Econometrics 1 (1): 49–59.

Altonji, Joseph G. (1986), "Intertemporal substitution in labor supply: evidence from micro data", Journal of Political Economy 94 (3): S176–S215.

Altonji, Joseph G. and Lewis M. Segal (1996), "Small-sample bias in GMM estimation of covariance structures", Journal of Business and Economic Statistics 14 (3): 353–366.

Anderson, Patricia M. (1993), "Linear adjustment costs and seasonal labor demand: evidence from retail trade firms", Quarterly Journal of Economics 108 (4): 1015–1042.

Anderson, Patricia M. and Bruce D. Meyer (1994), "The extent and consequences of job turnover", Brookings Papers on Economic Activity: Microeconomics: 177–236.

Anderson, T.W., Naoto Kunitomo and Takamitsu Sawa (1982), "Evaluation of the distribution function of the limited information maximum likelihood estimator", Econometrica 50: 1009–1027.

Angrist, Joshua D. (1990), "Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records", American Economic Review 80: 313–335.

Angrist, Joshua D. (1995a), "Introduction to the JBES symposium on program and policy evaluation", Journal of Business and Economic Statistics 13 (2): 133–136.

Angrist, Joshua D. (1995b), "The economic returns to schooling in the West Bank and Gaza strip", American Economic Review 85 (5): 1065–1087.

Angrist, Joshua D. (1998), "Estimating the labor market impact of voluntary military service using social security data on military applicants", Econometrica 66 (2): 249–288.

Angrist, Joshua D. and William N. Evans (1998), "Children and their parents' labor supply: evidence from exogenous variation in family size", American Economic Review, in press.

Angrist, Joshua D. and Guido W. Imbens (1991), "Sources of identifying information in evaluation models", Technical working paper no. 117 (NBER, Cambridge, MA).

Angrist, Joshua D. and Guido W. Imbens (1995), "Two-stage least squares estimates of average causal effects in models with variable treatment intensity", Journal of the American Statistical Association 90 (430): 431–442.

Angrist, Joshua D. and Alan B. Krueger (1991), "Does compulsory school attendance affect schooling and earnings?" Quarterly Journal of Economics 106: 979–1014.

Angrist, Joshua D. and Alan B. Krueger (1992), "The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples", Journal of the American Statistical Association 87 (418): 328–336.

Angrist, Joshua D. and Alan B. Krueger (1995), "Split-sample instrumental variables estimates of the returns to schooling", Journal of Business and Economic Statistics 13 (2): 225–235.

Angrist, Joshua D. and Victor Lavy (1998), "Using maimonides rule to estimate the effects of class size on scholastic achievement", Quarterly Journal of Economics, in press.

Angrist, Joshua D. and Whitney K. Newey (1991), "Over-identification tests in earnings functions with fixed effects", Journal of Business and Economic Statistics 9 (3): 317–323.

Angrist, Joshua D., Guido W. Imbens and Kathryn Graddy (1995), "Non-parametric demand analysis with an application to the demand for fish", Technical working paper no. 178 (NBER, Cambridge, MA).

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin (1996), "Identification of causal effects using instrumental variables", Journal of the American Statistical Association 91 (434): 444–455.

Angrist, Joshua D., Guido W. Imbens and Alan B. Krueger (1998), "Jackknife instrumental variables estimation", Journal of Applied Econometrics, in press.

Arellano, Manuel and Costas Meghir (1992), "Female labour supply and on-the-job search: an empirical model estimated using complementary datasets", Review of Economic Studies 59 (3): 537–559.

Ashenfelter, Orley A. (1978), "Estimating the effect of training programs on earnings", Review of Economics and Statistics 60 (1): 47–57.

Ashenfelter, Orley A. (1984), "Macroeconomic analyses and microeconomic analyses of labor supply", Carnegie-Rochester Series on Public Policy 21: 117–155.

Ashenfelter, Orley A. and David E. Card (1985), "Using the longitudinal structure of earnings to estimate the effect of training programs", Review of Economics and Statistics 67 (4): 648–660.

Ashenfelter, Orley A. and Alan B. Krueger (1994), "Estimates of the economic return to schooling from a new sample of twins", American Economic Review 84 (5): 1157–1173.

Ashenfelter, Orley A. and Joseph D. Mooney (1968), "Graduate education, ability and earnings", Review of Economics and Statistics 50 (1): 78–86.

Ashenfelter, Orley A. and David J. Zimmerman (1997), "Estimates of the returns to schooling from sibling data: fathers, sons and brothers", Review of Economics and Statistics 79 (1): 1–9.

Baker, George, Michael Gibbs and Bengt Holmstrom (1994), "The internal economics of the firm: evidence from personnel data", Quarterly Journal of Economics 109 (4): 881–919.

Barnow, Burt S., Glen G. Cain and Arthur Goldberger (1981), "Selection on observables", Evaluation Studies Review Annual 5: 43–59.

Becketti, Sean, William Gould, Lee Lillard and Finis Welch (1988), "The panel study of income dynamics after fourteen years: an evaluation", Journal of Labor Economics 6 (4): 472–492.

Behrman, Jere, Zdenek Hrubec, Paul Taubman and Terence Wales (1980), Socioeconomic success: a study of the effects of genetic endowments, family environment and schooling (North-Holland, Amsterdam).

Behrman, Jere R., Mark R. Rosenzweig and Paul Taubman (1994), "Endowments and the allocation of schooling in the family and in the marriage market: the twins experiment", Journal of Political Economy 102 (6): 1131–1174.

Behrman, Jere R., Mark R. Rosenzweig and Paul Taubman (1996), "College choice and wages: estimates using data on female twins", Review of Economics and Statistics 78 (4): 672–685.

Bekker, Paul A. (1994), "Alternative approximations to the distributions of instrumental variables estimators", Econometrica 62 (3): 657–681.

Bielby, William, Robert Hauser and David Featherman (1977), "Response errors of non-black males in models of the stratification process", in: D.J. Aigner and A.S. Goldberger, eds., Latent variables in socioeconomic models (North-Holland, Amsterdam) pp. 227–251.

Björklund, Anders and Robert Moffitt (1987), "The estimation of wage gains and welfare gains in self-selection models", The Review of Economics and Statistics 69 (1): 42–49.

Borjas, George J. (1980), "The relationship between wages and weekly hours of work: the role of division bias", Journal of Human Resources 15 (3): 409–423.

Borjas, George J., Richard B. Freeman and Kevin Lang (1991), "Undocumented Mexican-born workers in the United States: how many, how permanent?" in: John M. Abowd and Richard B. Freeman, eds., Immigration, trade and the labor market (National Bureau of Economic Research Project Report, University of Chicago Press, Chicago, IL).

Borjas, George J., Richard B. Freeman and Lawrence F. Katz (1997), " How much do immigration and trade affect labor market outcomes?" Brookings Papers on Economic Activity 10 (1): 1–67.

Bound, John (1989), "The health and earnings of rejected disability insurance applicants", American Economic Review 79 (3): 482–503.

Bound, John (1991), "The health and earnings of rejected disability insurance applicants: reply", American Economic Review 81 (5): 1427–1434.

Bound, John and Alan B. Krueger (1991), "The extent of measurement error in longitudinal earnings data: do two wrongs make a right?" Journal of Labor Economics 9 (1): 1–24.

Bound, John, et al. (1994), "Evidence on the validity of cross-sectional and longitudinal labor market data", Journal of Labor Economics 12 (3): 345–368.

Bound, John, David Jaeger and Regina Baker (1995), "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak", Journal of the American Statistical Association 90 (430): 443–450.

Bowen, William G. and Derek Bok (1998), The shape of the river: long-term consequences of considering race in college and university admissions (Princeton University Press, Princeton, NJ).

Bronars, Stephen G. and Jeff Grogger (1994), "The economic consequences of unwed motherhood: using twins as a natural experiment", American Economic Review 84 (5): 1141–1156.

Brown, Charles, Greg J. Duncan and Frank P. Stafford (1996), "Data watch: the panel study of income dynamics", Journal of Economic Perspectives 10 (2): 155–168.

Burtless, Gary (1995), "The case for randomized field trials in economic and policy research", Journal of Economic Perspectives 9 (2): 63–84.

Buse, A. (1992), "The bias of instrumental variable estimators", Econometrica 60 (1): 173–180.

Campbell, Donald T. (1969), "Reforms as experiments", American Psychologist XXIV: 409–429.

Campbell, Donald T. and J.C. Stanley (1963), Experimental and quasi-experimental designs for research (Rand-McNally, Chicago. IL).

Card, David E. (1989), "The impact of the Mariel boatlift on the Miami labor market", Working paper no. 253 (Industrial Relations Section, Princeton University).
Card, David E. (1990), "The impact of the Mariel boatlift on the Miami labor market", Industrial and Labor Relations Review 43: 245–257.
Card, David E. (1995), "Earnings, schooling and ability revisited", in: Solomon W. Polachek, ed., Research in labor economics (JAI Press, Greenwich, CT).
Card, David E. (1996), "The effect of unions on the structure of wages: a longitudinal analysis", Econometrica 64 (4): 957–979.
Card, David E. and Alan B. Krueger (1994), "Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania", American Economic Review 84 (4): 772–784.
Card, David E. and Alan B. Krueger (1998), "A reanalysis of the effect of the New Jersey minimum wage increase on the fast-food industry with representative payroll data", Working paper no. 6386 (NBER, Cambridge, MA).
Card, David E. and Daniel Sullivan (1988), "Measuring the effect of subsidized training on movements in and out of employment", Econometrica 56 (3): 497–530.
Center for Drug Evaluation and Research (1988), Guideline for the format and content of the clinical and statistical sections of a new drug application (US Food and Drug Administration, Department of Health and Human Services, Washington, DC).
Chamberlain, Gary (1977), "Education, income and ability revisited", Journal of Econometrics 5 (2): 241–257.
Chamberlain, Gary (1978), "Omitted variables bias in panel data: estimating the returns to schooling", Annales de l'INSEE 30–31: 49–82.
Chamberlain, Gary (1980), "Discussion", American Economic Review 70 (2): 47–49.
Chamberlain, Gary (1984), "Panel data", in: Zvi Griliches and Michael D. Intriligator, eds., Handbook of econometrics (North-Holland, Amsterdam).
Chamberlain, Gary and Guido W. Imbens (1996), "Hierarchical Bayes models with many instrumental variables", Discussion paper no. 1781 (Department of Economics, Harvard University).
Chamberlain, Gary and Edward E. Leamer (1976), "Matrix weighted averages and posterior bounds", Journal of the Royal Statistical Society, Series B 38: 73–84.
Chay, Kenneth Y. (1996), "An empirical analysis of black economic progress over time" PhD Thesis (Department of Economics, Princeton University).
Cochran, William G. (1965), "The planning of observational studies of human populations (with discussion)", Journal of the Royal Statistical Society, Series A 128: 234–266.
Coder, John and Lydia Scoon-Rogers (1996), "Evaluating the quality of income data collected in the annual supplement to the March Current Population Survey and the Survey of Income and Program Participation", Census working paper no. 215 (US Bureau of the Census, Washington, DC).
Dawid, A.P. (1979), "Conditional independence in statistical theory", Journal of the Royal Statistical Society, Series B 41: 1–31.
Deaton, Angus (1985), "Panel data from a time series of cross-sections", Journal of Econometrics 30: 109–126.
Deaton, Angus (1995), "Data and econometric tools for development analysis", in: Hollis Chenery and T.N. Srinivasan, eds., Handbook of development economics (North-Holland, Amsterdam).
Deaton, Angus (1997), The analysis of household surveys: a microeconometric approach to development policy (Johns Hopkins University Press, Baltimore, MD).
Deaton, Angus and Christina Paxson (1998), "Economies of scale, household size and the demand for food", Journal of Political Economy, in press.
Dehejia, Rajeev H. and Sadek Wahba (1995), "Causal effects in nonexperimental studies: re-evaluating the evaluation of training programs", Mimeo. (Department of Economics, Harvard University).
Dickens, William T. (1985), "Error components in grouped data: why it's never worth weighting", Technical working paper no. 43 (NBER, Cambridge, MA).
Dominitz, Jeff and Charles F. Manski (1997), "Using expectations data to study subjective income expectations", Journal of the American Statistical Association 92: 855–867.

Donald, Steven and Whitney K. Newey (1997), "Choosing the number of instruments", Mimeo. (Department of Economics, MIT).

DuMouchel, William H. and Greg Duncan (1983), "Using sample survey weights in multiple regression analyses of stratified samples", Journal of the American Statistical Association 78, 535–543.

Duncan, Greg J. and Daniel H. Hill (1985), "An investigation of the extent and consequences of measurement error in labor-economic survey data", Journal of Labor Economics 3 (4): 508–532.

Duncan, Greg T. and Robert W. Pearson (1991), "Enhancing access to microdata while protecting confidentiality: prospects for the future", Statistical Science 6 (3): 219–239.

Durbin, J. (1953), "A note on regression when there is extraneous information about one of the coefficients", Journal of the American Statistical Association 48: 799–808.

Durbin, J. (1954), "Errors in variables", Review of the International Statistical Institute 22: 23–32.

Farber, Henry S. and Alan B. Krueger (1993), "Union membership in the United States: the decline continues", Working paper no. 306 (Industrial Relations Section, Princeton University).

Fitzgerald, John, Peter Gottschalk and Robert Moffit (1998), "An analysis of sample attrition in panel data: the Michigan panel study of income dynamics", Journal of Human Resources, in press.

Freeman, Richard B. (1984), "Longitudinal analyses of the effects of trade unions", Journal of Labor Economics 2: 1–26.

Freeman, Richard B. (1989), Labor markets in action (Harvard University Press, Cambridge, MA).

Freeman, Richard B. (1990), "Employment and earnings of disadvantaged young men in a labor shortage economy", Working paper no. 3444 (NBER, Cambridge, MA).

Freeman, Richard B. and Brian Hall (1986), "Permanent homelessness in America?", Working paper no. 2013 (NBER, Cambridge, MA).

Freeman, Richard B. and Harry J. Holzer (1986), "The black youth employment crisis: summary of findings", in: Richard B. Freeman and Harry J. Holzer, eds., The black youth employment crisis (National Bureau of Economic Research Project Report, University of Chicago Press, Chicago, IL).

Freeman, Richard B. and Morris M. Kleiner (1990), "The impact of new unionization on wages and working conditions", Journal of Labor Economics 8 (1): S8–S25.

Friedberg, Rachel M. and Jennifer Hunt (1995), "The impact of immigrants on host country wages, employment and growth", Journal of Economic Perspectives 9 (2): 23–44.

Friedman, Milton (1957), A theory of the consumption function (Princeton University Press, Princeton, NJ).

Fuchs, Victor, Alan B. Krueger and James M. Poterba (1998), "Why do economists disagree about policy? The roles of beliefs about parameters and values", Journal of Economic Perspectives, in press.

Fuller, Wayne A. (1987), Measurement error models (Wiley, New York).

Girshick, M.A. and Trygve Haavelmo (1947), "Statistical analysis of the demand for food: examples of simultaneous estimation of structural equations", Econometrica 15 (2): 79–110.

Goldberger, Arthur S. (1972), "Selection bias in evaluating treatment effects: some formal illustrations", Discussion paper (Institute for Research on Poverty, University of Wisconsin) pp. 123–172.

Goldberger, Arthur S. (1991), A course in econometrics (Harvard University Press, Cambridge, MA).

Gorseline, Donald E. (1932), The effect of schooling upon income (University of Indiana, Bloomington, IN).

Griliches, Zvi (1977), "Estimating the returns to schooling: some econometric problems", Econometrica 45 (1): 1–22.

Griliches, Zvi (1979), "Sibling models and data in economics: beginnings of a survey", Journal of Political Economy 87 (5): S37–S64.

Griliches, Zvi (1986), "Economic data issues", in: Zvi Griliches and Michael D. Intriligator, eds., Handbook of econometrics (North-Holland, Amsterdam).

Griliches, Zvi and Jerry A. Hausman (1986), "Errors in variables in panel data", Journal of Econometrics 31 (1): 93–118.

Griliches, Zvi and Jacques Mairesse (1998), "Production functions: the search for identification", in: Zvi Griliches, ed., Practicing econometrics: essays in method and application (Edward Elgar, Cheltenham, UK).

Griliches, Zvi and William M. Mason (1972), "Education, Income and Ability", Journal of Political Economy 80 (3): S74–S103.

Grosh, Margaret E. and Paul Glewwe (1996), "Household survey data from developing countries: progress and prospects", American Economic Review 86 (2): 15–19.

Grosh, Margaret E. and Paul Glewwe (1998), "Data watch: the World Bank's living standards measurement study household surveys", Journal of Economic Perspectives 12 (1): 187–196.

Groves, Robert M. (1989), Survey errors and survey costs (Wiley, New York).

Groves, Robert M. (1998), Non-response in household interview surveys (Wiley, New York).

Gruber, Jonathan (1997), "The consumption smoothing benefits of unemployment insurance", American Economic Review 87 (1): 192–205.

Hahn, Jinyong (1998), "On the role of the propensity score in the efficient estimation of average treatment effects", Econometrica 66: 315–332.

Hahn, Jinyong, Petra Todd and Wilbert van der Klaauw (1998), "Estimation of treatment effects with a quasi-experimental regression-discontinuity design: with application to evaluating the effect of federal antidiscrimination laws on minority employment in small U.S. firms", Mimeo. (Department of Economics, University of Pennsylvania).

Hall, Alastair R., Glenn D. Rudebusch and David W. Wilcox (1996), "Judging instrument relevance in instrumental variables estimation", International Economic Review 37 (2): 283–298.

Hansen, Lars Peter (1982), "Large sample properties of generalized method of moments estimators", Econometrica 50 (4): 1029–1054.

Hansen, W. Lee, Burton A. Weisbrod and William J. Scanlon (1970), "Schooling and earnings of low achievers", American Economic Review 60 (3): 409–418.

Hausman, Jerry A. and William E. Taylor (1981), "Panel data and unobservable individual effects", Econometrica 49 (6): 1377–1398.

Hearst, Norman, Thomas Newman and Steven Hulley (1986), "Delayed effects of the military draft on mortality: a randomized natural experiment", New England Journal of Medicine 314: 620–624.

Heckman, James J. (1978), "Dummy endogenous variables in a simultaneous equations system", Econometrica 46 (4): 931–959.

Heckman, James J. and V. Joseph Hotz (1989), "Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training", Journal of the American Statistical Association 84 (408): 862–874.

Heckman, James J. and Thomas E. MaCurdy (1986), "Labor econometrics", in: Orley Ashenfelter and Richard Layard, eds., Handbook of labor economics (North-Holland, Amsterdam).

Heckman, James J. and Brook S. Payner (1989), "Determining the impact of antidiscrimination policy on the economic status of blacks: a study of South Carolina", American Economic Review 79 (1): 138–177.

Heckman, James J. and Richard Robb, Jr. (1985), "Alternative methods for evaluating the impact of interventions", in: James J. Heckman and Burton Singer, eds., Longitudinal analysis of labor market data, Econometric society monographs series no. 10 (Cambridge University Press, Cambridge, MA).

Heckman, James J. and Jeffrey A. Smith (1995), "Assessing the case for social experiments", Journal of Economic Perspectives 9 (2): 85–110.

Heckman, James J., Hidehiko Ichimura and Petra E. Todd (1997), "Matching as an econometric evaluation estimator: evidence from evaluating a job training programme", Review of Economic Studies 64 (4): 605–654.

Heckman, James J., Lance Lochner and Christopher Taber (1998), "Tax policy and human-capital formation", American Economic Review 88 (2): 293–297.

Holland, Paul W. (1986), "Statistics and causal inference", Journal of the American Statistical Association 81: 945–970.

Hurd, Michael, et al. (1998), "Consumption and savings balances of the elderly: experimental evidence on survey response bias", in: D. Wise (ed.), Frontiers in the economics of aging (University of Chicago Press, Chicago, IL) pp. 353–387.

Imbens, Guido W. and Joshua D. Angrist (1994), "Identification and estimation of local average treatment effects", Econometrica 62 (2): 467–475.

Imbens, Guido W. and Tony Lancaster (1994), "Combining micro and macro data in microeconometric models", Review of Economic Studies 61 (4): 655–680.

Imbens, Guido W. and Wilbert van der Klaauw (1995), "The cost of conscription in the Netherlands", Journal of Business and Economic Statistics 13 (2): 207–215.

Imbens, Guido W., Donald B. Rubin and Bruce I. Sacerdote (1997), "Estimating income effects: evidence from a survey of lottery players", Mimeo. (Economics Department, UCLA).

Jacobson, Louis S., Robert J. Lalonde and Daniel G. Sullivan (1994), "Earnings losses of displaced workers", American Economic Review 83 (4): 685–709.

Jaeger, David (1993), "The new current population survey education variable: a recommendation", Research report no. 93-289 (Population Studies Center, University of Michigan).

Jakubson, George (1986), "Measurement error in binary explanatory variables in panel data models: why do cross section and panel estimates of the union wage effect differ?" Working paper no. 209 (Industrial Relations Section, Princeton University).

Jakubson, George (1991), "Estimation and testing of the union wage effect using panel data", Review of Economic Studies 58 (5): 971–991.

Jappelli, Tuillio, Jorn-Steffen Pischke and Nicholas Souleles (1998), "Testing for liquidity constraints in euler equations with complementary data sources", Review of Economics and Statistics 80, 251–262.

Juhn, Chinhui, Kevin M. Murphy and Brooks Pierce (1993), "Wage inequality and the rise in returns to skill", Journal of Political Economy 101 (3): 410–442.

Juster, F. Thomas and James P. Smith (1997), "Improving the quality of economic data: lessons from the HRS and AHEAD", Journal of the American Statistical Association 92 (440): 1268–1278.

Kane, Thomas J., Cecilia Elena Rouse and Douglas Staiger (1997), "Estimating returns to schooling when schooling is misreported", Unpublished paper.

Katz, Lawrence F. and Bruce Meyer (1990), "Unemployment insurance, recall expectations and unemployment outcomes", Quarterly Journal of Economics 105 (4): 973–1002.

Katz, Lawrence F. and Kevin M. Murphy (1992), "Changes in relative wages, 1963–1987: supply and demand factors", Quarterly Journal of Economics 107 (1): 35–78.

Keane, Michael P. and Kenneth Wolpin (1997), "Introduction to the JBES special issue on structural estimation in applied microeconomics", Journal of Business and Economic Statistics 15 (2): 111–114.

Kling, Jeffrey (1998), "Interpreting instrumental variables estimates of the returns to schooling", in: Identifying causal effects of public policies, PhD thesis (Department of Economics, MIT).

Kremer, Michael (1997), "Development datasets", Mimeo. (Department of Economics, MIT).

Krueger, Alan B. (1990a), "Incentive effects of workers' compensation insurance", Journal of Public Economics 41: 73–99.

Krueger, Alan B. (1990b), "Workers' compensation insurance and the duration of workplace injuries", Working paper no. 3253 (NBER, Cambridge, MA).

Krueger, Alan B. and Douglas Kruse (1996), "Labor market effects of spinal cord injuries in the dawn of the computer age", Working paper no. 349 (Industrial Relations Section, Princeton University).

Krueger, Alan B. and Jorn Steffen Pischke (1992), "The effect of social security on labor supply. A cohort analysis of the notch generation", Journal of Labor Economics 10 (2): 412–437.

Krueger, Alan B. and Lawrence H. Summers (1987), "Efficiency wages and the inter-industry wage structure", Econometrica 56 (2): 259–293.

Lalonde, Robert J. (1986), "Evaluating the econometric evaluations of training programs using experimental data", American Economic Review 76 (4): 602–620.

Lang, Kevin (1993), "Ability bias, discount rate bias and the return to education", Mimeo. (Department of Economics, Boston University).

Lazear, Edward P. (1992), "The job as a concept", in: William J. Bruns, Jr., ed., Performance measurement, evaluation and incentives (Harvard Business School Press, Boston, MA).

Leamer, Edward E. (1982), "Let's take the con out of econometrics", American Economic Review 73 (1): 31–43.
Lester, Richard A. (1946), "Shortcomings of marginal analysis for wage-employment problems", American Economic Review 36: 63–82.
Levy, Frank (1987), Dollars and dreams: the changing american income distribution (Russell Sage Foundation, New York).
Lewis, H. Gregg (1963), Unionism and relative wages in the United States: an empirical inquiry (University of Chicago Press, Chicago, IL).
Lewis, H. Gregg (1986), Union relative wage effects (University of Chicago Press, Chicago, IL).
Lillard, Lee, James P. Smith and Finis Welch (1986), "What do we really know about wages? The importance of nonreporting and census imputation", Journal of Political Economy 94 (3): 489–506.
Lusardi, Ann Maria (1996), "Permanent income, current income and consumption: evidence from two panel datasets", Journal of Business and Economic Statistics 14 (1): 81–90.
MaCurdy, Thomas E. (1981), "An empirical model of labor supply in a life-cycle setting", Journal of Political Economy 89 (6): 1059–1085.
Maddala, G.S. (1971) "The likliehood approach to pooling cross-section and time-series data", Econometrica 39: 939–953.
Marquis, K.H., J.C. Moore and K. Bogen (1996), "An experiment to reduce measurement error in the SIPP: preliminary results", Mimeo. (Bureau of the Census).
Marshall, Alfred (1890), Principles of economics (Macmillan, London).
McCarthy, P.J. (1979), "Some sources of error in labor force estimates from the current population survey", in: National Commission on Employment and Unemployment Statistics, Counting the labor force, appendix, Vol. II (US Government Printing Office, Washington, DC).
Medoff, James L. and Katharine G. Abraham (1980), "Experience, performance and earnings", Quarterly Journal of Economics 95 (4): 703–736.
Mellow, Wesley and Hal Sider (1983), "Accuracy of response in labor market surveys: evidence and implications", Journal of Labor Economics 1 (4): 331–344.
Meyer, Bruce D. (1995), "Natural and quasi-experiments in economics", Journal of Business and Economic Statistics 13 (2): 151–161.
Meyer, Bruce D., W. Kip Viscusi and David L. Durbin (1995), "Workers' compensation and injury duration: evidence from a natural experiment", American Economic Review 85: 322–340.
Mincer, Jacob and Yoshio Higuchi (1988), "Wage structures and labor turnover in the U.S. and in Japan", Journal of the Japanese and International Economy 2 (2): 97–133.
Morgenstern, Oskar (1950), On the accuracy of economic observations (Princeton University Press, Princeton, NJ).
Murphy, Kevin M. and F. Welch (1992), "The structure of wages", Quarterly Journal of Economics 197 (1): 285–326.
Newey, Whitney K. (1985), "Generalized method of moments estimation and testing", Journal of Econometrics 29 (3): 229–256.
Nickell, Stephen J. (1981), "Biases in dynamic models with fixed effects", Econometrica 49 (6): 1417–1426.
NLS Handbook (1995), NLS handbook (Center for Human Resource Research, The Ohio State University, Columbus, OH).
Park, Jin Huem (1994), "Returns to schooling: a peculiar deviation from linearity", Working paper no. 339 (Industrial Relations Section, Princeton University).
Parsons, Donald O. (1980), "The decline in male labor force participation", Journal of Political Economy 88 (1): 117–134.
Parsons, Donald O. (1991), "The health and earnings of rejected disability insurance applicants: comment", American Economic Review 81 (5): 1419–1426.
Passell, P. (1992), "Putting the science in social science", New York Times.
Pindyk, Robert S. and Daniel L. Rubinfeld (1991), Econometric models and economic forecasts (McGraw-Hill, New York).

Polivka, Anne (1996), "Data watch: the redesigned current population survey", Journal of Economic Perspectives 10 (3): 169–181.

Polivka, Anne (1997), "Using earnings data from the current population survey after the redesign", Working paper no. 306 (Bureau of Labor Statistics).

Polivka, Anne and Stephen Miller (1995), "The CPS after the redesign: refocusing the economic lens", Mimeo. (Bureau of Labor Statistics).

Poterba, James M. and Lawrence H. Summers (1986), "Reporting errors and labor market dynamics", Econometrica 54 (6): 1319–1338.

Powell, James L., James H. Stock and Thomas M. Stoker (1989), "Semiparametric estimation of index coefficients", Econometrica 57 (6): 1403–1430.

Riddell, W. Craig (1992), "Unionization in Canada and the United States: a tale of two countries", Mimeo. (Department of Economics, University of British Columbia).

Rosenbaum, Paul R. (1984), "The consequences of adjustment for a concomitant variable that has been affected by the treatment", Journal of the Royal Statistical Society Series A 149: 656–666.

Rosenbaum, Paul R. (1995), Observational studies (Springer-Verlag, New York).

Rosenbaum, Paul R. and Donald B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", Biometrika 70: 41–55.

Rosenbaum, Paul R. and Donald B. Rubin (1984), "Reducing bias in observational studies using subclassification on the propensity score", Journal of the American Statistical Association 79: 516–524.

Rosenbaum, Paul R. and Donald B. Rubin (1985), "Constructing a control group using multi-variate matching methods that include the propensity score", American Statistician 39: 33–38.

Rosenzweig, Mark R. and Kenneth I. Wolpin (1980), "Testing the quantity-quality model of fertility: the use of twins as a natural experiment", Econometrica 48 (1): 227–240.

Rosovsky, Henry (1990), The university: an owner's manual (W.W. Norton and Company, New York).

Rothgeb, Jennifer M. and Sharon R. Cohany (1992), "The revised CPS questionnaire: differences between the current and the proposed questionnaires", Paper presented at the Annual Meeting of the American Statistical Association.

Rubin, Donald B. (1973), "Matching to remove bias in observational studies", Biometrics 29 (1): 159–183.

Rubin, Donald B. (1974), "Estimating causal effects of treatments in randomized and non-randomized studies", Journal of Educational Psychology 66: 688–701.

Rubin, Donald B. (1977), "Assignment to a treatment group on the basis of a covariate", Journal of Educational Statistics 2: 1–26.

Rubin, Donald B. (1983), "Imputing income in the CPS: comments on 'measures of aggregate labor cost in the United States'", in: Jack E. Triplett, ed., The measurement of labor cost (University of Chicago Press, Chicago, IL).

Sawa, Takamitsu (1969), "The exact sampling distribution of ordinary least squares and two-stage least squares estimators", Journal of the American Statistical Association 64 (327): 923–937.

Sawa, Takamitsu (1973), "An amost unbiased estimator in simultaneous equations systems", International Economic Review 14 (1): 97–106.

Siegel, Paul and Robert Hodge (1968), "A causal approach to the study of measurement error", in: Hubert Blalock and Ann Blalock, eds., Methodology in social research (McGraw-Hill, New York) pp. 28–59.

Siegfried, John J. and George H. Sweeney (1980), "Bias in economics education research from random and voluntary selection into experimental and control groups", American Economic Review 70 (2): 29–34.

Singer, Eleanor and Stanley Presser (1989), Survey research methods (University of Chicago Press, Chicago, IL).

Solon, Gary R. (1985), "Work incentive effects of taxing unemployment benefits", Econometrica 53 (2): 295–306.

Stafford, Frank (1986), "Forestalling the demise of empirical economics: the role of microdata in labor economics research", in: Orley Ashenfelter and Richard Layard, eds., Handbook of labor economics (North-Holland, Amsterdam).

Staiger, Douglas and James H. Stock (1997), "Instrumental variables regression with weak instruments", Econometrica 65 (3): 557–586.
Stigler, Stephen M. (1977), "Do robust estimators work with real data?" Annals of Statistics 5 (6): 1055–1098.
Stoker, Thomas M. (1986), "Aggregation, efficiency and cross-section regression", Econometrica 54 (1): 171–188.
Sudman, Seymour and Norman Bradburn (1991), Asking questions: a practical guide to survey design (Jossey-Bass Publishers. San Francisco, CA).
Taubman, Paul (1976), "Earnings, education, genetics and environment", Journal of Human Resources 11 (Fall), 447–461.
Taussig, Michael K. (1974), Those who served: report of the Twentieth Century Fund Task Force on policies towards veterans (The Twentieth Century Fund, New York).
Thurow, Lester C. (1983), Dangerous currents: the state of economics (Random House, New York).
Topel, Robert H. (1991), "Specific capital, mobility and wages: wages rise with job seniority", Journal of Political Economy 99 (1): 145–176.
Trochim, William K. (1984), Research design for program evaluation: the regression-discontinuity approach (Sage, Beverly Hills, CA).
Tufte, Edward R. (1992), The visual display of quantitative information (Graphics Press, Cheshire, CT).
Tukey, John W. (1977), Exploratory data analysis (Addison-Wesley Publishing Company, Reading, MA).
US Bureau of the Census (1992), Current population survey, March 1992. Technical documentation (Bureau of the Census, Washington, DC).
US Bureau of the Census (1996), Census of population and housing, 1990 United States: public use microdata sample: 5 percent sample. Third ICPSR release (US Department of Commerce,Washington, DC).
van der Klaauw, Wilbert (1996), "A regression-discontinuity evaluation of the effect of financial aid offers on college enrollment", Unpublished manuscript (Department of Economics, New York University).
Vroman, Wayne (1990), "Black men's relative earnings: are the gains illusory?" Industrial and Labor Relations Review 44 (1): 83–98.
Vroman, Wayne (1991), "The decline in unemployment insurance claims activity in the 1980s", UI occasional paper no. 91-2 (Employment and Training Administration, US Department of Labor).
Wald, A. (1940), "The fitting of straight lines if both variables are subject to error", Annals of Mathematical Statistics 11: 284–300.
Welch, Finis (1975), "Human capital theory: education, discrimination and life-cycles", American Economic Review 65: 63–73.
Welch, Finis (1977), "What have we learned from empirical studies of unemployment insurance?" Industrial and Labor Relations Review 30: 451–461.
Westergard-Nielsen, Niels (1989), "Empirical studies of the European labour market using microeconomic datasets: introduction", European Economic Review 33 (2/3): 389–394.
White, Halbert (1980), "Using least squares to approximate unknown regression functions", International Economic Review 21 (1): 149–170.
Willis, Robert J. and Sherwin Rosen (1979), "Education and self-selection", Journal of Political Economy 87 (5): S7–S36.
Yitzhaki, Shlomo (1996), "On using linear regressions in welfare economics", Journal of Business and Economic Statistics 14: 478–486.