John Bound, Charles Brown, and Nancy Mathiowetz

Measurement Error in Survey Data

Report No. 00-450

Research Report
The Population Studies Center at the University of Michigan is one of the oldest population centers in the United States. Established in 1961 with a grant from the Ford Foundation, the Center has a rich history as the main workplace for an interdisciplinary community of scholars in the field of population studies. Today the Center is supported by a Population Research Center Core Grant from the National Institute of Child Health and Human Development (NICHD) as well as by the University of Michigan, the National Institute on Aging, the Hewlett Foundation, and the Mellon Foundation.

PSC Research Reports are prepublication working papers that report on current demographic research conducted by PSC associates and affiliates. The papers are written by the researcher(s) for timely dissemination of their findings and are often later submitted for publication in scholarly journals. The PSC Research Report Series was begun in 1981 and is organized chronologically. Copyrights are held by the authors. Readers may freely quote from, copy, and distribute this work as long as the copyright holder and PSC are properly acknowledged and the original work is not altered.
Measurement Error in Survey Data

John Bound
Professor, Department of Economics and the Populations Studies Center, University of Michigan and
the National Bureau of Economic Research

Charles Brown
Professor, Department of Economics and the Survey Research Center, University of Michigan and
the National Bureau of Economic Research

Nancy Mathiowetz
Assistant Professor, Joint Program in Survey Methodology, University of Maryland
and
Adjunct Assistant Research Scientist, Institute for Social Research, University of Michigan

April 2000

Acknowledgments: We are grateful to, Joe Altonji, Dan Hamermesh, Jeff Wooldridge, Gary Solon,
Shinichi Sakata, participants at the conference for this volume and especially to Jim Heckman and
Ed Leamer for extremely helpful comments on previous versions of this paper and to Mari Ellis for
help with the preparation of the manuscript.
Contents

Abstract
1.0 Introduction
2.0 The Impact of Measurement Error on Parameter Estimates
   2.1 Special Cases
   2.2 General Results – Linear Model
   2.3 Differential Measurement Error -- an Example
   2.4 Bounding Parameter Estimates
   2.5 Contaminated and Corrupted Data
   2.6 Measurement Error in Categorical Variables
   2.7 Nonlinear Models
3.0 Correcting for Measurement Error
   3.1 Instrumental Variables in the Bivariate Linear Model
   3.2 Multivariate Linear Model
   3.3 Nonlinear Models
   3.4 The Contribution of Validation Data
4.0 Approaches to the Assessment of Measurement Error
5.0 Measurement Error and Memory: Findings from Household-Based Surveys
   5.1 Cognitive Processes
   5.2 Social Desirability
   5.3 Essential Survey Conditions
   5.4 Applicability of Findings to the Measurement of Economic Phenomena
6.0 Evidence on Measurement Error in Survey Reports of Labor-related Phenomena
   6.1 Earnings
      6.1.1 Annual Earnings
      6.1.2 Monthly, Weekly, and Hourly Earnings
   6.2 Transfer Program Income
   6.3 Assets
   6.4 Hours Worked
   6.5 Unemployment
      6.5.1 Current Labor Force Status and Transitions to and from Unemployment
      6.5.2 Retrospective Unemployment Reports
   6.6 Industry and Occupation
   6.7 Tenure, Benefits, Union Coverage, Size of Establishment, and Training
   6.8 Measurement Error in Household Reports of Health-Related Variables
      6.8.1 Health Care Utilization, Health Insurance, and Expenditures
      6.8.2 Health Conditions and Health/Functional Status
   6.9 Education
7.0 Conclusions
Abstract

Economists have devoted increasing attention to the magnitude and consequences of measurement error in their data. Most discussions of measurement error are based on the “classical” assumption that errors in measuring a particular variable are uncorrelated with the true value of that variable, the true values of other variables in the model, and any errors in measuring those variables. In this survey, we focus on both the importance of measurement error in standard survey-based economic variables and on the validity of the classical assumption.

We begin by summarizing the literature on biases due to measurement error, contrasting the classical assumption and the more general case. We then argue that, while standard methods will not eliminate the bias when measurement errors are not classical, one can often use them to obtain bounds on this bias. Validation studies allow us to assess the magnitude of measurement errors in survey data, and the validity of the classical assumption. In principle, they provide an alternative strategy for reducing or eliminating the bias due to measurement error.

We then turn to the work of social psychologists and survey methodologists which identifies the conditions under which measurement error is likely to be important. While there are some important general findings on errors in measuring recall of discrete events, there is less direct guidance on continuous variables such as hourly wages or annual earnings.

Finally, we attempt to summarize the validation literature on specific variables: annual earnings, hourly wages, transfer income, assets, hours worked, unemployment, job characteristics like industry, occupation, and union status, health status, health expenditures, and education. In addition to the magnitude of the errors, we also focus on the validity of the classical assumption. Quite often, we find evidence that errors are negatively correlated to true values.

The usefulness of validation data in telling us about errors in survey measures can be enhanced if validation data is collected for a random portion of major surveys (rather than, as is usually the case, for a separate convenience sample for which validation data could be obtained relatively easily); if users are more actively involved in the design of validation studies; and if micro data from validation studies can be shared with researchers not involved in the original data collection.
1.0 Introduction

Empirical work in economics depends crucially on the use of survey data. The evidence we have, however, makes it clear that survey responses are not perfectly reliable. Even such salient features of an individual’s life as years of schooling seem to be reported with some error. While economists have been aware of the errors in survey data for a long time, until recently most empirical studies tended to ignore it altogether. However, perhaps stimulated by increases in the complexity of the models we have been estimating, and in particular, with the increasing use of panel data that can seriously exacerbate the effect of measurement error on our estimates, economists have been paying an increasing amount of attention to measurement error.\(^1\)

Most assessments of the consequences of measurement error and methods for correcting the biases it can cause have emphasized models that make strong—and exceedingly convenient—assumptions about the properties of the error. Most frequently, measurement error in a given variable is assumed to be independent of the true level of that and all other variables in the model, measurement error in other variables, and the stochastic disturbance. We will refer to such purely random measurement error as “classical” measurement error. In some applications—such as the case where the error is a sampling error in estimating a population mean—these assumptions can be justified. But in most micro data analyses using survey data, they reflect convenience rather than conviction.

From these assumptions comes much of the conventional wisdom about the effects of measurement error on estimates in linear models: (i) error in the dependent variable neither biases nor renders inconsistent the parameter estimates but simply reduces the efficiency of those estimates; (ii) error in the measurement of an independent variables produces downward-biased (attenuated) and inconsistent parameter estimates of its effect, while inadequately controlling for the confounding effects of this variable on the well measured variables; and (iii) the inclusion of other independent variables that are correlated with the mis-measured independent variable accentuates the downward bias.\(^2\)

In fact, these conclusions need to be qualified. The bias introduced by measurement error depends both on the model under consideration (e.g., whether it is linear) and on the joint

\(^1\)Thus, for example, the volatility of earnings and consumption data have often been attributed measurement error (MaCurdy, 1982; Abowd and Card, 1987, 1989, Hall and Mishkin, 1982; Shapiro, 1982). On the other hand a variety of authors have rationalized a dramatic drop in the magnitude of coefficient estimates associated with the move to fixed effects models in terms of measurement error in key variables and have used a variety of techniques to undo the presumed damage (Freeman, 1984 and Card, 1996 follow this kind of strategy when using fixed effect models to estimate union premia, Krueger and Summers (1988) do so when estimating industry premia, and Ashenfelter and Krueger (1994) do so when estimating educational premia).

\(^2\)The notion that fixed effect models tend to seriously accentuate the effect of measurement error on parameter estimates represents an important special case of this last point.
distribution of the measurement error and all the variables in the model. The effect of measurement error can range from the simple attenuation described above to situations where (i) real effects are hidden; (ii) observed data exhibit relationships that are not present in the error free data; and (iii) even the signs of the estimated coefficients are reversed.

Standard methods for correcting for measurement error bias, such as instrumental variables estimation, are valid when errors are classical and the underlying model is linear, but not, in general, otherwise. While statisticians and econometricians have been quite clear about the assumptions built into procedures they have developed to correct for measurement error, empirical economists have often relied on such procedures without giving much attention to the plausibility of the assumptions they are explicitly or implicitly making about the nature of measurement error. Not only can standard fixes not solve the underlying problem, they can make things worse!

Twenty years ago, analysts would typically have ignored the possibility that the data they were using was measured with considerable error. Rarely, if ever would such researchers acknowledge, let alone try to justify their tacit assumption that measurement error in the data they were using was negligible. More recently, it has become quite common for analysts to correct for measurement error. However, when doing so, researchers virtually always rely on the assumption that measurement error is of the classical type, usually with no justification at all. If we are to be serious regarding measurement error in our data, we need to understand the relationship between the constructs that enter our models and the measures we use to proxy them. This is a tall order. However, even when this “gold standard” is unattainable it will often be possible to put some kind of plausible bounds on the extent and nature of the measurement error of key variables, and use these bounds to work out bounds for estimated parameters of interest.

In addition to providing some evidence about the magnitude of measurement errors, validation studies that compare survey responses to more accurate data such as payroll records permit one to determine whether measurement errors are indeed uncorrelated with other variables. In principle—though this possibility has been realized only incompletely in practice—validation studies can provide more general information on the relationships among errors in measuring each variable, its true value, and the errors and true values in each of the other variables. The research summarized in this chapter is based on direct observation of the measurement error properties of interview reports for a wide range of economic measures. The evidence provides much information to challenge the conventional wisdom.

One general conclusion from the available validation evidence is that the possibility of non-classical measurement error should be taken much more seriously by those who analyze survey data, both in assessing the likely biases in analyses that take no account of measurement error and in devising procedures that “correct” for such error. A second result is that it is important to be at least as explicit about one’s model of the errors in the data as about the relationship among the “true” variables that we seek to estimate. Unless one is comfortable assuming that the classical assumptions apply, arguing informally based on that standard case may be dangerous, and writing out the alternative model that better describes one’s data can often give real insight into the biases one faces and the appropriateness of traditional cures. A third finding is that, all too often, validation studies are not as helpful to data analysts as they ought to be. Even for the relatively simple goal of
assessing the extent of measurement error in individual variables, the “extent” of the error is often
not summarized in ways that are suggested by the simple models that guide our thinking. Too few
studies take the next step of relating errors in one variable to true values and errors in other variables.
We hope that by contrasting the information that is often provided with that which would be most
helpful to analysts, we can increase the contribution made by future validation studies. Given the
difficulty of mounting a successful validation effort, maximizing the payoff from such efforts is
important. In addition to these general themes, we present a variable-by-variable summary of what is
known about the accuracy of survey measures.

We begin by reviewing what is known about the impact of measurement error on parameter
estimates in section 2, and possible corrections for the effect of such error in section 3. This review
is not meant as an exhaustive survey the large statistical literature on this subject, but rather is meant
to introduce the reader to various issues and to set the stage for our discussion of the validation
studies we review.3 What summary measures of the errors in survey data would be most valuable to
an analyst for deciding how important such errors are for his/her analysis? How appropriate are
standard techniques for “correcting” for measurement errors, given what validation studies can tell
us about such errors? In section 4 we briefly discuss the design of validation studies, while section 5
reviews what is known about the circumstances under which phenomena are likely to be well
reported by survey respondents. Finally section 6 reviews validation studies across a large range on
substantive areas. This review is organized by variable, so readers can concentrate on the variables
that are most important in their own research. We offer some conclusions in section 7.

2.0 The Impact of Measurement Error on Parameter Estimates

We start with the presumption that we are interested in using survey data to estimate some
parameters of interest. These parameters might be means or medians, as is the case when we are
interested in tracking the unemployment rate or median earnings, but will often represent more
complicated constructs such as differences in means between groups or regression slope coefficients.
Measurement error in survey data will typically introduce biases into such estimates. If what we are
interested in is the estimate of simple means, then, as long as measurement error is mean 0, it will
not bias our estimates. However, as is well known, if we are interested in parameters that depend on
relationships between variables, then even mean 0 measurement error will typically bias our
estimates.

In what follows we will focus primarily on the impact of measurement on parameter
estimates within the context of the linear model.4 Most of the statistics and econometrics literature

3 Fuller (1987) contains a thorough discussion of the biases measurement error introduces
into parameter estimates and on standard methods for correcting such biases within the context of
the linear model, when measurement error is random (classical). Carroll, Ruppert and Stefanski
(1995) contains a more general discussion of the same issues within the context of non-linear
models, with considerable attention to models in with errors are not purely random.

4 The framework we present here derives from Bound et al. 1994.
on the subject has dealt with this case, presumably because it is in this case that the impact of measurement error on parameter estimates can be well characterized.\(^5\) There is a growing literature focusing on the impact of measurement error on parameter estimates within the context of non-linear models; however it remains unclear the extent to which the intuitions we develop within the context of the linear model remain true within this context (see section 2.6 and 3.3 for further discussion of this point).

Assume the true model is

\[
y^* = X^* \beta + \varepsilon
\]

and \(\varepsilon\) are both scalars and \(X^*\) and \(\beta\) are vectors. We will maintain the assumption that \(\varepsilon\) is uncorrelated with \(X^*\). The motivation for this assumption is largely strategic -- we are interested in the impact that measurement error has on our estimates and so focus on the case where our estimates would be unbiased in its absence. Instead of \(X^*\) and \(y^*\), we observe \(X\) and \(y\), where

\[
X = X^* + \mu

y = y^* + \nu.
\]

In general, we will not assume \(\mu\) and \(\nu\) are uncorrelated with \(X^*\), \(y^*\) or \(\varepsilon\). We will use the term classical measurement error to refer to the case where \(\mu\) and \(\nu\) are assumed to be uncorrelated with \(X^*\), \(y^*\) or \(\varepsilon\), and the term nondifferential (Carroll, Ruppert and Stefanski, 1995) measurement error (in explanatory variables) to refer to the case where, conditional on \(X^*\), \(X\) contains no information about \(y^*\), implying that \(\mu\) is uncorrelated with either \(y^*\) or \(\varepsilon\).\(^8\)

---

\(^5\) Fuller’s excellent monograph focuses solely on the linear mode.

\(^6\) For the more general case (i.e., nonlinear modes), this condition needs to be strengthened to refer to the case where \(\mu\) and \(\nu\) are assumed to be independent of \(X^*\), \(y^*\) or \(\varepsilon\).

\(^7\) More technically, measurement error in \(X^*\) is referred to a nondifferential if the distribution of \(y^*\) given \(X^*\) and \(X\) depends only on \(X^*\) (i.e., \(f(y^*|X^*,X) = f(y^*|X^*)\)).

\(^8\) A few examples my clarify the kind of contexts in which differential measurement can occur. Kaestner, Joyce, and Wehbeh (1996) estimate the effect of maternal drug use on an infant’s birth weight. They find that self-reported drug use has a larger estimated effect on birth weight than does drug use as assessed from clinical data. They argue that this is because casual users tend to under-report drug use. Thus, if \(X^*\) is a binary measure of drug use based on clinical data and \(X\) the self report, and \(y^*\) is birth weight, \(E(y^*|X^*,X)\) is increasing in \(X\). It is also plausible for measurement error to be differential in the context in which \(X\) is does not merely represent a mismeasured version of \(X^*\), but is a separate variable representing a proxy for \(X^*\).
Measurement error in the above sense can occur for a number of reasons that are worth keeping distinct. Respondents can simply misreport on a measure because, for example, their memory is flawed. Here it is possible to imagine obtaining a perfectly measured and therefore valid measure of the quantity in question. An example of this might be pre-tax wage and salary income (i.e., earnings) for a specific calendar year. Alternatively we may be using \( X \) and \( y \) to proxy for the theoretical constructs of our economic models. Thus, for example, we might use reported years of educational attainment as a proxy for human capital. In this case, the errors will importantly include the gap between what the survey intended to measure and our theoretical construct. While something of the same statistical apparatus can be used to analyze the impact of either kind of error on parameter estimates, clearly validation data can shed light on only the first kind of error.

In the absence of validation data, the analyst observes only \( X \) and \( y \). We will be primarily interested in the effect of measurement error on the consistency of our estimates. For this reason, we will not distinguish between populations and samples. The “least squares estimator” of \( \beta \) is

\[
\beta_{yX} = [X'X]^{-1}X'y
\]  

### 2.1 Special Cases

We will present a general approach to dealing with measurement errors in \( X^* \) and \( y^* \) which are correlated with the true \( X, y \) and \( \epsilon \). Before doing so, however, it is useful to highlight a few results that can be derived from the general approach for the biases due to measurement errors when convenient assumptions hold. To simplify discussion of the various biases, we assume throughout that the \( X \)'s have been defined so that \( \beta_{yX} \geq 0 \). Consider three special cases.

First, if there is classical measurement error in only one independent variable \( x_j \), the proportional bias in estimating \( \beta_j \) depends on the noise to total variance ratio, \( \sigma_{uj}^2 / \sigma_{x_j}^2 \). In particular, with only one independent variable in the regression, the proportional bias is just equal to this ratio,

\[
\beta_{jx_j} = \beta [1 - \frac{\sigma_{uj}^2}{\sigma_{x_j}^2 + \sigma_{uj}^2}].
\]  

---

(Carroll, Ruppert and Stefanski, 1995). Thus, for example, if there are contextual effects, the use of aggregate proxies for micro level constructs can exaggerate causal effects (Loeb and Bound, 1996; Geronimus, Bound and Neidert, 1996).
With other variables in the regression:

\[
\beta_{yx|Z} = \beta_j \left[ 1 - \frac{\sigma_{uj}^2}{\sigma_{xj|^2}^2 \left( 1 - R_{xj|^2, Z}^2 \right) + \sigma_{\mu j}^2} \right]
\]

(5)

where \( Z^* \) represents the elements of \( X^* \) other than \( x_j^* \) (\( X^* = [x_j^* | Z^*] \)), \( R_{xj|^2, Z}^2 \) represents the \( R^2 \) from the regression of \( x_j^* \) on the remaining elements of \( X^* \), and \( b_{yx|Z} \) represents the least squares regression of \( y \) on \( x_j \) holding \( Z \) constant. Thus, classical measurement error in just one explanatory variable attenuates estimates of the effect of this variable on outcomes. The magnitude of this attenuation depends both on the noise to signal ratio and on the extent of multi-collinearity between the error ridden variable and the other variables in the equation (Levi, 1973; Garber and Klepper, 1980).

The measurement error in \( x_j \) biases not just estimates of \( \beta_j \), but also the coefficients on the accurately measured variables. Letting \( \Pi \) represent the coefficient vector from the least squares regression of \( x_j \) on \( Z^* \), then

\[
\beta_{yx|Z} = \beta_{i|j} + (\beta_j - \beta_{yx|Z}) \Pi.
\]

(6)

Thus, classical measurement error in \( x_j \) implies that using \( x_j \) as a proxy for \( x_j \) will partially, but only partially, control for the confounding effects of \( x_j \) on the estimates of the effect of other variables on outcomes (McCallum, 1972; Wickens, 1972; Garber and Klepper, 1980).

Second, even if the error, \( \mu_j \), is correlated with the true \( x_j \) (or other \( X^* \)'s), but is uncorrelated with \( \epsilon \), the proportional downward bias is equal to the regression coefficient from a hypothetical regression of \( \mu_j \) on the set of measured \( X \)'s. If there is only one independent variable in the model, this reduces to the simple regression coefficient \( \beta_{\mu X} \),

\[
\beta_{yx} = \beta [1 - \beta_{\mu X}].
\]

When \( \mu \) and \( X^* \) are uncorrelated, \( \beta_{\mu X} \) is equal to the variance ratio \( \sigma_{\mu}^2 / (\sigma_{y|x}^2 + \sigma_{\mu}^2) \) and, as such, will be between 0 and 1. More generally, this will not be true. In particular if \( \mu \) and \( X \) are negatively correlated (the error \( \mu \) is "mean reverting"), \( \beta_{\mu X} \) will be smaller than in the classical case and can even be negative -- that is, \( \beta_{\mu X} > \beta \). More generally, if the error in one variable is correlated with other variables in the model, the biases on various coefficients depend on the direction of the partial correlation between the error and the various variables in the model.
Third, if the dependent variable $Y$ is measured with error, and that error is correlated with the true $y^*$ (where $v = \delta y + v^*$ and $v^*$ is uncorrelated with $X^*$ and $\varepsilon$), and the $X^*$s are measured without error, then the proportional bias in estimating $\beta$ is just equal to $\delta$. To emphasize the similarity to the previous case, note that $\delta$ is just the regression coefficient $\beta_{vy}$.

Each of the above results applies to cross-section analysis, and to panel data by substituting $\Delta X^*$ for $X^*$, etc. But when one uses $\Delta y$ and $\Delta X$ as one's dependent and independent variables, respectively, another aspect of the data becomes important -- the correlation over time in the true values (the correlation between $y$ at time $t$ and at time $t-1$, and similarly for $X$) and in the measurement errors (the correlation between $v$ at time $t$ and at time $t-1$, and similarly for $\mu$). A general result is that, if the variance of a variable (say, $X^*$) is the same in both years, the variance of $\Delta X^*$ is equal to $2\sigma^2_X(1-r_{X^*_tX^*_{t-1}})$ which is greater or less than $\sigma^2_X$, as $r_{X^*_tX^*_{t-1}}$ is less than or greater than one-half. A common concern, usually expressed in the context of classical measurement errors, is that true values of $X$ will be highly correlated over time, while the measurement errors will be more or less uncorrelated. In this case, $\sigma^2_{\Delta X}$ will be less than $\sigma^2_X$, while $\sigma^2_{\Delta \mu}$ will be greater than $\sigma^2_{\mu}$, so that moving from "levels" to "changes" intensifies the bias due to errors in measuring the independent variable(s).\footnote{See Griliches and Hausman (1986) for an illuminating discussion of these issues.}

There is one more special case worth noting. Suppose that $x_j$ represents a component of $x_j^*$, with $r_{\mu,x_j}=0$, and that other variables (both $y^*$ and the other $X^*$s) are measured without error. Take first the case where $x_j^*$ represents the only explanatory variable in the model. (1) can now be rewritten:

$$y = \beta [x_j + \mu_j] + \varepsilon = \beta x_j + [\beta \mu_j + \varepsilon].$$

Since, by assumption, $\mu_j$ is orthogonal to $x_j$, the composite error in (7) will be orthogonal to $x_j$ and $b_{\beta x_j}$ will consistently estimate $\beta_j$. With other variables in the equation, OLS will no longer consistently estimate $\beta$. Rewriting (1) in this case we have:

\footnote{This variance component framework fits many different kinds of contexts. Thus, for example, we might imagine that $x_j^*$ represents schooling, with $x_j$ representing the observed quantity of schooling obtained and $\mu_j$ representing the unobserved quality of this schooling (here one might question the orthogonality of $\mu_j$ and $x_j$). Alternatively, $x_j$ might represent cell means of $x_j^*$ (we use industry specific injury rates as a proxies for job specific injury rates). Here $\mu_j$ and $x_j$ are orthogonal by construction. More detailed discussions of this latter case can be found in Dickens and Ross (1984) and Geronimus, Bound and Neidert (1996).}
\[ y = \beta_j [x_j + \mu_j] + Z\gamma + \varepsilon \]
\[ = \beta_j x_j + Z\gamma + [\beta_j \mu_j + \varepsilon]. \]  

While \( \mu_j \) is orthogonal to \( x_j \), we do not expect \( \mu_j \) to be orthogonal to \( Z \). Thus, in this case the exclusion of \( \mu_j \) from our estimating equation represents a specification error, and both \( \beta_{yx_j} \) and \( \beta_{yZ,x_j} \) will be biased. If the signs of the partial correlations between \( \mu_j \) and \( Z \) are the same as signs of the partial correlations between \( x_j \) and \( Z \), then using \( x_j \) as a proxy for \( x_j^* \) will only partially control for the confounding effect of \( Z \) on \( \beta \). -- as an estimate of \( \beta \), \( \beta_{yx_j} \) will be still be biased in the same direction as is \( \beta_{yZ,x_j} \).

### 2.2 General Results – Linear Model

Having highlighted some special cases in which the consequences of measurement error can be summarized succinctly, we turn to a more general model. With \( \mu \) and \( \nu \) potentially correlated with \( X^* \) and \( y^* \), the least squares regression coefficient can be rewritten as

\[ \beta_{yx} = (X'X)^{-1}X'(X\beta - \mu \beta + \nu + \varepsilon) \]
\[ = \beta + (X'X)^{-1}X'(-\mu \beta + \nu + \varepsilon). \]  

Therefore, the bias of the least squares estimator of \( \beta \) is

\[ \beta_{yx} - \beta = (X'X)^{-1}X'(-\mu \beta + \nu + \varepsilon). \]  

It is useful to collect the measurement errors and their coefficients. Define

\[ \gamma = \begin{bmatrix} -\beta \\ 1 \\ 1 \end{bmatrix} \]

\[ \omega = [\mu \mid \nu \mid \varepsilon] \]

\[ \beta_{yx_j} > \gamma \] and \( \beta_{yZ,x_j} \leq \beta_j \).
Then (10) can be rewritten as
\[ \beta_{\mathbf{x}X} - \beta = (X'X)^{-1}X'y = A\gamma. \]

If there are \( k \) separate variables in the independent-variable matrix \( X \), then \( A \) is \( k \) by \( k + 2 \). It can be rewritten in a more intuitive form as
\[ A = [\beta_{\mu X} | \beta_{\nu X} | \beta_{\varepsilon X}] \]

where the \( j \)th column of \( \beta_{\mu X} \) consists of the coefficients from regressing \( \mu_j \) on \( X \), and \( \beta_{\nu X} \) and \( \beta_{\varepsilon X} \) represent the set of coefficients from regressing \( \nu \) and \( \varepsilon \) on \( X \).

If there is measurement error in only one independent variable \( X_j^* \) and if this error is uncorrelated with \( \varepsilon \), only one column of \( A \) will be nonzero, and \( A_{\mu_j} = \beta_{\mu X} \), as claimed in our discussion of special cases. If \( \nu = \delta y' + \nu' = \delta X'\beta + \delta \varepsilon + \nu' \), and \( \nu' \) is uncorrelated with the other variables of the model, and the independent variables are measured without error, then \( \beta_{\mu X} \) and \( \beta_{\varepsilon X} \) are matrix and vector of zeros, and \( \beta_{\nu X} = \delta \beta \). Thus, the proportional bias for each coefficient equals \( \delta \).

As the above expression makes clear, with measurement error in more that one explanatory variable, the bias on any particular coefficient will involve multiple terms, and is hard to characterize. What should be clear is that without some knowledge of the distribution of the errors (\( \mu \) and \( \nu \)), the situation is hopeless -- the data put no restrictions on possible values of \( \beta \).

Even with classical assumptions, measurement error in more than one explanatory variable does not necessarily attenuate the coefficients on the variables measured with error. Theil (1961) derives a useful approximation to the bias in the context of where two variables are measured with error. He imagines we are interested in estimating the relationship:
\[ y^* = \beta_1 x_1^* + \beta_2 x_2^* + \varepsilon, \]  

but observe only error ridden proxies for the \( x^* \)’s, \( x_1 (x_1 = x_1^* + \mu_1) \) and \( x_2 (x_2 = x_2^* + \mu_2) \). The errors (the \( \mu \)’s) are assumed to be independent of each other, the \( x \)’s and \( \varepsilon \) and the \( x^* \)’s are scaled to have unit variance. Theil shows that when the errors are small

\[ \beta_{yX_1,X_2} - \beta = \frac{-\beta_1 \lambda_{11}}{1 - \rho^2} + \frac{\beta_2 \lambda_{21}}{1 - \rho^2} \]

\[ \beta_{yX_2,X_1} - \beta = \frac{-\beta_2 \lambda_{12}}{1 - \rho^2} + \frac{\beta_1 \lambda_{12}}{1 - \rho^2} \]
where \( \rho \) represents the correlation between the \( x^* \)'s, and the \( \lambda \)'s represent the error to total variance ratios for the two variables \( (\lambda_j = \sigma_{\mu_j}^2 / \sigma_{x_j}^2) \). Thus, in the multivariate case, the bias on a particular coefficient depends on factors that, as long as \( \rho \) is positive, tend to offset each other. In fact, it should be clear that in the two variable case, the bias on the estimated coefficient on the variable measured with less error can be positive.\(^{12}\)

### 2.3 Differential Measurement Error -- an Example

In many cases, assuming that measurement error is classical is a simple (and potentially dangerous) expedient when we have little a priori reason to believe that any other particular assumption would be more plausible. In other situations, however, we have good reason to believe that the errors are differential, and the basis for this belief can help us write down relatively detailed but still manageable models. The growing literature on labor supply of older workers provides a useful example, both because it is relevant for our discussion of survey measures of health and because doing so will allow us to highlight the potential importance of differential measurement error.\(^{13}\)

A large fraction of the men and women who leave the workforce before the age of 62 report health as the reason they do so. Though health is, no doubt, an important determinant of the age at which men and women retire, there are a variety of reasons not to take these self-reports at face value. It seems plausible that men and, to a less extent women, rationalize retirement in terms of health even when they retire primarily for other reasons.\(^{14}\) Myers (1982) has gone so far as to argue that there is no useful information in self-evaluated health. At the same time, for want of alternative measures, econometric analyses of the labor supply decisions of older men and women have generally used respondents' self-assessment of their health. There remain important questions about the validity of self-reported measures of health and therefore of the inferences that can be drawn from studies that use them.

The most common health measures used in retirement research have been global questions such as, "Does health limit the amount or kind of work you can perform?" or "How would you rate your health? Is it excellent, very good, good, fair or poor?" There are a number of reasons to be suspicious of such survey measures (Parsons, 1982; Anderson and Burkhauser, 1984, 1985; Bound, 1991; Waidmann et al., 1995). First, respondents are being asked for subjective judgments and there

---

\(^{12}\) When more than one variable is measurement with error, not only is it no longer true that the coefficients on these variables are necessarily attenuated but it is also no longer true that the inclusion of one of the error ridden variables will necessarily reduce the bias on coefficients on accurately measured variables. See Garber and Klepper (1980) for a succinct discussion of these issues.

\(^{13}\) The discussion here follows Bound (1991) closely.

\(^{14}\) Plausibly, this rationalization is not entirely conscious.
is no reason to expect that these judgments will be entirely comparable across respondents. Second, responses may not be independent of the labor market outcomes we may wish to use them to explain. Third, since health may represent one of the few 'legitimate' reasons for a working aged man to be out of work, men out of the labor force may mention health limitations to rationalize their behavior. Lastly, since early retirement benefits are often available only for those deemed incapable of work, men and women will have a financial incentive to identify themselves as disabled, an incentive that will be particularly high for those for whom the relative rewards from continuing to work are low.

Each of these problems will lead to a different kind of bias. The lack of comparability across individuals represents measurement error that is likely to lead to our underestimating the impact of health on labor force participation, while the endogeneity of self-reported health is likely to lead to our exaggerating its impact. Biases in our estimation of health's impact on outcomes will also induce biases on coefficients of any variables correlated with health. Finally the dependence of self-reported health on the economic environment will induce a bias on estimates of the impact of economic variables on participation, regardless of whether we correctly measure the impact of health itself.

As an alternative to using global self-reported health measures, a variety of authors have argued for the use of what have been perceived to be more objective indicators of health: responses to questions about specific health conditions or limitations, doctors' reports or information on subsequent mortality. Such proxies are presumed to be more objective than self-reported health measures, though this does not mean that reports of specific conditions are completely reliable (see Section 6.8.2). Moreover, even with perfectly accurate measures of health conditions or mortality, it is not clear that their use as proxies for health give us an accurate indication of the impact of health on labor supply. Part of the problem with 'objective' measures is that they measure health rather than work capacity. As long as these health proxies are not perfectly correlated with work capacity -- the aspects of health that affect an individual's capacity of work -- they will suffer from errors in variables problems. With self-reported health measures we have biases working in opposite directions and, as such, they will have a tendency to cancel each other out. With objective measures there is only one bias, and, as long as the correlation between the proxy and actual health isn't close to perfect, the bias will be quite substantial.

The issues here are important for our understanding not only of the importance of health, but also of the impact of economic variables on early retirement. Both subjective and objective health indicators are correlated with such things as education, race, pre-retirement earnings, and pre-retirement occupation. These factors are also important indicators of early labor market withdrawal. One interpretation of these correlations is that it is those in poor health who leave the workforce

---

15 While responses to questions about specific health conditions or limitations still represent self-reports, the presumption has been that such measures are less susceptible to measurement and endogeneity problems since the questions are narrower and more concrete and, unlike questions about work limitations, are not linked to employment behavior.

16 For a review of the literature on the effects of health on labor supply decisions see Currie and Madrian, 1999.
before normal retirement age. Alternatively these correlations could be interpreted as reflecting the fact that poor labor market prospects induce men to leave the labor force, but that they then rationalize this behavior by identifying themselves as limited in their ability to work.

The literature that has compared results using a variety of different health measures has tended to find that health seems to play a smaller role and economic variables a greater one when the more objective proxies are used. Most authors have interpreted these results as an indication of the biases inherent in using self-reported measures (Parsons, 1982; Anderson and Burkhauser, 1984, 1985; Chirikos and Nestel, 1981; Lambrinos, 1981). These authors have typically either ignored the possible biases inherent in the use of a proxy, or have assumed that these biases are small in comparison to the ones introduced by the use of self-reported measures.

Others have argued in favor of using self-reported information (Burtless, 1987; Sickles and Taubman, 1986). These authors emphasize the flaws inherent in most objective measures of health while pointing to the clinically oriented research supporting the reliability and predictive validity of self-reported health measures (Idler and Benyamini, 1997; Nagi, 1969; Maddox and Douglas, 1973; LaRue et al., 1979; Ferraro, 1980; Mossey and Shapiro, 1982; Manning et al., 1982). These authors ignore the fact that even if self-reported health is a reliable indicator of actual health, this may not be enough to guarantee that it will give sensible results when used as a proxy for health in retirement equations. At issue is whether self-reported health measures are systematically biased, with those out of work being substantially more likely to report health problems than those working. Were this the case, the use of self-reported measures might give misleading information on the reasons why men retire early even if these measures were highly correlated with actual health.

To make these comments precise, we consider a simple model for the labor supply of older men or women. The choice of hours of work, \( y \), depends on the relative rewards of doing so, \( w \), exogenous income (which for simplicity we ignore), unobserved health status, \( \eta \), and other random components:\(^{17}\)\( \varepsilon \):

\[
y = \beta_1 w + \lambda_1 \eta + \varepsilon. \tag{13}
\]

We are interested in consistently estimating \( \beta_1 \) and perhaps \( \lambda_1 \). We expect \( \beta_1 \) to be positive. Since \( \eta \) is unobserved, the sign of \( \lambda_1 \) is arbitrary, but if larger values of \( \eta \) are associated with better health then we would expect that \( \lambda_1 \) should be positive as well.

\(^{17}\)The notational conventions we use in this section are somewhat different than the conventions we use elsewhere. To focus attention on the impact of differential measurement error in our health measure, we are abstracting from potential measurement error in other variables. Thus, the reader should think of \( y \) as representing well measured hours and \( w \) as well measured compensation. On the other hand, \( h \) and \( d \) represent error ridden measures of health, \( \eta \).
We also have an indicator of \( \eta \), self-reported health, \( h \). \( h \) depends on health status \( \eta \), but also on the economic rewards for continuing to work, \( w \), and, again on other random components \( \mu_1 \),

\[
h = \beta_2 w + \lambda_2 \eta + \mu_1
\]

We expect both \( \beta_2 \) and \( \lambda_2 \) to be positive.

We assume that \( \eta \) is orthogonal to both \( \varepsilon \) and \( \mu_1 \) but, as long as there are common unobserved components that affect both \( h \) and \( \nu \), as there will be if the two are definitionally related or if health limitations act as a rationalization for retirement, \( \varepsilon \) and \( \mu_1 \) will be positively correlated.

As long as \( \eta \) and \( w \) are positively correlated, ignoring \( \eta \) in estimating equation (13) will lead to overestimates of the importance of economic incentives in determining labor force participation. The obvious alternative would be to use \( h \) as a proxy for \( \eta \) but there are a variety of econometric problems with doing so. The correlation between \( \varepsilon \) and \( \mu_1 \) introduces a simultaneity bias while variance in \( \mu_1 \) introduces errors-in-variables biases on \( \hat{\lambda}_1 \). Errors in estimates of \( \lambda_1 \) translate into errors in estimates of \( \hat{\beta}_1 \), while the dependence of \( h \) on \( w \) introduces an additional bias on \( \hat{\beta}_1 \). In particular, treating \( \nu \) and \( h \) as if they were observable, letting \( r_{\eta,w} \) represent the correlation between \( \eta \) and \( w \), and \( \rho \) the correlation between \( \varepsilon \) and \( \mu_1 \) and normalizing \( \lambda_2 \) to equal 1, it is easy to show that:

\[
\hat{\lambda}_1 = \frac{\lambda_1 \sigma_\eta^2 (1 - r_{\eta,w}^2) + \sigma_\varepsilon \sigma_\mu_1 \rho}{\sigma_\eta^2 (1 - r_{\eta,w}^2) + \sigma_\mu_1^2}
\]

\[
\hat{\beta}_1 = \beta_1 + (\lambda_1 - \hat{\lambda}_1) \frac{\sigma_{\eta,w}}{\sigma_w^2} - \hat{\lambda}_1 \beta_2
\]

As long as \( \rho > 0 \), this correlation will impart an upward bias on \( \hat{\lambda}_1 \), while \( \sigma_\mu_1^2 \) will impart the standard errors-in-variables downward bias on \( \hat{\lambda}_1 \). Which one dominates depends on the relative strength of these two forces. The bias on \( \hat{\beta}_1 \) will depend both on the bias on \( \hat{\lambda}_1 \) and on \( \hat{\beta}_2 \). Thus, even if the errors-in-variables and the simultaneity biases on \( \hat{\lambda}_1 \) were to cancel, we might still tend to underestimate \( \beta_1 \).

The above expressions make clear that the biases on \( \hat{\lambda}_1 \) and \( \hat{\beta}_1 \) may be quite substantial even when \( h \) is a reliable measure of \( \eta \) (i.e., even when \( \sigma_{\eta,w}^2 \) is quite small). They also make clear that the magnitude and even the direction of the bias depends on the magnitude of several different
correlations. Even if self-reported health is highly correlated with actual health estimates using it as a proxy for health may not give reliable results. Likewise, even if self-reported health often represents rationalization, the use of self-reports may not necessarily exaggerate the role of health in retirement. Beliefs about the kinds of bias involved using self-reported health as a proxy for actual health implicitly reflect judgments about all quantities involved in the above expressions.

Now consider a somewhat more complete model where we have added an equation to make explicit the correlation between \( w \) and \( \eta \) and have some more objective indicator of health status, \( d \), which for concreteness sake we will imagine to be subsequent mortality. We have:

\[
\begin{align*}
y &= \lambda_1 \eta + \beta_1 w + \epsilon \\
h &= \lambda_2 \eta + \beta_2 w + \mu_1 \\
d &= \lambda_3 v + \mu_2 \\
w &= \lambda_4 \eta + \zeta \\
\eta &= v + \xi.
\end{align*}
\]

In this model health, \( \eta \), has two components -- \( \nu \), which influences both longevity and work capacity (e.g., heart problems) and, \( \xi \), which influences only the capacity for work (e.g., arthritis). The implicit assumption imbedded in the variance components formulation (\( \eta = \nu + \xi \)) is that, up to factors of proportionality (\( \lambda_1/\lambda_2 \) and \( \lambda_4/\lambda_2 \)), \( \nu \) and \( \xi \) enter the labor force, health and compensation equations with identical coefficients. This assumption seems a natural one as we are thinking of \( \eta \) as capacity for work, and \( h \) as a self-report on this capacity. \( \epsilon \), \( \mu_1 \) and \( \mu_2 \) are assumed to be uncorrelated with \( w \), while all four errors (\( \epsilon \), the \( \mu \)’s, and \( \zeta \)) are assumed to be uncorrelated with \( \eta \) or its components \( \nu \) and \( \xi \). \( \mu_2 \) is assumed to be uncorrelated with either \( \epsilon \), \( \mu_1 \) or \( \zeta \). These assumptions imply that \( \zeta \) is also uncorrelated with either \( \epsilon \) or \( \mu_1 \). Lastly, \( \nu \) and \( \xi \) are assumed to be uncorrelated with each other. This assumption is mostly definitional -- \( \xi \) is the piece of \( \eta \) that is uncorrelated with \( d \).

\( d \) is objective in two ways that \( h \) is not: \( d \) does not depend directly on \( w \) nor is \( \mu_2 \) correlated with \( \epsilon \). Still, as long as the date of death is not perfectly correlated with an individual's capacity for work, using it as a proxy for health will not adequately control for health, in a regression of \( y \) on \( w \) (and \( d \)). In particular, normalizing \( \lambda_3 \) to equal 1 we have:

\[
\hat{\lambda}_1 = \lambda_1 \frac{\sigma^2 \nu \nu \nu}{\sigma^2 \nu \nu \nu + \sigma^2 \mu_2}.
\]
As long as there are disabling conditions that are not life threatening (e.g., severe back problems, mental illness) controlling for \(d\) will still leave an omitted variable bias on \(\hat{\beta}_1\), while as long as current capacity for work does not perfectly predict date of death there will be errors-in-variables biases on both \(\hat{\lambda}_1\) and \(\beta_1\).

\[
\hat{\beta}_1 = \beta_1 + (\lambda_1 - \hat{\lambda}_1) \frac{\sigma_{v,w} \sigma_{\nu,\lambda}}{2 \sigma_{w}^2} - \hat{\lambda}_1 \frac{\sigma_{\nu,\lambda}^2}{\sigma_{w}^2}
\]

To summarize, using mortality information as a health proxy will tend to underestimate the effects of health and overestimate the effects of economic variables on the labor force participation decision. In contrast, using self-reported health status can either over- or underestimate the impact of either health or economic variables on such decisions.

### 2.4 Bounding Parameter Estimates

While, without some restrictions on the nature of the measurement error, the data puts no bounds on \(\beta\), there has been considerable work done putting bounds on \(\beta\) under the assumption that measurement error is classical. The oldest, and best know of such results is due to Gini (1921). Working with the simple bivariate regression and under the assumption that the errors, \(\nu\) and \(\mu\) are uncorrelated with each other, with \(y^*\) and \(x^*\), and with \(\varepsilon\), it is easy to show that

\[
\frac{1}{\hat{\beta}_{xy}} = \beta \left[ 1 + \frac{\sigma_{v}^2 + \sigma_{y}^2}{\beta^2 \sigma_{x}^2} \right].
\]

Thus

\[
\beta_{yx} \leq \beta \leq \frac{1}{\hat{\beta}_{xy}}.
\]

Under the assumptions of classical measurement error, \(\beta_{yx}\) and \(1/\hat{\beta}_{xy}\) bound \(\beta\), with the tightness of the bounds being a function of the \(R^2\) between \(y^*\) and \(x^*\). More generally, if we allow \(r_{xy,\mu} = 0\) but maintain the other assumptions it is possible to show that as long as the correlation between \(x\) and \(x^*\) is positive \(\beta_{yx}\) will be correctly signed (Weinberg, Umbach and Greenland, 1994).
Under the assumption that only one of the explanatory variables is measured with error, it is easy to generalize Gini’s result to regressions with multiple explanatory variables. On the other hand, in the context in which multiple explanatory variables are all measured with error, the situation is more complex. Klepper and Leamer (1984) derive results under the assumption that the errors are independent of each other and of the unobserved correctly measured variables18.

We start by illustrating Klepper and Leamer’s result within the context of a model with two explanatory variables,

\[ y^* = \beta_1 x_1^* + \beta_2 x_2^* + \varepsilon. \] (20)

For ease of discussion, we will assume that \( x_1^* \) and \( x_2^* \) have been normalized in such a way that \( \beta_1 \) and \( \beta_2 \) are both non-negative. We can imagine several possible “estimates” of \( \beta_1 \) and \( \beta_2 \). The estimates from the direct regression:

\[
\hat{\beta}_1^0 = \frac{\beta_{y|x_1,x_2}}{} \\
\hat{\beta}_2^0 = \frac{\beta_{y|x_1,x_2}}{}
\]

the estimates from the reverse regression of \( x_1 \) on \( y \) and \( x_2 \),

\[
\hat{\beta}_1^1 = \frac{1}{\beta_{x_1^*y|x_2}} \\
\hat{\beta}_2^1 = \frac{\beta_{x_1^*y}}{\beta_{x_1^*x_2}}
\]

and the estimates from the regression of \( x_2 \) on \( y \) and \( x_1 \),

\[
18\text{Some of the results developed by Klepper and Leamer had been developed previously by Koopmans (1937), Reiersol (1945), Dhondt (1960), and Patefield (1981), however Klepper and Leamer’s treatment of these issues is both the clearest and the most complete.}
\]
Klepper and Leamer show that, if all the variables involved are normal, the bounds they derive are tight and that every point within these bounds represents a maximum likelihood estimate of the regression parameters.

Recall that we have normalized $X^*$ in such a way that $\beta \geq 0$. More generally, the condition that Klepper and Leamer (1984) derive implies that the data puts bounds on $\beta$ only if the coefficients from all $k$ possible reverse regressions (regressions of $x_j$ on $y$ and all the other $x$'s) have the pattern of signs as does the original regression of $y$ on $X$.

Klepper and Leamer also show that, if all the variables involved are normal, every vector of parameter estimates with the convex hull represents a maximum likelihood estimate of the model parameters.
Krasker and Pratt (1986) take a different approach. In the context of multiple regression where only one of the variables is measure with error, they ask how highly correlated must the error ridden proxy, $x_j$, be to the unobserved correctly measured variable $x_j^*$ to guarantee that $\beta_{yx;Z}$ will be of the correct sign. No assumptions are made about possible correlations between the error $\mu_j$ and either $y^*$ or any of the elements of $X^*$. Krasker and Pratt show that as long as

$$r^2_{x_jx_j^*} > R^2_{x_j^*,y^*,Z^*} + 1 - R^2_{x_j^*,y^*,Z^*},$$

$\beta_{yx;Z}$ will have the correct sign. For the two variable case (where only one is measured with error) they also derive results for $\beta_{yZ,x_j}$. Here, correlations often have to be quite high to guarantee that estimates will be correctly signed.

### 2.5 Contaminated and Corrupted Data

The measurement error represented in the typical text book and that has received the most treatment in the statistics literature represents “chronic errors” that affect every observation (the error distributions have no mass point at 0). On the other hand, there are situation in which in may be natural to assume that, while in general a variable is well measured, occasional observations are afflicted with potentially gross errors. While, formally speaking, our treatment of measurement error in the proceeding sections encompasses this case, intermittent errors are worth some a bit of attention on their own.

If one has some notion as to the probability that intermittent errors occur, it is often possible to put bounds on the distribution of the variable of interest. Horowitz and Manski (1995) formalize some quite intuitive ideas. HM study the situation in which the researcher is interested in making inference about the marginal distribution of a variable, $y_1$. However, the researcher does not observe, $y_1$, but rather a random variable $y$,

$$y = y_1 z + y_0 (1-z),$$

where, $z$ represents a random variable that take on the value of 1 with probability $p$, 0 with probability $1-p$, and $y_0$ a random variable whose distribution is unknown. HM refer to the case in which $z$ is independent of $y_1$ as “contaminated sampling”, while the case in which this is not true is referred to as “corrupted sampling.”

---

To mention some trivial examples, interviewer errors such as recording that a person was paid 10 dollars per year, rather than per hour, or that person has roughly 10,000, rather than 10 dollars in the bank can lead to occasional gross errors. Imputations for missing data when the researcher is not told which observations include the imputations would be another.
While it is not clear that HM’s ideas can practically be applied, in general, to the regression context, it is possible within specific contexts to apply similar ideas to the estimation of causal parameters (Hotz, Mullin and Sanders, 1997). As HM note, their discussion relates quite closely to discussions within this statistics literature of estimators that are designed to minimize the impact of “contaminated” or “corrupted” data on parameter estimates (Huber, 1981; Hampel, Rousseeuw, and Stahel, 1986). As far as we know Siegal and Hodge (1968) were the first to make this point.

2.6 Measurement Error in Categorical Variables

While, strictly speaking, the analysis presented in the previous sections applies to both continuous and categorical variables, errors in categorical variables are more naturally thought of as classification errors. Thus, for example, if $x^*$ is a dichotomous, 0/1 variable, it seems natural to think in terms of the probabilities of false positives ($\pi_{10} = \text{prob}(x=1|x^* = 0)$) and false negatives ($\pi_{01} = \text{prob}(x=0|x^* = 1)$). In this context, measurement error cannot be classical. If $x^* = 1$, then $x - x^* \leq 0$, while if $x^* = 0$, $x - x^* \geq 0$, so it must be the case that $\sigma_{x^*,\mu} < 0$. Thus, errors in binary variables must be mean reverting. More generally, if $x^*$ has a limited range, as is often the case with the constructs we deal with (e.g., educational attainment) there will be a tendency for $\sigma_{x^*,\mu} < 0$ since when $x^*$ is at the maximum (minimum) of its range, reporting errors can only be negative (positive).
Pr(x=z_j|x^*=z_k, y) = Pr(x=z_j|x^*=z_k), \tag{22}

where z_j \in \{0, 1\}. This is a strong and often implausible assumption. Suppose, for example, that x represents a chronic health condition -- x*=1 if a person suffers from the chronic condition and 0 otherwise. It seems plausible that the severity of a person’s condition will have an effect on the probability that a person recognizes that they suffer from the condition as well as on outcomes. In this case Pr(x=1|x^*=1, y) will be a function of y, and the random error assumption is violated.

At any rate, under the nondifferential measurement error assumption Aigner (1973) shows that:

\[\beta_{yx} = \beta \left[ 1 - Pr(x^*=1|x=0) - Pr(x^*=0|x=1) \right] \]

\[= \beta \left[ 1 - \frac{\pi_{01}}{\pi_{01} + (1-\pi_{10})(1-\pi)} - \frac{\pi_{10}(1-\pi)}{\pi_{10} + (1-\pi_{01})(1-\pi)} \right] \tag{23}\]

where \(\pi\) represents the true fraction of 1's in the population (\(\pi = Pr(x^*=1)\)) and the second line is derived using Bayes rule.\(^{26}\) Since all the \(\pi\)'s lie between 0 and 1, the expression in parenthesis must be less than 1 and \(\beta_{yx}\) will be biased towards 0. In fact, for sufficiently high mis-classification rates (i.e. if \(\pi_{01} + \pi_{10} > 1\), \(\beta_{yx}\) can be wrong signed. Bollinger (1996) has worked out bounds for \(\beta_{yx}\) in this model. Under the assumption that \(\pi_{01} + \pi_{10} < 1\) and the normalization that \(\beta > 0\), Bollinger shows that

\[\beta_{yx} \leq \beta \leq \max \left( [\beta_{yx} \mu_x + \beta_{yx} (1-\mu_x)], [\beta_{yx} \mu_x + \beta_{yx} (1-\mu_x)] \right),\]

where \(\mu_x = Pr(x=1)\). Bollinger also shows how these bounds can be tightened when prior information exists about \(\pi_{01}\) and \(\pi_{10}\).

Classification error in a dependent variable will also typically bias estimates. Take the case where \(y^*\) is a dichotomous, 0/1 variable, and we are interested in estimating \(Pr(y^*=1|x^*)\). We have accurate measures of \(x^* (x=x^*)\) but \(y\) suffers from classification error that is independent of \(x^*\), with \(\pi_{10} = Pr(y=1|y^*=0)\) and \(\pi_{01} = Pr(y=0|y^*=1)\). Since, in this context, the measurement error in the dependent variable is negatively correlated with the accurately measured variable, it should come as no surprise that classification error in a dichotomous dependent variable will tend to bias downward estimates of the effect of \(y^*\) on \(x^*\). In fact, it is easy to see that:

\(^{26}\)If the two kinds of classification error are of the same magnitude (i.e. if \(\pi_{01} \pi = \pi_{10} (1-\pi)\)), then the expression in square brackets in (22) simplifies considerably to \(1-\pi_{10} \pi_{01}\).
\[
\frac{\partial \Pr(y=1|x)}{\partial x} = [1-(\pi_{10}+\pi_{01})] \frac{\partial \Pr(y^*=1|x^*)}{\partial x^*}
\]

(24)

More generally, random misclassification of the dependent variable in a discrete-response setting will bias downwards estimated response functions (Hausman, Abrevaya and Scott-Morton, 1998; Abrevaya and Hausman, 1997).

Categorical variables are often thought of as the discrete indicators of continuous latent variables. Thus, we might imagine that \( y^*=1 \) if \( \xi>0 \) and 0 otherwise. We are interested in estimating \( \Pr(y^*=1|x^*) \), but do not observe \( y^* \). Instead we observe \( y \) where \( y=1 \) if \( \xi+v>0 \) and 0 otherwise. We assume that \( v \) represents normally distributed random “measurement error” in \( \xi \) (i.e., \( v \) is assumed to be independent of both \( \xi \) and \( x^* \)). The probability of classification error in this model depends not just on \( y^* \), but also on \( \xi \) and thus on \( x^* \). To keep things simple, we assume that

\[
\xi = \beta x^* + \epsilon,
\]

(25)

where \( \epsilon \) is a normally distributed, mean 0, random variable. We also assume that \( x^* \) is well measured (\( x=x^* \)). Were we to directly observe \( y^* \), we could consistently estimate \( \beta/\sigma_\epsilon \).

As it is, however, we can consistently estimate only \( \beta/(\sqrt{\sigma_\epsilon^2 + \sigma_v^2}) \). Retrieving \( \beta/\sigma_\epsilon \) requires estimated knowledge of \( \sigma_v^2 \).

Alternatively, imagine that we have a categorical indicator of an latent continuous right hand side variable. Here we imagine the underlying model in terms of the latent variables:

\[
y^* = \beta \eta + \epsilon
\]

(26)

We have a reliable indicator of \( y^* \), \( y(y=y^*) \), but observe only a categorical indicator of \( \eta, x \), where \( x=1 \) if \( \eta>0 \) and 0 otherwise. In this case \( E(y|x=1) = \beta E(\eta|\eta>0) \), while \( E(y|x=0) = \beta E(\eta|\eta<0) \). Thus \( b_{yx} \) consistently estimates \( \beta[E(\eta|\eta>0)-E(\eta|\eta<0)] \). Now suppose \( x \) only imperfectly indicates whether \( \eta>0 \). In particular, we assume that \( x=1 \) if \( \eta+\mu>0 \), where \( \mu \) represents random measurement error. In this case

\[
\beta_{yx} = \beta[E(\eta|\eta+\mu>0) - E(\eta|\eta+\mu<0)]
\]

(27)

\[
< \beta[E(\eta|\eta>0) - E(\eta|\eta<0)].
\]

21
Thus, once again, the use of a noisy explanatory variable tends to lead to an underestimation of the magnitude of the parameter of interest.

2.6 Nonlinear Models

While there is growing literature on the impact of measurement error on parameter estimates within the context of non-linear models, discussions universally occur within the context of specific models. For this reason, it is not possible to summarize results in quite the same way as we were when talking about the linear model. Broadly speaking, the results that do exist suggest that (i) results based on linear models are often approximately true within the context of the non-linear models that have been explicitly studied and (ii) if anything, non-linearities tend exacerbate biases introduced by measurement error.

We have seen that with multiple covariates measured with error, even in the context of the linear model, the effects of measurement error are not easily summarized. On the other hand, in the context of classical measurement error in one variable the bias is always in the form of attenuation. With multiple variables measured with error or if measurement error is not classical, attenuation may not hold.

Weinberg, Umbach and Greenland (1994) study the effect of non-differential measurement error in an explanatory variable within the context of a simple bivariate model, \( f(y^* | x^*) \), where the “dose-response” is monotonic (i.e., \( E(y^* | x^*) \) monotonically increases (decreases) with \( x^* \)). Recall that nondifferential measurement error in \( x^* \) implies that \( f(y^* | x^*, x) = f(y^* | x^*) \). Weinberg, Umbach and Greenland show that as long as \( E(x|x^*) \) increases monotonically with \( x^* \), \( \sigma_{y^*|x^*} \), and \( \sigma_{y^*|x} \) must have the same sign. To paraphrase Weinberg, Umbach and Greenland, as long as the measurement of \( x^* \) is good enough that the population mean of measured “exposure” goes up when true “exposure” does, trend reversal can not occur.

While Weinberg, Umbach and Greenland’s results suggests that in simple models non-differential measurement error of the kind they describe can not cause trend reversal, monotonicity is not necessarily maintained. Hwang and Stefanski (1994) show that even within the context of classical measurement error, it is possible to find situations where the regression of \( y^* \) on \( x^* \), \( E( y^* | x^* ) \), is monotonically increasing (decreasing) in \( x^* \), but that the regression of \( y^* \) on \( x \), \( E( y^* | x ) \) is not.

There is also evidence within the context of specific models that non-linearities tend to exacerbate the magnitude of the bias introduced by measurement error. Griliches and Ringstad (1970) analyzed the situation where \( y^* \) is a quadratic function of \( x^* \):

\[
y^* = \beta_0 + \beta_1 x^* + \beta_2 x^{*2} + \epsilon
\] (28)
$y^*$ is assumed to be well measured ($y = y^*$), but $x^*$ is not ($x = x^* + \mu$). Under the assumption that both $x^*$ and $\mu$ are normally distributed and that $\mu$ is uncorrelated with either $x^*$ or $\varepsilon$, Griliches and Ringstad showed that:

$$
\beta_{y^*x^2} = \beta_1(1-\lambda)
$$
$$
\beta_{y^*x} = \beta_2(1-\lambda)^2
$$
(29)

where, as before, $\lambda = \sigma^2_\mu / \sigma^2_x$. Thus, the coefficient on the quadratic term is more severely biased than is the coefficient on the linear term.

Yatchew and Griliches (1985) derive results for the probit model with one mismeasured explanatory variable. Once again, assuming all variables are distributed normally and that measurement error is classical, they show that simply using $x$ in place of $x^*$ produces estimates that converge to:

$$
\beta = \frac{\sigma^2_x / (\sigma^2_x + \sigma^2_\mu)}{\sqrt{\sigma^2_\varepsilon + \beta^2 \sigma^2_\mu / \sigma^2_x}}
$$
(30)

As is evident from (29), the usual bias towards zero that is present in the linear model is compounded by the term appearing after the plus sign in the denominator.

In the linear model, biases due to measurement error do not depend on whether that error is normal or homoskedastic. However, in non-linear models, this is potentially important, and can induce biases that run counter to our intuitions in the linear case. Consider, for example, a Tobit model

$$
y^* = x^* \beta + \varepsilon
$$
$$
y = y^* \text{ if } y^* > 0
$$
$$
y = 0, \text{ otherwise}
$$

Our explanatory variable $x^*$ is measured with error, and suppose the error $\mu$ is heteroskedastic. Then we can re-write the model for the latent variable $y^*$ as

$$
y^* = x\beta + (-\mu\beta + \varepsilon)
$$
where the “error term” in parentheses is heteroskedastic. Given that heteroskedasticity by itself leads to inconsistent parameter estimates in Tobit models, and can, in plausible cases lead to over-estimating $\beta$ (Maddala, 1983, p. 179), it seems quite possible that heteroskedastic measurement error could lead to upward-biased parameter estimates.

### 3.0 Correcting for Measurement Error

Under the assumption that measurement error is classical, statisticians and econometricians have developed a number of methods to deal with the biases introduced into our estimators when measurement error is present. In particular, under such assumptions, knowing the marginal distribution of the $u_j$’s is sufficient to allow the researcher to undo the biases introduced by measurement error. Alternatively, if one has exogenous determinants of the error ridden explanatory variables or, in some cases, multiple indicators of the same outcome, one can use these as instruments.

We wish to emphasize three points about such general strategies. The first is that these strategies are not as distinct as they might first seem. The second is that these strategies for obtaining consistent estimates of the parameters of interest work if measurement is classical, but do not, in general do so otherwise. Third, even when the correction does not produce consistent estimates, it may produce a bound; and if OLS and IV are biased in different directions or IV is less biased than OLS, this additional information may be very valuable.

### 3.1 Instrumental Variables in the Bivariate Linear Model

---

27 The methods mentioned all involve introducing external information. As long as the measurement error in $X^*$ is classical, and $X^*$, itself, is not normally distributed, $\beta$ is formally identified (Reiersol, 1950; Kapteyn and Wansbeek, 1983). Under the assumption that $X^*$ is not normal a number of authors have suggested instrumental variable estimators that use third or higher moments of the various variables as instruments for $X$ (Geary, 1942; Pal, 1980; Cragg, 1997; Dagenais and Dagenais, 1997; Lewbel, 1997). However, these methods depend crucially on the assumption that $E(y^*|X^*)$ is a strict linear function of $X^*$, and, as such, estimates will be sensitive to specification error. At any rate, such methods have seldom been used in practice. Alternatively, Wald (1940) suggested an estimator of $\beta$ that involved grouping the data. However, unless one has some external information that can be used to form groups (i.e. an instrument), the resulting estimator will typically be no less biased than OLS (Pakes, 1982).

28 The focus of this section is on point estimation. As such, we ignore sampling variability of the various estimators we discuss. In many cases, the estimators are or can be interpreted as instrumental variable estimators. More generally a discussion of the distribution of these estimators can be found in Fuller (1987), Carroll, Ruppert and Stefanski (1995), and Newey and McFadden (1994).
To illustrate these points we will focus on the bivariate linear regression model. To further simplify things, we will also assume that all variables are measured as deviations around their respective means. Thus our model becomes:

\[ y^* = \beta x^* + \varepsilon. \]  \hfill (31)

We assume that we measure \( y^* \) without error (\( y=y^* \)). On the other hand, we have two error ridden indicators of \( x^* \), \( x_1 = x^* + \mu_1 \) and \( x_2 = x^* + \mu_2 \), with \( \mu_1 \) and \( \mu_2 \) uncorrelated with \( x^* \).

Using either \( x_1 \) or \( x_2 \) as proxies for \( x^* \) will lead to estimates of \( \beta \) that are biased towards 0. One alternative would be to use the multiple measures of \( x \) to first gauge the magnitude of the errors and then to correct the bias introduced by these errors. In particular, under the assumptions that \( \mu_1 \) and \( \mu_2 \) are uncorrelated with all the other variables in the system (including each other), \( \sigma_x^2 = \sigma_{x_1x_2}^2 \).

Define

\[ \lambda_1 = \frac{\sigma_{x_1}^2}{\sigma_{x^*}^2 + \sigma_{\mu_1}^2} = \frac{\sigma_{x_1x_2}^2}{\sigma_{x_1}^2} \]  \hfill (32)

where \( \lambda_1 \) represents the signal to total variance ratio for \( x_1 \). Similarly,

\[ \lambda_2 = \frac{\sigma_{x_2}^2}{\sigma_{x^*}^2 + \sigma_{\mu_2}^2} = \frac{\sigma_{x_1x_2}^2}{\sigma_{x_2}^2} \]  \hfill (33)

\( \beta_{yx_1} = \lambda_1 \beta \) and \( \beta_{yx_2} = \lambda_2 \beta \). Under the assumptions of the model, data on \( y, x_1 \) and \( x_2 \) allow one to consistently estimate \( \beta_{yx_1}, \beta_{yx_2}, \lambda_1 \) and \( \lambda_2 \) and thence \( \beta \). In fact, two such estimates are available, giving us some capacity to test the underlying assumptions of the model. In particular, our assumption that \( \mu_1 \) and \( \mu_2 \) are uncorrelated with \( x^* \) and \( \varepsilon \) implies \( \sigma_{x_1\varepsilon} = \sigma_{x_2\varepsilon} \), which is testable.

Alternatively, one might choose to use \( x_2 \) to instrument \( x_1 \):\(^{29}\)

\[ \beta_{y|x_2}^1 = \frac{\sigma_{y|x_2}}{\sigma_{x_1|x_2}}. \]  \hfill (34)

\(^{29}\)Of course, instruments don’t necessarily have to be alternative indicators of \( x^* \). Any variable \( w \), such that \( \sigma_{x^*,w} \neq 0, \sigma_{w,\varepsilon} = 0, \) and \( \sigma_{w,\mu} = 0 \) represents a valid instrument for \( x \).
Notice that $\beta_{iv}^1 = \beta_{yx}/\lambda_2$. Thus using $x_2$ to instrument $x_1$ is equivalent to regressing $y$ on $x_2$ and then using an estimate of $\lambda_2$ to disattenuate the resulting estimate of $\beta$.

Under what circumstances will $\beta_{iv}^1$ represent a consistent estimate of $\beta$? To see, we first write out $\beta_{iv}^1$ in terms of the $x^*$:

$$
\beta_{iv}^1 = \frac{\beta [\sigma^2_{x^*} + \sigma_{x^*,\mu_2}] + \sigma_{\mu_2,e}}{[\sigma^2_{x^*} + \sigma_{x^*,\mu_1}] + \sigma_{x^*,\mu_1} + \sigma_{\mu_1,\mu_2}}. \quad (35)
$$

Thus $\beta_{iv}^1 = \beta$ if $\sigma_{\mu_2,e} = \sigma_{x^*,\mu_1} = \sigma_{\mu_1,\mu_2} = 0$. In other words $\beta_{iv}^1 = \beta$ if $x_2$ is exogenous, the measurement in $x_1$, $\mu_1$ is uncorrelated with $x^*$, and the measurement errors in $x_1$ and in $x_2$ are uncorrelated with each other. These are clearly strong assumptions.

The assumption that $\sigma_{\mu_2,e} = 0$ means that reporting errors in $x_2$ are unrelated to factors other than the $x^*$ affecting $y$. There are circumstances where this assumption may be a sensible one, but others in which it is clearly not. For example, if $x_1$ and $x_2$ represent two self-reported measures of health, and $y$ represents a measure of labor supply, we might expect that reporting (the $\mu$’s) would be correlated with the equation error ($\epsilon$).

The assumption that the two errors in reporting $x^*$ are uncorrelated (i.e., that $\sigma_{\mu_1,\mu_2} = 0$) will also often be open to question. For example, if $x_1$ and $x_2$ represent two reports on $x^*$ taken from the same individual but at different times, it seems likely that the two errors will be positively correlated. Even if $x_1$ and $x_2$ represent two reports on $x^*$ taken from different individuals it will often be possible that the errors will be positively correlated. Thus, for example, two siblings’ reports on their parent’s education will usually both be based on what that parent told the two. If the parent exaggerates their educational attainment (e.g., claims to have finished college, even though she did not), it seems likely that this exaggeration will be common to both siblings’ reports as well as to the parent’s. Moreover, if part of the problem is not simply that individuals inaccurately report $x$, but that our measures do not accurately reflect our constructs (we are interested in human capital, but ask

---

30We have been talking as if $y, x_1$ and $x_2$ all come from the same sample, but what is often the case is that a researcher has only one measure of $x^*$ in the primary data set of interest, but has an estimate of $\lambda$ from some other data set which included multiple measures of $x$. Using an estimate of $\lambda$ based on one sample to correct regression estimates from another is fine as long as one can justifiably interpret the samples as representing similar samples from similar populations.

31Fuller (1987) states these conditions somewhat differently. Using $x_2$ to instrument $x_1$ will consistently estimate $\beta$ if (1) $\sigma_{x^*,e} = 0$ and (2) $\sigma_{x^*,\mu_1} = \sigma_{\mu_1,\mu_2} = 0 \Rightarrow \sigma_{x^*,\mu_1} = 0$. Since $\sigma_{x^*,\mu_1} = 0$ and $\sigma_{\mu_1,\mu_2} = 0$ represent conceptually distinct conditions, we think it makes sense to distinguish the two when discussion conditions for the consistency of the IV estimator.
about educational attainment in years), once again it seems likely that the errors from the separate reports will be positively correlated. In all these situations we expect $\sigma_{\mu_1, \mu_2} > 0$.

Thus, it seems likely that in many situations reporting errors will be positively correlated with each other. The good news here is that, as long as it is true that $\sigma_{\mu_1, x} = \sigma_{\mu_2, x} = 0$, then $\beta_{y_1} \geq \beta_{y_1}^*$ . Thus, correcting for measurement error will tighten our bounds on the true parameter. In addition, with more than two measures of $x^*$ it is possible to begin to relax some, but not all of the assumptions regarding the independence of reporting errors.

Finally, what about the assumption that $\sigma_{x, \mu} = 0$? This assumption is really at the heart of classical measurement error model. There are situations where this assumption seems quite reasonable. Thus, for example, if $x$ represents a sample mean and $x^*$ a population mean, then there may be good reason to believe that $\mu = x - x^*$ is independent of $x^*$. Alternatively, if $x^*$ represents IQ and $x$ the performance on a specific test, then, again, it may be natural to assume that $\mu$ (testing error) is uncorrelated with the truth (here one might want to claim that this is true by construction). However, in the context of survey measurements, there does not seem to be any compelling reason to believe that measurement error is uncorrelated with the truth. Moreover, there are a number of circumstances where it seems likely that reporting errors are negatively correlated with the truth $\sigma_{x, \mu} < 0$. For example, if, as may often be the case, $x$ represents a component of $x^*$, it may be as natural to assume that $\mu$ and $x$ are uncorrelated as it does that $\mu$ and $x^*$ are uncorrelated. Of course, $\sigma_{x, \mu} = 0$ implies that $\sigma_{x, \mu} < 0$.

As we have already mentioned, if $\sigma_{x, \mu} < 0$, then it is no longer necessarily the case that $\beta_{y_1} < \beta$. If it is still true that $\sigma_{\mu, x} = \sigma_{\mu_1, x_2} = 0$, then $\beta_{y_1} \geq \beta$. More generally, if all we know is that $\sigma_{x, \mu} = 0$, then $\beta_{y_1}^*$ could either over or under estimate $\beta$ and exactly the same could be said for $\beta_{y_1}$. Short of some clear notions regarding the nature of measurement error, it is unclear whether standard methods of correcting for biases introduced into our estimates by such errors take us any closer to the truth.

An interesting example of the situation where $\sigma_{x, \mu} < 0$ occurs in the situation discussed above where $x^*$ is dichotomous, and errors are therefore errors of classification. Now suppose one has two
There are situations in which the researcher knows or has estimates of \( \pi_{01} \) and \( \pi_{10} \) from external information. Thus, for example, researchers studying the impact of training programs on the employment and earnings of those trained sometimes do not have an explicit control group, but use nationally representative samples instead. In this context, the control group sample will be “contaminated” with individuals who received training. However, in these situations, the researcher will typically have reliable information on the fraction of the population that receives training, and can use this as an estimate of \( \pi_{01} \). At any rate, in this kind of situation it is reasonably straightforward to derive consistent estimators of the parameters of interest. For a discussion of the case where misclassification occurs in an explanatory variable see Aigner (1973), Freeman (1984), Heckman and Robb (1985) and Heckman, LaLonde and Smith (1999). For the case where the misclassification occurs in the dependent variable, see Poterba and Summers (1986, 1995).

\[ Pr(x_1 = z_j | x^* = z_k, y, x_2) = Pr(x_1 = z_j | x^* = z_k) \]  

(36)

and

\[ Pr(x_2 = z_j | x^* = z_k, y, x_1) = Pr(x_2 = z_j | x^* = z_k). \]  

(37)

In other words, we are assuming the measurement error in \( x^* \) is nondifferential. Here one might be tempted to use \( x_2 \) to instrument \( x_1 \); however, as our discussion above will have made clear, this procedure will tend to produce estimates of \( \beta \) that are too large in magnitude. In fact, it is easy to show that:

\[ \beta_{iv}^1 = \beta \frac{1}{1 - (\pi_{01} + \pi_{10})}, \]  

(38)

which will be greater than \( \beta \) as long as there is any measurement error in \( x_1 \).

However, under the specified assumptions, it is possible to derive consistent estimates of \( \beta \) using GMM methods (Kane, Rouse and Staiger, 1999; Black, Berger and Scott, 1998, also mention this possibility). To see the plausibility that this is the case, it is sufficient to count parameters and moments. The “structural” model includes three parameters: the constant term, the slope coefficient

---

32There are situations in which the researcher knows or has estimates of \( \pi_{01} \) and \( \pi_{10} \) from external information. Thus, for example, researchers studying the impact of training programs on the employment and earnings of those trained sometimes do not have an explicit control group, but use nationally representative samples instead. In this context, the control group sample will be “contaminated” with individuals who received training. However, in these situations, the researcher will typically have reliable information on the fraction of the population that receives training, and can use this as an estimate of \( \pi_{01} \). At any rate, in this kind of situation it is reasonably straightforward to derive consistent estimators of the parameters of interest. For a discussion of the case where misclassification occurs in an explanatory variable see Aigner (1973), Freeman (1984), Heckman and Robb (1985) and Heckman, LaLonde and Smith (1999). For the case where the misclassification occurs in the dependent variable, see Poterba and Summers (1986, 1995).
and the error variance. In addition, there are four distinct error rates as well as the probability that 
\( x^* = 1 \) -- a total of eight parameters in all. With data on \( y, x_1 \) and \( x_2 \), we have 8 independent moments. 
The cross tabulation of \( x_1 \) and \( x_2 \) give us three, the mean of \( y \) conditional on \( x_1 \) and \( x_2 \) gives us four 
more, and the variance of \( y \) gives us one -- eight in all.\(^{33}\)

More generally, if one is working with a linear model that includes categorical variables and if one has 
multiple, error-ridden indicators of such variables where the errors are independent of either the 
outcome or the other explanatory variables in the system, it is possible to get consistent estimates 
of the parameter of the model using GMM techniques (Kane, Rouse and Staiger, 1999).\(^{34}\)

While the assumption that \( \sigma_{\mu_2, \epsilon} = \sigma_{x^*, \mu_1} = \sigma_{\mu_1, \epsilon} = 0 \) is sufficient to identify \( \beta \), it is not 
sufficient to fully identify the model. Counting sample covariances makes this clear. \( \text{Var}(y, x_1, x_2) \) 
contains a total of 6 separate terms. However, even with the stated restrictions, our model includes 
seven distinct parameters (\( \beta, \sigma_{x^*, \epsilon}, \sigma_\epsilon, \sigma_{\mu_1}, \sigma_{\mu_2}, \sigma_{x^*, \mu_1}, \) and \( \sigma_{x^*, \mu_2} \)). In particular, the conditions necessary 
for the consistent estimation of \( \beta \) are not sufficient to allow us to separately identify \( \sigma_{x^*, \epsilon}, \sigma_\epsilon, \sigma_{\mu_1}, \) and \( \sigma_{x^*, \mu_2} \). The IV estimator allows us to solve both the pure errors in variable and the endogeneity 
problems associated with the use of \( x_1 \) as a proxy for \( x^* \), but does not allow us to separate out these 
two effects. If, in addition to the assumptions we have already made, we assume that \( \mu_2 \) is 
uncorrelated with \( x^* (\sigma_{x^*, \mu_2} = 0) \), or that \( \mu_1 \) is uncorrelated with \( \epsilon (\sigma_{\mu_1, \epsilon} = 0) \) then the model is fully 
identified.

As Goldberger (1972) and Griliches (1974; 1986) have emphasized, it is often also possible to 
consistently estimate errors in variables models in a multi equation setting. We illustrate with an 
extremely simple model. Suppose

\[^{33}\text{Kane, Staiger and Rouse’s work echoes earlier work Goodman (1974a, 1974b), Haberman (1977), Andersen (1982) and others on what Goodman refers to as latent structural models. Goodman showed that in a context in which one observed multiple independent discrete indicators of a (latent) discrete random variable it was often possible to identify the distribution of both the underlying latent variable and the transition matrices that stochastically map the latent variable into observable indicators. The correspondence between latent structural models and the model proposed by Kane, Rouse and Staiger is remarkably close. However, the models that Goodman and his colleagues worked with involve solely discrete variables and have been mostly ignored by economists.}\]

\[^{34}\text{It is worth noting that the discussion has been of models in which } x^* \text{ is, itself, categorical. Such models need to be distinguished from models in which } x^* \text{ is conceptualized as continuous (e.g. health status), but we have only categorical indicators of } x^*. \text{ If } x \text{ represents an error ridden categorical indicator of } x^* (\text{i.e. if } x = k \text{ iff } c_{k-1} < x^* + \mu \leq c_k) \text{ there may be no particular reason to believe that } \mu \text{ is correlated with } x^*. \text{ In fact, in this case, the models are linear in latent variables. For this reason, the intuitions and insights obtained from work on the linear errors in variables model still holds. The case where } x \text{ represents a categorical indicator of an underlying continuous variable has been extensively analyzed (e.g. Heckman, 1978; Lee, 1982a, 1982b; Muthen, 1983).}\]
35Accurately measured $x$’s can be included as elements of $W$. ($39$)

The error terms (the $\varepsilon$’s and $\mu$) are assumed to be uncorrelated with each other and with $x^*$. Under these assumptions, $\beta_1$ can be consistently estimated by using $y_2$ as an instrument for $x$ in the regression of $y_1$ on $x$ ($\beta_1 = \text{cov}(y_1, y_2)/\text{cov}(x, y_2)$). $\beta_2$ can be estimated in a similar fashion. Chamberlain and Griliches (1975) used more sophisticated multi-equation models to control for ‘ability’ when estimating the effect of education on earnings. However, as Griliches has emphasized, estimates based on such models are only as good as the models themselves. In this kind of setting, minor specification errors can have significant effects on parameter estimates.

### 3.2 Multivariate Linear Model

The methods we have been discussing generalize to the multivariate case. Suppose, for example, one is willing to assume that errors in both the outcome and the explanatory variables ($\nu$ and $\mu$) are uncorrelated with both the actual (accurately measured) outcome and the explanatory variables and that one has prior knowledge of their joint distribution, then

$$\hat{\beta} = (S_{XX} - \hat{\Sigma}_{\mu,\mu})^{-1}(S_{XY} - \hat{\Sigma}_{\mu,\nu})$$

(40)

will consistently estimate $\beta$, where $S_{XX}$ represents the sample variance of $X$, $S_{XY}$ the sample covariance of $X$ and $y$, $\hat{\Sigma}_{\mu,\mu}$ a consistent estimate of the variance of the $\mu$’s and $\hat{\Sigma}_{\mu,\nu}$ a consistent estimate of the covariance between $\mu$ and $\nu$.

Alternatively, if one has as many instruments ($W$’s) as one has as one has explanatory variables ($X$’s)$^{35}$, $\nu$ is uncorrelated with $y^*$, and $\sigma_{W,\mu} = \sigma_{W,\nu} = \sigma_{W,\varepsilon} = 0$ then the IV estimator,

$$\beta_{iv} = [W'X]^{-1}W'y$$

(41)

consistently estimates $\beta$. Of course, if the assumptions are violated and $W$ is correlated with $\mu$, $\nu$ or $\varepsilon$, $\beta_{iv}$ will be inconsistent. One special case is worth noting. Take the situation where only one element of $X$ is measured with error (denote this variable as $x$) while the rest are accurately measured (denote this vector as $Z^*$). We are interested in estimating the equation:

---

$^{35}$Accurately measured $x$’s can be included as elements of $W$. 30
Following the example of section 2.3 in detail, with two indicators of \( \eta \) we might be tempted to use one to instrument the other, but this will not work. As long as \( \beta_2 \neq 0 \) using \( d^* \) to instrument \( h^* \) will purge \( h^* \) of its dependence on \( \varepsilon \) and so will correctly estimate \( \lambda_1 \) but will tend to underestimate \( \beta_1 \) by \( \beta_2 \lambda_1 \). The intuition that we should be able to use \( d^* \) to instrument \( h^* \) arises from the similarity of this model to the classical errors-in-variables model, in which one error prone measure can be used to instrument another. This model differs from the classical errors-in-variables model in that the endogeneity of \( h^* \) causes the error in this indicator to be correlated with the other regressor in the model, \( w \). The instrumental variable procedure uses the projection of \( h^* \) onto \( w \) and \( d^* \) as a proxy for \( \eta \). What we need, instead, is the projection of \( \eta \) on \( w \) and \( d^* \). With \( h^* \) as the dependent variable, the estimated coefficient on \( w \) will reflect not only the errors in \( d^* \) but also \( w \)'s direct effect on \( h^* \), \( \beta_2 \). This, in turn, will induce the downward bias on \( \beta_1 \) of \( \beta_2 \lambda_1 \). We could sort all of this out if we had a consistent estimate of \( \beta_2 \), but this requires either knowledge of the reliability of \( d^* \) as a proxy for \( \eta \) or another indicator of \( \eta \). Thus, using mortality information to instrument self-reported disability status

\[
y^* = \beta x^* + Z^* \gamma + \varepsilon
\]
3.3 Nonlinear Models

Correcting for the bias created by errors in variables is more difficult in non-linear than in linear models. Typically, instrumental variable methods work well only when errors are relatively small in magnitude (Amemiya, 1985, 1990). Thus, for example, if one is interested in estimating the non-linear model,

\[ y = g(x^*; \theta) + \varepsilon, \]

where we assume that \( \varepsilon \) is independent of \( x^* \), and that \( \theta \) is a parameter vector. We observe a proxy for \( x^*, x \), where \( \mu = x - x^* \) is independent of \( x^* \). We also have available instruments, \( w \), that are correlated with \( x^* \), but are independent of both \( \mu \) and \( \varepsilon \). We might imagine trying to estimate \( \theta \) by non-linear instrumental variables (Amemiya, 1974). However, if \( g \) is non-linear not just in parameters, but in variables, this procedure will not consistently estimate \( \theta \) (Amemiya, 1985, 1990; Hsiao, 1989).

For linear models there is a close tie between simultaneous equations and errors in variables models. However, for non-linear models, the analogy breaks down. To see why, imagine that \( x^* \) is a linear function of \( w \), \( x^* = \pi w + \nu \), with \( \nu \) orthogonal to \( w \) by construction. For the linear model we have:

\[ y = x^\prime \beta + \varepsilon \]

\[ = \pi w \beta + \beta \nu + \varepsilon. \]

\( \beta \nu \) is orthogonal to \( w \), so using \( \pi w \) in place of \( x^* \) will consistently estimate \( \beta \). For the nonlinear model we have:

\[ y = g(x^*; \theta) + \varepsilon \]

\[ = g(\pi w; \theta) + [g(x^*; \theta) - g(\pi w; \theta)] + \varepsilon. \]

will correctly estimate the impact of health but tend to underestimate the impact of economic variables on such decisions. In contrast, using mortality information alone to construct a health proxy will tend to underestimate the effects of health and overestimate the effects of economic variables on the labor force participation decision, while using self-reported health status can either over- or underestimate the impact of either health or economic variables on such decisions (Bound, 1991).
Amemiya (1985) and Hsiao (1989) give more formal versions of this argument. Simulation techniques can greatly facilitate such estimation (Lavy, Palumbo and Stern, 1998; Stinebrickner, forthcoming).

Amemiya studies the asymptotic behavior of the nonlinear instrumental variables estimator as $\sigma_\mu^2$ converges to 0, and finds that with standard regularity conditions, the estimator will not, in general, be a linear function of $\nu$ and thus there is no guarantee that it will be orthogonal to $g(\pi w; \theta)$.

In general, consistent estimation of non-linear errors-in-variables models requires the researcher to know or be able to consistently estimate the conditional distribution of $x^*$ given $x$, $f(x^*|x;\delta)$. With $f$ known, the mean of $y$ conditional on $x$ becomes:

$$E(y|x) = \int g(x^*; \theta) f(x^*|x;\delta) dx^* = G(x;\gamma),$$  \hspace{1cm} (46)

where $\gamma=(\theta,\delta)$. Substituting $G(x; \gamma)$ for $g(x^*; \theta)$, we obtain a model in terms of observables:

$$y = G(x;\gamma) + \nu$$ \hspace{1cm} (47)

where

$$\nu = \epsilon + g(x^*; \theta) - G(x;\gamma).$$ \hspace{1cm} (48)

By construction $E(\nu|x) = 0$. In principle this model can be estimated by maximum likelihood. Hsiao (1989) proposed computationally simpler minimum distance and two step estimators of the model. Alternatively, one can imagine using multiple imputation techniques (Rubin, 1987; Little and Rubin, 1987; Brownstone, 1998) to first impute estimates of $x^*$ and then use these in a second stage to estimate $\theta$.

The availability of an instrument, $w$, is not sufficient to allow the researcher to estimate the distribution of $x^*$ conditional on $x$ (or $w$, for that matter). We have $x = x^* + \mu = \pi w + \nu + \mu$. The regression of $x$ or $w$ allows us to consistently estimate $\pi$, but not the distribution of $\nu$. Thus, this first stage regression does not allow us to identify the distribution of $x^*$ conditional on $w$. Without knowledge of the distribution of $x^*$ conditional on observables, it is not possible to consistently estimates $\theta$. However, the estimator that simply uses $\hat{\pi} w$ as a proxy for $x^*$ often works well (Amemiya, 1985; Carroll and Stefanski, 1990) as an approximation.

---

37Amemiya (1985) and Hsiao (1989) give more formal versions of this argument.

38Simulation techniques can greatly facilitate such estimation (Lavy, Palumbo and Stern, 1998; Stinebrickner, forthcoming).

39Amemiya studies the asymptotic behavior of the nonlinear instrumental variables estimator as $\sigma_\mu^2$ converges to 0, and finds that with standard regularity conditions, the estimator
3.4 The Contribution of Validation Data

So far we have been discussing approaches to measurement error that use multiple, possibly error ridden, indicators of the key variables we are interested in to gauge the reliability of these measures. As we have seen, estimates of the reliability of key measures can be used to gauge the effect of measurement error on our estimates under the assumption that measurement error is, in one way or another, independent of the constructs that enter our models. An alternative is to compare the survey estimate with other, more accurate empirical data. The promise of validation studies is that they give some direct evidence on the nature of the measurement error in survey data, by allowing comparison of survey responses to “true” values if the same variables. Often, the “true” values are obtained from employer or administrative records. Thus, \( X^* \) will be referred to as the “record” data.

Consider first the simplest case, where the required validation data is quite modest. Suppose we wish to consistently estimate the effect of a single explanatory variable, \( x^* \), on \( y^* \), but our survey measure for \( x^* \) is measured with error. If the error is classical we know \( \beta_{yx} \). Data from a validations study, which includes both the survey response, \( x \), and an accurate measure of \( x^* \), (for example, based on checking reliable administrative records) can give us estimates of \( \sigma^2_\mu \) or \( \sigma^2_\varepsilon / \sigma^2_x \), which can be used to correct the estimate based on the original survey data. Even better, we could not assume the measurement error is classical; as long as it is uncorrelated with \( y^* \), we know that \( \beta_{yx} = \beta (1 - \beta_\mu \mu) \). The validation data allows us to estimate \( \beta_\mu \mu \) directly.

More ambitiously, validation data allows us to identify parameter estimates in the presence of arbitrary patterns of measurement error. Suppose that we have error ridden data for a random (primary) sample of the population. For a distinct random sample of the population we have validation data. We are imagining that this validation data contains both the error ridden and error free data. We can then use the validation data to compute the distribution of \( y^*, X^* \) given \( y, X \) (\( f(y^*, X^* | y, X) \)). This conditional distribution can then be used to impute the distribution of \( y^* \) and \( X^* \) in the primary data set. What is clearly crucial for such a procedure to be valid is that the distribution of \( y^*, X^* \) given \( y, X \) be the same in the primary and validation data set (Carroll, Ruppert and Stefanski, 1995, refer to this as transportability).

To be somewhat more concrete within the context of the linear model, validation data allow us to calculate empirical analogues to \( \beta_{\mu X}, \beta_{eX}, \beta_{eX}, b_{\mu X}, b_{eX}, \) and \( b_{eX} \). Assume to begin with that one’s measure of \( y \) in the primary data set is error free and that \( X \) is exogenous (\( \beta_{eY} = 0 \)). Also let \( \beta_{X^*, X} \) represent the matrix of regression coefficients from the regression of \( X^* \) on \( X \) in the validation sample (\( \beta_{X^*, X} = I + \beta_{\mu X} \)). A consistent estimate of \( \beta \) can be obtained by first using \( \beta_{X^*, X} \) calculated in the validation sample to transform \( X \) in the primary sample, \( \hat{X} = \beta_{X^*, X} X \), and then regressing \( y \) on \( \hat{X} \). Note that under these circumstances consistent estimation of \( \beta \) requires validation data on \( X \), but does not require validation data on \( y \). In fact, as the expressions make clear, the validation data on \( X \) can come from a separate sample that contains no information on \( y \), as long as both the primary sample and the validation sample are random samples from the same population.

\( \sigma^2_\mu \) approaches consistency as \( \sigma^2_\mu \) approaches 0.
More generally, if $\beta_{vX} \neq 0$ and $\beta_{v\varepsilon} \neq 0$, then one can obtain consistent estimates of $\beta$ by transforming $y$ as well as $X$. Let $\hat{y} = y - [\hat{\beta}_v + \hat{\beta}_\varepsilon] X$. Then

$$\hat{\beta} = [\hat{X}'\hat{X}]^{-1}\hat{X}'\hat{y} \quad (49)$$

consistently estimates $\beta$.\(^{40}\)

Lee and Sepanski (1995) generalize (43) to the nonlinear context. They consider the nonlinear regression:

$$y^* = g(x^*, \theta) + \varepsilon. \quad (50)$$

In the primary data set, the researcher has a random sample of error-ridden versions of $y^*$ and $x^*$, which, following our general notation, we will refer to as $y$ and $x$. The researcher also has available a validation data set that contains a random sample of both accurately measured and error ridden versions of $y^*$ and $x^*$, $y_v^*$, $x_v^*$, $x_v^*$, $y_v^*$, where the $v$ subscript is used to indicate the data come from the validation data.

Consider first the case where either $y^*$ is accurately measured ($y=y^*$) or where measurement error in $y$ is classical and so can be absorbed in the error term and where the measurement error in $x^*$ is nondifferential. Lee and Sepanski (1995) propose an estimator of $\theta$ that minimizes:

$$\hat{\theta} = \min_0 \left[ y - x (x_v^* x_v^*)^{-1} x_v^* g(x_v^*; \theta) \right]^2. \quad (51)$$

They show that under standard regularity assumptions, $\hat{\theta}$ consistently estimates $\theta$ and derive its asymptotic distribution. In the context where $y^*$ suffers from non-classical measurement error or where the measurement error in $x^*$ is differential, (51) can be modified to consistently estimate $\theta$. Define $w=[y:x]$ and $\hat{y} = y - w'(w_v w_v)^{-1} w_v (y_v - y_v^*)$. Then

$$\hat{\theta}' = \min_0 \left[ \hat{y} - x (x_v^* x_v^*)^{-1} x_v^* g(x_v^*; \theta) \right]^2 \quad (51')$$

will consistently estimate $\theta$.

\(^{40}\)These ideas are developed formally and generalized to the non-linear setting in Lee and Sepanski (1995).
Measurement error in key variables can be thought of as a special case of missing data -- in a literal sense the researcher is missing valid data on the variables measured with error. Much of the voluminous literature on handling missing data has focused on the case where data are missing for a subset of the data. Within the context of measurement error this is akin to having validation data available. Thus, the techniques that have been developed to deal with missing data (Little, 1992) could be applied to estimating models with error ridden data as well.41

In the general context, the impact of measurement error on parameter estimates is model dependent. As we have seen, within the context of the linear model, the impact will depend on the association between the measurement error in the key variables and all the other variables included in a model. More generally, one needs to be able to estimate $f(y^*, X^* | y, X)$, where $y^*$ and $X^*$ include all the variables of interest. Thus, the value of validation studies is enhanced if they include not just data on the key variables being validated but also on other variables that researchers would typically use in conjunction with these variables.

Validation studies report information regarding the magnitude of the measurement error involved in survey measures -- typically the mean and some measure of the dispersion in the measure. Correlations between the survey and validation study measures of the same variable will also often be reported and can be thought of as measures of the validity of the survey measures (the validity of a measure is the correlation between the measure and the actual underlying construct that the measure is intended to be a measure of). While information on the marginal distributions of the error is sufficient to allow researchers to use such studies to estimate the impact of measurement error on parameter estimates if measurement error is classical, our discussion should make clear that one of the real values of a validation study is to allow us to relax such assumptions. Studies sometimes report not only summary statistics but also sample regressions. However, even these regressions will provide information regarding the impact of measurement error on estimates only for models similar to the ones reported on in the validation study report. Perhaps, such tabulations should be seen as illustrative. While, in general, it will not make sense or be possible to report $f(y^*, X^* | y, X)$, it will often be possible to make the validation study data available to researchers thus allowing individual researchers to study the impact of measurement error on whatever kind of model they are interested in estimating. Indeed some of the most interesting results in the literature using validation studies have been done by individuals who were not originally involved in collecting the validation data but who use such data to examine the impact of measurement error on parameter estimates within the context of a specific research question.

While validation data has considerable promise, it is important to bear in mind the limitations of such data as well. We have in mind two distinct issues. First validation data presumably has higher validity than survey measures -- indeed the very value of such data depends on this presumption -- however this does not mean that it is completely without error. Even administrative data or payroll records will include errors. Equally important, validation data may not tap exactly the

41See Carroll, Ruppert and Stefanski (1995) for a discussion of the link between missing data and measurement error models. See Brownstone and Valletta (1996) for the implementation of these ideas within the context of an economic example.
same construct as does the survey measure and some of the discrepancies between the survey and validation data measures may involve discrepancies between the constructs the two capture. Neither the survey measure nor the validation study measure may adequately capture the construct we are interested in.

Second, validation study data collected in one context, may not generalize to another. In some contexts the issues are obvious. Thus, for example, data collected from a single firm may not be that informative about the nature of measurement error in nationally representative data both because of idiosyncracies regarding the firm and because the data misses any between firm variation. In other cases, issues are more subtle. Existing methodological work (see section 5) suggests that for many items the extent of measurement error will be context dependent. For example, the extent of measurement error in reported earnings and employment status appears to depend on the business cycle (see section 6.1).

4.0 Approaches to the Assessment of Measurement Error

In order to use the procedures outlined in section 3, one needs either data that includes multiple indicators of variables measured with error or validation data that includes both accurate and error ridden versions of the analysis variables. As the above discussion should make clear, the use of multiple measures to correct for biases introduced by measurement error requires the use of strong assumptions about the nature of the measurement error involved. What is nice about validation data is that it allows the researcher to relax such assumptions.

Multiple indicator or validation data is sometimes collected as part of the primary data collection effort. However, more commonly, such data comes from external independent studies. It should be clear that internal multiple indicator or validation data is to be preferred over external data. With the use of external data one is always making an assumption about the transportability of models from the external to the primary data.

Most of the research involving validation data incorporates one of two designs: (1) obtaining external data for the individuals included in the survey or (2) comparing external population based parameters or estimates with those derived from the survey. We examine empirical studies that encompass four separate approaches to the assessment of the quality of household reported economic phenomena:

42Carroll, Ruppert and Sepanski (1995) emphasize the value of having validation data collected on a random sub-samples of the primary data (“We cannot express too forcefully that if it is possible to construct an internal validation data set, one should strive to do so. External validation can be used ... but one is always making an assumption when transporting such models to the primary data.”). Validation data collected as a subsample of the primary data is practically nonexistent in the data typically used by economists, but such data has sometimes been collected in other contexts (see the examples discussed by Carroll, Ruppert and Sepanski).
i. Validation studies which involve micro-level comparisons of household-reported data with external measures of the phenomena, such as employer’s records or administrative records;

ii. Micro-level comparisons of response variance which involve the comparison of individual survey respondents’ reports at time t with reports obtained at time t+x, under the same essential survey conditions;

iii. Micro-level comparisons of response differences involving the comparison of the individual survey respondents’ reports at time t with reports obtained at time t +/-x, involving either administrative records (e.g. comparison to tax returns) or the collection of survey data under different (and supposedly preferred) survey conditions; and

iv. Macro-level comparisons of estimates based on survey reports with aggregate estimates generated under different (and supposedly preferred) survey conditions or from aggregate administrative records.

Each of these approaches to the assessment of data quality suffers from potential limitations; these limitations are outlined in the discussion that follows.

Validation studies which permit micro-level comparisons can be classified as one of three types of studies: (1) a reverse record check, in which elements are sampled from the administrative (or validation) records and then interviewed; (2) prospective record checks in which elements are interviewed and then administrative records are checked to confirm the reported behaviors; or (3) complete record check studies, in which all elements in the population have a probability of selection into the sample and administrative records or other validation information are obtained for all sampled elements, regardless of whether the behavior of interest has been reported or not. If the measure of interest is a discrete event (e.g. hospitalization, industrial accident related to a particular job), reverse record check studies are quite adequate in measuring underreporting, but are often insensitive to overreports, since the administrative records may not include the complete universe of events. Prospective record check studies attempt to verify affirmative survey responses; thus these designs are better for assessing overreporting of discrete events but less adequate than reverse record check studies for assessing underreporting of events since it may be difficult to obtain validation data from all potential sources. A complete or full record check, provided all of the relevant records can be located, provides the best means for assessing both underreporting as well as overreporting. However, such studies are rare, requiring a closed universe from which one can obtain the validation information and be confident that the records include an accounting for the entire universe of behaviors.

Regardless of the design of the validation study, most empirical investigations incorporating validation data attribute differences between the respondent report and the validation data to the respondent and thus may overstate the level of response error. There are two separate issues here. First, various factors may contribute to measurement error, including the interviewer, the wording of a particular question, the context of the questionnaire, as well as the essential survey conditions such as
the mode of data collection; however, differences between survey reports and administrative records are often discussed in terms of response error. Recognizing the alternative sources of errors is a first step in modeling them properly. Second, as noted above, differences between respondent reports and the validation data may reflect deficiencies in the latter. Most record check studies fail to assess or even discuss the level of potential error in the records or the error introduced via the matching of survey and record reports. Comparisons of survey reports with self-reported administrative records (e.g. tax records) may show discrepancies because of errors in the administrative records. Finally, it is rare to see a thorough discussion of the impact of definitional differences between the two sources of information on the level of apparent error.

In contrast to most validation studies, micro-level comparisons of survey reports for discrete events occurring before time t obtained at two points in time under the same essential survey conditions focus on simple response variance over time. However, the accuracy of the data at either time t or time t+x can not be assessed. Empirical investigations of this type usually attribute differences in the two estimates to error in the reports obtained at time t+x (under the assumption that the quality of retrieval declines over time).

Micro-level comparisons which entail survey estimates produced as a result of different survey designs similarly tend to attribute differences in the estimates to response error for the estimates produced under the less optimal design. Hence, the later comparison requires a priori knowledge of the design most likely to produce the most accurate data.

Macro-level comparisons are fraught with several potential confounding factors, including differences in the population used to generate the estimates, definitional differences, and differences in the reference period of interest. Benchmark data are themselves potentially subject to various sources of errors and omissions, complicating the interpretation of differences between the two sources. Finally, whereas micro-level validation can compare survey responses to external data, comparisons of survey data with aggregate benchmark data requires some assumptions about non-response, either reweighting the available responses or imputing values for non-respondents. Perfectly accurate survey responses can appear to diverge from benchmark totals if the handling of non-respondents is in error; incorrect survey responses could even add to correct control totals if response error and errors in nonresponse corrections are offsetting.

5.0 Measurement Error and Memory: Findings from Household-Based Surveys

The assessment of measurement error across various substantive disciplines has provided a rich empirical foundation for understanding under what circumstances survey responses are most likely to be subject to measurement error. The theoretical framework for most of these investigations draws from the disciplines of cognitive and social psychology. Although these investigations have provided insight into the factors associated with measurement error, there are few fundamental principles which inform either designers of data collection efforts or analysts of survey data as to the circumstances, either individual or design-based, under which measurement error is most likely to be significant or not. Those tenets which appear to be robust across substantive areas are outlined below.
5.1 Cognitive Processes

Tourangeau (1984) as well as others (see Sudman, Bradburn, and Schwarz, 1996 for a review) have categorized the survey question and answer process as a four-step process involving comprehension of the question, retrieval of information from memory, assessment of the correspondence between the retrieved information and the requested information, and communication. In addition, the encoding of information, a process outside the control of the survey interview, determines a priori whether the information of interest is available for the respondent to retrieve from long-term memory.

Much of the measurement error literature has focused on the retrieval stage of the question answering process, classifying the lack of reporting of an event as retrieval failure on the part of the respondent, comparing the characteristics of events which are reported to those which are not reported. One of the general tenets from this literature concerns the length of the recall period; the greater the length of the recall period, the greater the expected bias due to respondent retrieval and reporting error. This relationship has been supported by empirical data investigating the reporting of consumer expenditures and earnings (Neter and Waksberg, 1964); the reporting of hospitalizations, visits to physicians, and health conditions (e.g., National Center for Health Statistics, 1961, 1967; Cannell, Fisher, and Bakker, 1965; Woolsey, 1953); reports of motor vehicle accidents (Cash and Moss, 1972), crime (Murphy and Cowan, 1976); and recreation (Gems, Ghosh, and Hitlin, 1982). However, even within these studies the findings with respect to the impact of the length of recall period on the quality of survey estimates are not consistent. For example, Dodge (1970) found that length of recall was significant in the reporting of robberies but had no effect on the reporting of various other crimes, such as assaults, burglaries, and larcenies. Contrary to theoretically-justified expectations, the literature also offers several examples in which the length of the recall period had no effect on the magnitude of response errors (see for example, Mathiowetz and Duncan, 1988; Schaeffer, 1994). These more recent investigations point to the importance of the complexity of the behavioral experience over time, as opposed to simply the passage of time, as the factor most indicative of measurement error.

Another tenet rising from the collaborative efforts of cognitive psychologists and survey methodologists concerns the relationship between true behavioral experience and retrieval strategies undertaken by a respondent. Recent investigations suggest that the retrieval strategy undertaken by the respondent to provide a “count” of a behavior is a function of the true behavioral frequency. Research by Blair and Burton (1987) and Burton and Blair (1991) indicate that respondents choose to count events or items (episodic enumeration) if the frequency of the event/item is low and they rely on estimation for more frequently occurring events. The point at which respondents switch from episodic counting to estimation varies by both the characteristics of the respondent as well as characteristics of the event. As Sudman et al. (1996) note, “no studies have attempted to relate individual characteristics such as intelligence, education, or preference for cognitive complexity to the choice of counting or estimation, controlling for the number of events” (p. 201). Work by Menon (1994) suggests that it is not simply the true behavioral frequency that determines retrieval strategies, but also the degree of regularity and similarity among events. According to her hypotheses, those events which are both regular and similar (brushing teeth) require the least amount of cognitive effort to report, with respondents relying on retrieval of a rate to produce a response. Those events which
The work on telescoping has focused on the effect of telescoping on the time of individual events. However, it seems likely that when respondents are asked to retrospectively recall the timing of various events in their past, errors in the reported timing of various events are correlated, creating something of a spurious coincidence of events. This is a potentially serious issue for event history analysis.

The impact of different retrieval strategies with respect to the magnitude and direction of measurement error is not well understood; the limited evidence suggests that errors of estimation are often unbiased, although the variance about an estimate (e.g., mean value for the population) may be large. Episodic enumeration, however, appears to lead to biased estimates of the event or item of interest, with a tendency to be biased upward for short recall periods and downward for long recall periods. In part, the direction of the estimation error related to episodic enumeration is a function of the misdating of the dates of retrieved episodes of behavior, a phenomenon referred to in the literature as telescoping (e.g., Sudman, et al. 1996). The evidence for telescoping comes from studies which have examined respondent’s accuracy in reporting dates of specific events. Forward telescoping refers to the phenomena in which respondents report the occurrence of an event as more recent than is true; backward telescoping refers to misdating in the opposite direction, that is, reporting the event as occurring earlier in time than is true. The direction of the misdating appears to be a function of the length of the reference period. Forward telescoping is most evident when the reference period is short (one or two weeks), whereas backward telescoping is more common for longer (one year or more) reference periods.

The misdating of episodic information in panel data collection efforts has given rise to a particular type of response error referred to as the “seam effect” (Hill, 1987). Seam effects refer to the phenomena in a disproportionate number of changes in respondent status (e.g., employment status) change at the “seam” between the end of the reference period for wave x of a study and the start of the reference period for wave x+1 of a study. For example, a respondent will report being employed at the time of the wave x interview; at wave x+1, the respondent reports being unemployed for the entire reference period. Hence his or her change in employment status occurred at the seam of the reference periods. Although the seam effect may arise as a function of the misdating of the start or end of a particular status, some have speculated that the effect is a result of respondents minimizing the level of effort associated with the respondent task by projecting the current status back to the beginning of the reference period of interest.

Finally, a third tenet springing from this same literature concerns the salience or importance of the behavior to be retrieved. Salience is hypothesized to affect the strength of the memory trace and subsequently the effort involved in retrieving the information from long-term memory. The stronger the trace, the lower the effort needed to locate and retrieve the information. In a study on the reporting of hospitalizations, Cannell, Fisher, and Bakker (1965) found that hospitalizations of longer duration were subject to lower levels of errors of omission than hospitalizations of one or two days in length; Waksberg and Vallient (1978) report a similar pattern with respect to injuries. Although

The work on telescoping has focused on the effect of telescoping on the time of individual events. However, it seems likely that when respondents are asked to retrospectively recall the timing of various events in their past, errors in the reported timing of various events are correlated, creating something of a spurious coincidence of events. This is a potentially serious issues for event history analysis.
salient information may be subject to lower levels of errors of omission, other research has indicated that salience may lead to overestimation on the part of the respondent (e.g. Chase and Harada, 1984). As is evident from the literature, overestimation or overreporting on the part of the respondent can result from either forward telescoping of events, that is, the misdating the event of interest counting events which occurred prior to the start of the reference period, or from misestimation, in part, due to the salience of the event of interest. Unfortunately, empirical investigations of response error in which overreporting is evident have not addressed the relative importance of forward telescoping and salience as the source of the response error.

5.2 Social Desirability

In addition to asking respondents to perform the difficult task of retrieving complex information from long term memory, survey instruments often ask questions about socially and personally sensitive topics. It is widely believed and well documented that such questions elicit patterns of underreporting (for socially undesirable behavior and attitudes) as well as overreporting (for socially desirable behaviors and attitudes). The determination of social desirability is a dynamic process, a function of the question topic, the immediate social context, and the broader social environment at the time the question is asked. Some topics are deemed, by social consensus, to be too sensitive to discuss in “polite” society. In the 1990s this is a much shorter list than was true in the 1950s, but most would agree that topics such as sexual practices, impotence, and bodily functions fall within this classification. Some hypothesize that questions concerning income also fall within this category (e.g. Tourangeau, Rips, and Rasinski, 1999). Other questions may concern topics which have strong positive or negative normative responses (e.g. voting; the use of pugnacious terms with respect to racial or ethnic groups) or for which there may be criminal retribution (e.g. use of illicit drugs; child abuse).

The sensitivity of the behavior or attitude of interest may affect both the encoding of the information as well as the retrieval and reporting of the material; little of the survey methodological research has addressed the point as which the distortion or measurement error occurs with respect to the reporting of sensitive material. The encoding of emotionally charged behaviors is hypothesized to include an encoding of the emotion associated with the event. The presence of the emotion may affect further retrieval of that information. Cognitive dissonance may lead the respondent to “undo” the details of the event, distorting the event in subsequent rehearsals, thereby encoding the distorted information with the behavior (Loftus, 1975). Even if the respondent is able to retrieve accurate information concerning the behavior of interest, he or she may choose to edit this information at the response formation stage as a means to reduce the costs, ranging from embarrassment to potential negative consequences beyond the interview situation, associated with revealing the information.

5.3 Essential Survey Conditions

The measurement process and the quality of survey data can also be affected by design features such as the mode of data collection (e.g. face-to-face, telephone, self-administered), the method of data collection (e.g. paper and pencil, computer assisted interviewing), the nature of the respondent (self vs. proxy response), characteristics of the interviewer (e.g., gender, race, voice
quality), cross section vs. longitudinal design, the frequency and time interval between interviews for longitudinal data collection, as well as the data collection organization and survey sponsor. Groves (1989) provides a thorough review of empirical literature related to these various sources of error. While there is evidence that at times, each of these factors may affect the quality of the data, the empirical literature is inconsistent as to the direction and magnitude of the error attributable to each of these design features.

5.4 Applicability of Findings to the Measurement of Economic Phenomena

One of the problems in drawing inferences from other substantive fields to that of economic phenomena is the difference in the nature of the measures of interest. As noted earlier, much of the assessment of the quality of household-based survey reports concerns the reporting of discrete behaviors; many of the economic measures that are the subject of survey inquiries are not necessarily discrete behaviors or even phenomena that can be linked to a discrete memory. Some of the phenomena of interest could be considered trait phenomena. Consider the reporting of occupation. We speculate that the cognitive process by which one formulates a response to a query concerning current occupation is different from the process related to reporting number of doctor visits during the past year.

For other economic phenomena, it is likely that individual differences in the approach to formulating a response impact the magnitude and direction of error associated with the measurement process. Consider the reporting of current earnings related to employment. For some respondents, the request to report current earnings requires little cognitive effort—it may almost be an automatic response. For these individuals, wages may be considered a characteristic of their self identity, a trait related to how they define themselves. For other individuals, the request for information concerning current wages may require the retrieval of information from a discrete episode (the last paycheck), a recent rehearsal of the information (the reporting of wages in an application for a credit card), or the construction of an estimate at the time of the query based on the retrieval of information relevant to the request.

Given both the theoretical and empirical research conducted within multiple branches of psychology and survey methodology, what would we anticipate are the patterns of measurement error for various economic measures? The response to that question is a function of how the respondent’s task is formulated and the very nature of the phenomena of interest. For example, asking a respondent to provide an estimate of the number of weeks of unemployment during the past year is quite different from the task of asking the respondent to report the starting and stopping dates of each unemployment spell for the past year. For individuals who are in a steady-state (constant employment or unemployment), neither task could be considered a difficult cognitive process. For these individuals, unemployment is not a discrete event but rather may become encoded in memory as a trait which defines the respondent. However, for the individual with sporadic spells of unemployment throughout the year, the response formulation process would most likely differ for the two questions. While the response formulation process for the former task permits an estimation strategy on the part of the respondent, the latter requires the retrieval of discrete periods of unemployment. For the reporting of these discrete events, we would hypothesize that patterns of response error evident in the reporting of episodic behavior across other substantive fields would be
observed. Similar patterns of differences may be observed as a function of requesting the respondent to report current earnings as compared to directing them to think about their last paycheck and report the gross earnings. With respect to social desirability, we would anticipate patterns similar to those evident in other types of behavior, overreporting of socially desirable behaviors and underreporting of socially undesirable behaviors.

6.0 Evidence on Measurement Error in Survey Reports of Labor-related Phenomena

6.1 Earnings

Empirical evaluations of household-reported earnings information include the assessment of annual earnings, usual earnings (with respect to a specific pay period), most recent earnings, and hourly wage rates. Validation data are generally based on employers’ or administrative records. Gradually, the focus of such studies has shifted. Early studies tended to focus on whether the mean error was near zero, and so whether the survey reports were unbiased. More recent studies focus on the variance of the error relative to true variation and, more generally, on the bias caused by errors when survey measures of individual earnings are used in linear models. As a result, it is hard to report results from the various studies we review in a consistent fashion. Ideally, we would like to report information on the distribution of errors (e.g., the mean and variance of errors) together with some measure of the potential biases introduced into simple models by the error. Our preferred measure of this potential bias is the slope coefficient from the regression of the record values on the survey values of the same variable. As we have seen, under the assumption that the record values are valid, one minus this coefficient gives a measure of the proportional downward bias introduced by the measurement error for simple bivariate linear regression models that use the variable in question as the explanatory variable. The range of summary measures in Table 1 reflects the considerable variation in what can be computed from studies from different disciplines that are motivated by different questions.

Overall, the findings suggest that annual earnings are reported with less error than hourly wage rates or weekly earnings. Mean estimates of annual earnings appear to be subject to relatively small levels of response error, whereas absolute differences indicate significant over- and under-reporting at the individual level. We also find consistent evidence that errors are mean-reverting, but less consistent evidence that errors are correlated with standard human capital and demographic variables.

\[44\] The measure will be valid if the employer’s or administrative records are valid and error free or if the errors in such records are completely random.
6.1.1 Annual Earnings

Nine of the studies reported in Table 1, representing six different data collection efforts, examine the quality of reports of annual earnings. For each of these studies, comparisons are made between the survey reports and either administrative data (IRS tax forms, Social Security Administration records) or employers’ records.

Miller and Paley (1958) compared 1950 Census reports and IRS data for a sample of Census respondents. Limiting attention to families for which each member over age 14 could be matched to an income tax report (including spouses on joint filings), they found that median earnings were $3,412 in the Census reports and $3,570 in the IRS data. Moreover, the two distributions appear quite similar (see Table 10 in the original paper). While Miller and Paley do not ask whether the errors in the Census reports are mean reverting, the similarity of the two distributions suggests they must be.

By focusing on IRS records for validation, Miller and Paley excluded those with earnings low enough that they do not file an income-tax report. In contrast, Borus (1970) focused on survey responses of residents in low-income Census tracts in Fort Wayne, Indiana. He experimented with two methods for collecting annual earnings from respondents, a set of two relatively broad questions concerning earnings and a detailed set of questions concerning work histories. The responses from both sets of questions were compared to data obtained from the Indiana Employment Security Division for employment earnings covered by the Indiana Unemployment Insurance Act (e.g., excludes agricultural employees, self-employed, and those working for relatives). The mean annual earnings among the respondents was $2,500; although the mean error for the two question types was relatively small, $47 and $39 for the work history and broad questions, respectively, the standard deviation of the mean error was large ($623 and $767). Over 10 percent of the respondents misreported annual earnings by $1,000. While these individual-level errors seem large relative to the mean values, they are similar in magnitude to more recent estimates based on nationally representative samples (e.g., Bound and Krueger, 1991).

In contrast to one of Borus’s conclusions, Smith (1997) finds that, among low-income individuals eligible to participate in federal training, earnings data based on adding up earnings on individual jobs leads to significantly higher values than data based on direct questions about annual earnings.

---

45 The Panel Study of Income Dynamics (PSID) Validation study is represented three times; see Duncan and Hill (1985), Rogers, Brown, and Duncan (1993), and Bound, Brown, Duncan and Rogers (1994); the CPS-SSA matched study is reported in Bound and Krueger (1991) and Bollinger (1998).

46 Of 7,091 families, only 3,903 were completely matched. One important reason for non-matches is income low enough that no Federal tax would be owed; except for this difference, Miller and Paley (1958) find the matched sample representative of the larger Census sample.

47 Note that for families with more than one earner, we are really comparing family rather than individual earnings.
earnings. In Smith’s data, this difference is due to higher values for hours worked and for irregular earnings (overtime, tips, and commissions). Comparisons with administrative data for the same individuals lead Smith to conclude that the estimates based on adding up earnings across jobs leads to over-reporting, rather than more complete reporting.

Carstensen and Woltman (1979) compared reports of annual earnings obtained in a special supplement to the January, 1977 Current Population Survey (CPS) with employers’ reports. Respondents in rotation group 7 (1/8 of the entire CPS sample)\(^{48}\) were asked to report earnings as well as report his or her employer’s complete name and address. While one of the major strengths of the design is the nationally-representative sample of household respondents, the effective response rate of 61 percent raises questions as representativeness of the sample of matched employer-employee information.\(^{49}\) The use of a mail questionnaire to obtain information from the employer suggests that comparisons between the employer and employee information must consider measurement (or reporting) error in the validation data as a potential source of the discrepancy. The study includes comparisons of annual earnings, as well as hourly, weekly, and monthly rates of pay and usual hours worked. With respect to annual earnings, the absolute difference in the two earnings sources was $800 (s.e.=$403), or about 5 percent of the mean annual earnings.\(^{50}\)

The Panel Study of Income Dynamics (PSID) Validation Study consisted of two waves of interviews with respondents sampled from a single large manufacturing firm and the corresponding record information for those respondents.\(^{51}\) Cooperation by the firm essentially eliminated problems of matching validation data to each respondent and allowed for the resolution of anomalies in the

\(^{48}\)Note that each rotation group of the CPS sample is a nationally-representative sample.

\(^{49}\)Of the 6,791 eligible persons in the CPS, 5,591 (82 percent) provided complete employer address data. Among the employers for whom address information was provided by the CPS respondent, 4,166 (75 percent) responded to the mail survey which included the same earnings and hours questions asked of the CPS household respondent, resulting in an effective response rate of 61 percent.

\(^{50}\)Respondents in the Carstensen and Woltman study could report earnings in the time unit of their choice, that is, annual, weekly, monthly, or hourly. The comparison of annual earnings was limited to those respondents for whom both the respondent and the employer reported the earnings as annual earnings.

\(^{51}\)The PSID-VS was conducted by telephone with workers at their homes, rather than administered at the workplace. Similar to other household-based studies, the PSID-VS suffered from nonresponse. The initial wave of interviewing was conducted in the summer of 1983 with 418 of the 534 sampled employees (78.3 percent). A second wave of interviewing was conducted in the summer of 1987. The sample consisted of respondents to the initial wave and a fresh sample of hourly workers; the response rate among the initial wave respondents was 82.4 percent and 74.7 percent among the new sample of hourly workers, resulting in an overall sample size of 492 completed interviews.
validations. The questionnaire used at both waves requested that the respondent provide information for the previous two calendar years. At the time of the initial interview (1983), the firm’s hourly workforce was fully unionized and virtually all employees, both hourly and salaried, worked full time. The workforce was considerably older and had more job tenure than was true of national sample of workers, in part due to layoffs and few new hires in the two years prior to the initial interview. These deviations were offset by a sampling procedure that disproportionately sampled younger and salaried workers. Comparisons between the two validation samples and data from the Panel Study of Income Dynamics for the respective years indicates that, with respect to annual and hourly earnings, the validation sample respondents have considerably higher means and lower variance than a national sample.

Using data from the first validation study, Duncan and Hill (1985) compared reports of annual earnings for calendar year 1981 and 1982 with information obtained from the employer’s records. For neither year is the mean of the simple difference between the two data sources statistically significant, although the absolute differences for each year indicate significant under- and over reporting. The average absolute difference between the interview and record reports of earnings for 1982 was $2,123, approximately 7 percent of mean earnings. The report of earnings for 1981 was of lower quality than for 1982; the absolute difference of the two reports of earnings for 1981 was $2,567, or approximately 8.5 percent of mean earnings. The error-to-true variance ratio showed a larger difference between the two years: for calendar year 1982 annual earnings it was quite small (.154) but significantly larger for 1981 (.301).

While the margin of difference depends on the measure employed, by all indications previous-year's earnings are reported more accurately than those of two years prior to the interview. While this is consistent with greater error for longer recall periods, it may also reflect the fact that 1981 was a year of economic disruption both for the economy and for this firm.

Comparison of measures of change in annual earnings based on the household report and the employer records indicate no difference in means. Error to true variance ratios are higher for changes than for levels (.50 vs. .15-.30), even though mean absolute errors are similar for changes and levels. Errors in reported changes would be higher but for the positive correlation between the errors in the two years, .43. Duncan and Hill (1985) emphasize that these changes are obtained from differencing reports for two calendar years in the same interview, not differencing reports of last year's earnings from two interviews in a longitudinal survey.

Although the findings noted above are often based on small samples drawn from either a single geographic area (Borus, 1970) or a single firm (Duncan and Hill, 1985), the results parallel the findings from nationally representative samples. Bound and Krueger (1991) created a longitudinal linked file based on the 1977 and 1978 March CPS questionnaires and earnings histories from Social Security Administration files. The study is restricted to those respondents classified as heads of

---

52 As part of a joint project of the Census Bureau and the Social Security Administration, survey responses for persons in the March, 1978 CPS Annual Demographic File were linked to their respective earnings information in SSA administrative records to create the CPS-Social
households for whom information for March of 1978 was successfully matched to data reported in March of 1977 and the Social Security records. Of the 27,485 persons classified as heads of households in matchable rotation groups (50 percent of the CPS rotation groups), the three-way link was made for 9,137 persons. Other limitations (e.g. private, covered employment and positive (non-imputed) CPS and SSA earnings in both years) further reduced the effective sample to approximately 3,500 persons. Bound and Krueger note that the matching process tends to eliminate those who misreport their Social Security number or other matching data, and so those who tend to give inaccurate responses to other questions (e.g., earnings) may be under-represented. Another caveat is that the Social Security earnings data refer to earnings taxable under the payroll tax, and nearly half of the males in their sample reach this limit. Consequently, many of the estimates reported below are based on models that correct for this truncation, based on the assumption that ln earnings are normally distributed.

Bound and Krueger (1991) examined error in annual ln earnings reports separately for men and women. Although the error was distributed about a near-zero mean for both men and women, the magnitude of the error was substantial. For men, the error variance exceeded .10 and represented 27.6 percent of the total variance in CPS earnings; for women the error variance was approximately .05 and represented less than 9 percent of the total variance in CPS earnings for women. One striking feature of the errors is that while they appear to be unimodal and symmetric, the tails are substantially thicker than one would expect with a normal distribution. Indeed, for those for whom errors were directly observable (those below the Social Security earnings limit), the standard deviation of the errors was three times the interquartile range. In addition, the distributions show a large spike near 0. For those below the earnings limit, 12 percent of men and 14 percent of women report earnings that exactly match their Social Security records, while more than 40 percent of each gender report earnings within 2.5 percent.

Despite these errors, the correlation between interview and record ln earnings is high in Bound and Krueger’s data (.88 for men and .96 for women). Errors are negatively related to the record value for men (-.42), and essentially uncorrelated for women (-.03). Because errors for men are mean-reverting and errors for women are small, they find that measurement error should not appreciably bias the coefficient of ln earnings in linear models. The regression of record on interview values gives coefficients very close to 1 (.97 for men and .96 for women).

Because their data include two CPS waves, they can compare interview and record reports of changes in earnings as well. Consistent with the conventional wisdom, differencing increases the error variance (from .1 to .12 for men, and from .05 to .09 for women), and reduces the true variance by about half. Positive correlation in the errors (.4 for men, .1 for women) limits the increase in error variance due to differencing. Consequently, although the ratio of error to total variance is substantial (.65 for men, .2 for women) the regression of record changes on interview changes (.77 and .85)

Security Records Exact Match File (CPS-SER). To create a longitudinal data set, the CPS-SER was matched to the 1977 March CPS Annual Demographic File, based on the respondent’s unique CPS identification number, age, education, sex, and race.
suggest that the bias due to measurement error when the change in \( \ln \) income is an explanatory variable is not overwhelming.

Bollinger (1998) extended the work of Bound and Krueger (1991), examining the measurement error associated with each of the cross-sectional samples encompassing Bound’s and Krueger’s panel sample, expanding the sample to include women who were not heads of households, and incorporating nonparametric estimation procedures. To a large extent, Bollinger’s findings confirm those of Bound and Krueger. In addition, he finds higher measurement error in the cross-section samples as compared to the panel used by Bound and Krueger, suggesting that constructing a panel from CPS lead to the selection of respondents who appear to be better reporters. Bollinger also finds that the negative correlation between measurement error in reports of annual earnings and record earnings appears to be driven by a small proportion of men with low income who grossly overreport their earnings—or whose earnings are largely unrecorded by Social Security. Of additional interest in the work by Bollinger is the finding that although mean response error is negatively related to earnings, median response error is zero across earnings levels, suggesting median wage regression to be more robust to the effects of response error.

Coder’s (1992) analysis compares reports by respondents to the Survey of Income and Program Participation and Federal tax returns. The study is limited to SIPP respondents who were married couples as of March 1991, who met the following criteria: (1) valid Social Security numbers were reported for both the husband and the wife; (2) the couple could be matched to a married-joint tax return; and (3) nonzero wage and salary income amount was reported either during the SIPP interview or on the tax return. Of the approximately 9,200 husband-wife couples in the SIPP, 62 percent (or approximately 5,700 couples) met the criteria. Coder finds little difference between mean estimates of annual earnings and the respective validation source. He reports a simple correlation between earnings reported in SIPP and IRS data as .83; the mean annual earnings based on SIPP averaged approximately 4 percent less than the mean based on matched tax records. Coder’s data has an unusually large discrepancy between the variance of interview and record data, with the latter smaller; this in turn implies a very strong negative correlation between the “error” (SIPP-IRS) and the “true” (IRS) value—so strong that the effects of earnings on other variables would be overstated due to (mean-reverting) errors in earnings. Alternatively, it is possible that errors in the IRS data contribute to these results (Rodgers and Herzog, 1987, p. 408).

Bound, Brown, Duncan, and Rodgers (1994) analyze data from both (1983 and 1987) waves of the PSID Validation Study. The correlation between interview reports and company-record data on \( \ln \) earnings is about .9 (.92 for 1982 earnings, .89 for 1986), but the negative correlation between error and record values is weaker for 1986 (-.08 vs -.30). Consequently, the regression of record on interview value is closer to 1.0 for 1982 than for 1986 (.96 vs .82).

The distribution of errors for the PSID validation study appear to be quite different than that found by Bound and Krueger (1991) using the matched CPS-Social Security Earnings data. Since virtually all the individuals in the PSID validation study are men, it seems natural to compare PSID
validation study results to those for men using the CPS-SSE matched data. The two error distributions have similar means and interquartile ranges, but the PSID validation study data shows neither the spike at 0 nor the thick tails shown by the CPS-SSE matched data. As a result of the thick tails in the CPS-SSE data the variance of errors is an order of magnitude larger in the CPS-SSE data than it is in the PSID validation study data! What accounts for the difference in the distribution of errors between the PSID validation study and CPS-SSE data is unclear (see Bound, Brown, Duncan and Rodgers, 1994, p. 357 for a further discussion of these issues).

Given that the change in ln earnings computed from the PSID Validation Study covers four years rather than one, the findings for this variable should be seen as complementing rather than replicating Bound and Krueger's. The general patterns are strikingly similar--increased error variance, with the increase somewhat limited by the correlation in the errors over time; negative correlation between the error and the true value of the change (−.32), and regression of true change on interview reports of .77. One interesting difference is that the correlation between the errors is lower in Bound, et al.’s (1994) data (.14) than in Bound and Krueger's (1991). To some extent, this might be expected if the factors that produce this error change gradually over time; on the other hand, it may also reflect the difference between economy wide and single firm samples.

Three of these studies–Duncan and Hill (1985), Bound and Krueger (1991), and Bound, Brown, Duncan and Rogers (1994)–explore the implications of measurement error for earnings models. Duncan’s and Hill’s model relates the natural logarithm of annual earnings to three measures of human capital investment: education, work experience prior to current employer, and tenure with current employer, using both the error ridden self-reported measure of annual earnings and the record-based measure as the left-hand-side variable. A comparison of the ordinary least squares parameter estimates based on the two dependent variables suggests that measurement error in the dependent variable has a sizeable impact on the parameter estimates. For example, estimates of the effects of tenure on earnings based on interview data were 25 percent lower than the effects based on record earnings data. Although the correlation between error in reports of earnings and error in reports of tenure was small (.05) and insignificant, the correlation between error in reports of earnings and actual tenure was quite strong (−.23) and highly significant, leading to attenuation in the estimated effects of tenure on earnings based on interview information.

Bound and Krueger (1991) also explore the ramifications of an error-ridden left-hand-side variable by regressing error in reports of earnings on a number of human capital and demographic variables, including education, age, race, marital status, region, and SSA. Similar to Duncan and Hill (1985), the model attempts to quantify the extent to which the correlation between measurement error in the dependent variable and right-hand-side variables biases the estimates of the parameters. However, in contrast to Duncan and Hill, Bound and Krueger conclude that mismeasurement of earnings leads to little bias when CPS-reported earnings are on the left-hand-side of the equation.

Bound, Brown, Duncan and Rodgers (1994) estimate separate earnings functions using both interview and record earnings for both waves of the Validation Study. They find some evidence that
errors in reporting ln earnings are negatively related to tenure in 1982, and positively related to education in 1986. Overall, though, they find no consistent pattern. Rodgers, Brown, and Duncan (1993) note, however, that if annual hours are included as an explanatory variable, its coefficient is severely biased by a number of factors (e.g. correlation between errors in reporting hours and earnings, in addition to problems with the reliability of hours per se, as discussed in section 6.1.2).

While there is not much evidence that errors in reported earnings are strongly related to standard explanatory variables in earnings functions, two cautions should be noted. First, the tendency for errors in reported earnings to be mean-reverting means that, if there are no other problems, coefficients of all explanatory variables are biased toward zero. This bias is about 20 percent of the true coefficient in both studies. Second, errors in other variables may be correlated with earnings, but there is very little evidence one way or the other on this score.

The CPS-SSA matched data and the PSID validation data can also be used to shed some light on the impact of measurement error on earnings dynamics. The short nature of the CPS-SSA matched data panel limits its usefulness for this purpose, but the PSID validation study includes a total of six years of data. Using these data Pischke (1995) found that a relatively simple model in which measurement error in earnings stems from the under reporting of transitory earnings fluctuations together with a white noise component did a good job of explaining basic patterns in the PSID validation study data.

Pischke’s model rationalizes a number of the stylized facts that have emerged from recent earnings validation studies. In particular his model accounts for the finding that despite mean reversion, measurement error in earnings does not seem to significantly bias the coefficients on explanatory variables in earnings regressions -- the explanatory variables in such regressions would be expected to explain permanent, but not transitory earnings.

In terms of the estimation of earnings dynamics, Pischke’s estimates imply relatively good news. The negative correlation of measurement error with transitory earnings attenuates the role of the white-noise component. Pischke estimates that surveyed earnings tend to exaggerate the actual fluctuation in earnings by between 20 and 45 percent depending on the year, but do a reasonably good job identifying the relative importance of the permanent component to earnings changes.

53With 11 free parameters, Pischke fits 28 free covariances quite well. He reports an overall chi-square statistic on the model of 23.8 (p-value: 0.124).

54It is certainly possible to doubt the general validity of Pischke’s conclusion. His estimates are based on a tightly parameterized model that was estimated on data from a single firm. However, Baker and Solon (1998) have recently estimated earnings dynamic patterns using administrative data that are remarkably similar to patterns other authors (Baker, 1997; Haider, forthcoming) have found using survey data. These estimates would seem to confirm Pischke’s...
There are a few things that are important to note about the Pischke study. First, his model implies reporting errors will tend to be more severe at some points in time as against others (i.e., reporting errors will tend to rise in magnitude as the transitory component of earnings rises). Second, as Pischke emphasizes, it is hard to know how to generalize his results to more representative samples. Even were the PSID validation study establishment representative of establishments in the country as a whole, earnings dynamics in the sample would miss the component that arises when individuals move across firms.

On balance, the validation evidence suggests little bias in estimating mean annual earnings, and this is quite consistent with the fact that survey-based estimates of earnings aggregated up to economy-wide estimates correspond quite closely to earnings as measured in the National Income and Product Accounts. Moreover, despite significant absolute differences between household reports and record reports of earnings as well as significant error to record variance ratios, the correlation between the various sources of data are quite high. Several of the studies indicate coefficients for the regression of household reports on record reports of annual earnings near 1.0, reflecting a negative correlation between error in the household reports and the record value for annual earning. Only one study addressed the deterioration of the quality of reports of annual earnings as a function of time (Duncan and Hill, 1985); similar to empirical investigations in other fields, their findings provide support for less accurate reporting for longer reference periods. The evidence with respect to the impact of error in household reports of earnings is mixed; Duncan and Hill (1985) report significant attenuation in a model examining the effects of human capital investment, whereas Bound and Krueger (1991) conclude that misreporting of earnings leads to little bias for models incorporating CPS-earnings on the left-hand-side of the equation.

What can account for the significant individual differences between household and record reported annual earnings? The reporting of annual earnings within the context of a survey is most likely aided by the number of times the respondent has rehearsed the retrieval and reporting process for this information. We contend that the memory for one’s annual earnings is reinforced throughout the calendar year, for example, in the preparation of federal and state taxes or the completion of applications for credit cards and loans. To the extent that these requests have motivated the respondent to determine and report an accurate figure, such information should be encoded in the respondent’s memory. Indeed, both CPS and PSID time their collection of annual earnings data to coincide with the time when households would have received earnings reports from employers and might have begun preparing their taxes. Subsequent survey requests should therefore be “routine” in finding that measurement error does not have dramatic effects on estimated earnings dynamics.

For example, CPS-based estimates of total wage and salary income were 97% of independent estimates based on NIPA in 1990 (U.S. Census Bureau, 1993). As noted earlier, such a comparison reflects several factors besides the mean level of error in the individual reports, such as the accuracy of CPS adjustments for non-response.
contrast to many of the types of questions posed to a survey respondent. Hence we would hypothesize that response error in such situations would result from retrieval of the wrong information (e.g., annual earnings for calendar year 1996 rather than 1997), social desirability issues (e.g. overreports related to presentation of self to the interviewer), or privacy concerns, which may lead to either misreporting or item nonresponse.

However, several cognitive factors may affect the quality of reports of annual earnings. Comprehension may impact the quality of the information; for example, respondents may misinterpret the request for earnings information as a request for net earnings as opposed to gross earnings. In addition, the wording of most earnings questions does not stress the need for the reporting of exact earnings; hence respondents may interpret the question as one in which they are to provide estimates as opposed to precise reports of earnings. Estimation on the part of the respondents, as noted by Sudman, Bradburn, and Schwarz (1996), often leads to reports that are noisy at the individual level but unbiased at the population level. Retrieval of earnings information for any one year may also be subject to interference with respect to stored information concerning earnings in previous years. If the source of the misreporting by respondents was due to social desirability bias, we would anticipate that the direction of the error would be toward overreporting of annual earnings, especially among those with low levels of earnings and possibly, underreporting among those at the highest levels of earnings. Although there is evidence of a negative correlation between response error and the true value overall, there is little evidence to support the existence of social desirability bias with respect to the reporting of annual earnings (e.g., Bollinger, 1998).

### 6.1.2 Monthly, Weekly, and Hourly Earnings

In contrast to the task of reporting annual earnings, the survey request to report most recent earnings or usual earnings, is more likely to be a relatively unique request and one which may involve the attempted retrieval of information that may not have been encoded by the respondent, the retrieval of information that has not been accessed by the respondent before, or the calculation of an estimate “on the spot.” Hence, we would anticipate that requests for earnings in any metric apart from a well-rehearsed metric would lead to significant differences between household reports and validation data. Moreover, the extent of rehearsal is likely to differ by type of worker; for example, those paid a monthly salary are more likely to have accessed information about monthly earnings than are those paid by the hour, while the reverse is likely for earnings per hour.

While annual earnings is the most frequently studied measure of labor market compensation in validation studies, Table 1 makes it clear that significant effort has also been devoted to validating other measures. Roughly speaking, we can divide these studies into two groups: those that study weekly or monthly pay, and those that study pay per hour.

Four of the earliest studies in Table 1 focus on the correlation between weekly or monthly earnings as reported by workers and their employer's reports. All four (Keating, Paterson, and Stone's (1950) study of jobs held in the past year by unemployed workers in St. Paul; Hardin and Hershey's
Keating, Peterson, and Stone show a cross-tabulation of interview vs. record reports which displays at least weak mean reversion for men. However, from their grouped data, in which 70 percent of the 115 cases are on the diagonal, it is hard to say anything more precise. Cartensen and Woltman (1979) compare worker and employer reports, using a supplement to the January 1977 CPS. Their survey instruments allowed both workers and employers to report earnings in whatever time unit they preferred (e.g. annual, monthly, weekly, hourly). As noted earlier, comparisons are limited to those reports for which the respondent and the employer reported earnings using the same metric. Curiously, when earnings are reported by both worker and employer on a weekly basis, workers underreport their earnings by 6 percent; but when both report on a monthly basis, workers over-report by 10 percent. When the various reports are converted to a common time unit (usual weekly earnings), they find workers report earning 11.7 percent less per week than their employers' reports. Unfortunately, they do not report correlations between worker and employer reports.

Studies of hourly wages or earnings per hour are less common, in part because it is difficult to obtain validation data for salaried workers. Typically, their pay is stated in weekly, monthly, or annual terms, and employers often do not have records of the weekly hours of their salaried workers (see section 6.4).

In their study of wages, Mellow and Sider (1983) utilized the January 1977 CPS data first analyzed by Carstensen and Woltman (1979). Hourly wages calculated from the CPS reported earnings and hours compared to employers' records indicate a small, but significant, rate of

---

56 Keating, Peterson, and Stone show a cross-tabulation of interview vs. record reports which displays at least weak mean reversion for men. However, from their grouped data, in which 70 percent of the 115 cases are on the diagonal, it is hard to say anything more precise.

57 In the CPS sample, validation data could be obtained only where the worker provided the name and address of the employer, and the employer provided the relevant data. Mellow and Sider note that validation data could be obtained for only about two thirds of the eligible sample. However, reported CPS earnings of those who refused to provide employer contact, or whose employers refused to provide validation data, were similar to earnings of those who did not refuse. The EOPP was actually two large studies: a survey of approximately 5,000 establishments and the other of approximately 30,000 households. Because of the geographic overlap between the two studies, it was possible to link a limited number (n=3,327) of worker and employer responses. The representativeness of the resulting sample is unclear, and was not discussed by Mellow and Sider.
underreporting (ln hourly wage as reported by the worker lower by .048). The variance of the difference between interview and record reports is .148, which is larger than Bound and Krueger's error variances for the logarithm of annual earnings in CPS data (.114 and .051 for men and women).

In a reanalysis of the same data used by Mellow and Sider, Angrist and Krueger (1999) report more details. In their basic sample they find that the variance in the difference between interview and record values of ln hourly earnings to be .24. In comparison they report the variance in the ln of survey earnings to be .36. While the ratio of these two numbers suggests a signal to total variance ratio of one third, Angrist and Krueger's tabulations suggest very substantial mean reversion. The regression of record on surveyed ln earnings suggests attenuation of about 25 percent.

Duncan and Hill's (1985) analysis of PSID Validation Study data investigates the accuracy of earnings per hour values calculated from workers' reports of annual earnings, weeks worked, and average hours per week. Because hours data were available only for hourly workers, their analysis excludes the firm's salaried workers. On average, calculated earnings per hour are relatively accurate (under-reported by about four percent). But the error to true variance ratio of 2.8 leads the authors to characterize the extent of measurement error as "enormous"--the unhappy result of annual earnings being less accurately reported for hourly than salaried workers and substantial error in reports of annual hours (see below).

Bound, Brown, Duncan, and Rodgers (1994) report similarly discouraging results for the logarithm of earnings per hour--error to true variance ratios of about 1.5 in both 1982 and 1986, and correlations between interview and record values of .51 and .64. Predictably, matters only get worse for the change in the logarithm of earnings per hour.

The correlations between interview and record values are strikingly lower than those for weekly or monthly earnings in company-based samples noted above. The earlier company-based studies focused on salaried workers, whereas the PSID Validation Study's hourly earnings information is available only for hourly workers. As it happens, these workers are unionized and the number of hours per week is relatively compressed. In a sense, the poor results for hourly pay occur not because the reporting errors are so large (the standard deviations of the errors are .11 and .16 in the two years) but because true variation is so limited (standard deviations of .09 and .13).

Rogers, Brown, and Duncan (1993), using data from the second wave of the PSID validation study, analyze the accuracy of the logarithm of reported earnings and calculated earnings per hour over three time intervals--annual, most recent pay period, and "usual". Their analysis is restricted to hourly workers, since record data on hours per week were unavailable for salaried workers. Two generalizations are evident from Table 1: the correlation between worker and record reports declines

---

58 Operationally, they define “usual” as the average over the preceding six two-week pay periods. They report, however, that their results are not very sensitive to the precise definition.
as one moves from annual to pay period to "usual"; and for any given time interval, earnings per hour are less accurately reported than earnings.

Since wage rates were calculated from reported hours and earnings the variance in the error associated with the wage rate can be decomposed into three parts: the variance of the error in reported earnings, the variance of the error in reported hours, and the covariance of those two reports. While the details vary with the time interval, in general all three of these components are important.\(^{59}\)

Two studies focus on the accuracy of reports of starting wage in a particular job. Branden and Pergamit (1994) evaluated the consistency of respondents’ reports of starting wages in the National Longitudinal Study by comparing responses reported at time \(t\) to those reported one year later. Only 42 percent of those studied reported the same starting wage for a particular job across the two years.\(^{60}\) Consistency varied as a function of the time unit used for reporting, with higher rates of consistency among those reporting their starting wage as an hourly or daily rate (47 percent and 52 percent consistent, respectively) as compared to a consistency rate of approximately 13 percent for those reporting a biweekly wage rate. In contrast, Barron, Berger, and Black (1997)\(^{61}\) find a high

---

\(^{59}\)Rodgers, Brown, and Duncan report these components normalized as shares of the relevant error variance. For wage rates derived from reports of annual earnings and annual hours, the contribution due to error in annual earnings and annual hours are about equal (.93 and .80). The errors are positively correlated (\(r=.43\)) and so the covariance is negative (-.74). For wage rates based on the most recent pay period, errors in reported earnings are about twice as important as errors in reported hours (1.36 and .62, respectively); the covariance is again negative (-.98). Based on usual pay, the contribution due to error in reports of earnings is 1.26, from error in reports of hours is .32, and the covariance is -.58.

\(^{60}\)The authors did not provide a definition of what was considered the “same starting wage.”

\(^{61}\)The study reported by Barron, Berger, and Black was based on a sample of establishments with 100 or more employees, screened to determine whether they were hiring at the time of the initial interview. The data collection encompassed three interviews with the firm and three with the newly hired employee of the firm. Of the 5,000 establishments originally sampled, no attempt was made to contact 1,603 establishments due to budgetary restrictions. Of the 1,554 establishments classified as eligible and for whom interviews were attempted, complete information was obtained from 258 (16.6 percent) employer-employee pairs. The low response rate, coupled with the lack of information for over 32 percent of the originally sampled establishments, raises serious concerns with the inferential limitations of the study. The authors report that the sample for which they could obtain information was similar to the original 5000 establishments in size and industry, but completions were more likely to come from rural areas and the Mountain and Pacific regions.
correlation between employers’ and employees’ reports of starting wages (.974). Differences in the length of the recall period (one year vs. at most four weeks) most likely contributes to the differences in the findings from the two studies. Unfortunately, given these relatively short recall periods, neither study gives much evidence on the question of recall accuracy for starting wages of those who have been employed for longer periods (e.g. typical of information collected as part of a retrospective event-history question sequence).

On the whole, the evidence suggests that reporting of weekly or monthly earnings are highly correlated with employer reports. Available evidence on earnings per hour is much less reassuring. Unfortunately, the cautions from the various PSID Validation Studies are--as their authors indicate--likely to be overly dramatic because the true variance of hourly earnings is considerably smaller in one firm than in a broader sample.

As was true for annual earnings, a few of the studies in Table 1 attempt to assess the importance of measurement errors in frequently-estimated linear models. Mellow and Sider (1983) examined the impact of measurement error in wage equations; they concluded that the structure of the wage determination process model was unaffected by the use of respondent- or employer-based information, although the overall fit of the model was somewhat higher with employer-reported wage information. Bound, Brown, Duncan, and Rodgers (1994) report estimates of simple "labor supply" equations (ln hours regressed on ln earnings per hour and demographic controls). Here, a number of potential biases are at work--due to the unreliability of hours reported as well as errors in hourly earnings--and their impact depends on the true supply elasticity. In the end, their results suggest such estimates may be badly biased, though the direction of the bias and the contribution of errors in measuring earnings per hour are less clear.62

The studies reported in Table 1 provide conflicting indications of the relative accuracy of survey reports of monthly or weekly earnings, with some relatively old studies showing quite high correlations with record values. The calculation of hourly earnings appears to be most prone to error; the correlations between interview and record values are significantly lower for hourly earnings than

62 French (1998) uses the PSID-VS data to correct estimates of the inter-temporal labor supply elasticity for measurement error. Within the context of his model, the covariance of the change in hours and the once lagged change in wages scaled by the variance in the transitory component of wages should give an estimate of the inter-temporal labor supply elasticity. The covariance terms involve covariances between current and twice lagged hours and wages. French allows for individuals to under-report the transitory component of wages and the transitory component of hours caused by the transitory component of wages to be under-reported, and for errors in wages and hours to be correlated, but otherwise that measurement error is classical. With these assumptions, French is able to use the PSID-VS to correct for measurement error. His results suggest that measurement error can not explain the failure of inter-temporal labor supply effects to explain short term movements in hours.
for weekly, monthly, or annual earnings. In most of the studies, however, hourly earnings are calculated from separate reports of earnings and hours rather than based on direct reports of hourly earnings by respondents. The error in hourly earnings is therefore a function not only of misreporting of earnings (annual, weekly, or monthly) but also a function of the reporting of hours worked, the later being subject to high levels of response error (see section 6.4). An empirical investigation that has not been reported to date is the comparison of the accuracy of direct reports of hourly earnings by household respondents with the hourly earnings reports calculated from reports of earnings and hours.

6.2 Transfer Program Income

Transfer program income can be categorized broadly as falling within one of two categories: relatively consistent recipiency status and income levels once eligibility has been established, and highly volatile recipiency status as well as income. As with most other episodic events, we expect that relatively stable behavioral experiences will be reported relatively accurately whereas complex behavioral experience (e.g., month to month changes in the receipt of AFDC transfer income) would be subject to high levels of response error. Respondents experiencing complex patterns of on/off recipiency status will most likely err on the side of failing to recall exceptions to the rule (e.g., the two months out of the year in which they were not covered by a particular program). Depending upon the usual status quo for these respondents (receipt or nonreceipt), both under- and overreporting may be evident. In addition, for some transfer program income subject to social desirability bias, we would hypothesize that respondents would err on the side of underreporting receipt. Finally, misunderstanding as to the exact type of transfer program income received by the respondent may lead to the misidentification of recipiency, leading to underreporting of one type of income receipt and a corresponding overreport of another type of income receipt.

For most surveys, the reporting of transfer program income is a two-stage process in which respondents first report recipiency (or not) of a particular form of income and then, among those who report recipiency, the amount of the income. One of the shortcomings of many studies which assess response error associated with transfer program income is the design of the study, in which the sample for the study is drawn from those known to be participants in the program. Responses elicited from respondents are then verified with administrative data. As noted earlier, retrospective or reverse record check studies limit the assessment of response error, with respect to recipiency, to determining the rate of underreporting; prospective or forward record check studies which only verify positive recipiency responses are similarly flawed since by design they limit the assessment of response error only to overreports. In contrast, a “full” design permits the verification of both positive and negative recipiency responses and includes in the sample a full array of respondents. Validation studies which sample from the general population and link all respondents, regardless of response, to the administrative record of interest, represent full study designs. These would include the studies by Bancroft (1940), Oberheu and Ono (1975), Halsey (1978), Hoaglin (1978), and the more recent studies by Marquis and Moore (1990), Grondin and Michaud (1994), Dibbs, Hale, Loverock, and Michaud (1995), Moore, Marquis, and Bogen (1996), and Yen and Nelson (1996). The findings from the other studies cited in Table 2, many of which indicate a preponderance for underreporting by

58
respondents with respect to receipt of a particular type of income, are to some extent, an artifact of the study design. Rather than interpret the findings from these studies as indicative of a consistent underreporting bias on the part of the respondents, a more conservative conclusion may be to view the findings as illustrative of the types and magnitude of errors recipients can make with respect to program receipt.

There are several different ways of summarizing the frequency of reporting errors, which can give very different impressions of the accuracy of the data. One is the fraction of cases for which interview and record data disagree. Another is the difference between the fraction reporting receipt in the interview data and the corresponding proportion according to the records, which is the extent of net under- or over-reporting. A third is the pair of conditional probabilities, \( \pi_{01} = \text{Prob}(\text{interview}=\text{no}|\text{record}=\text{yes}) \) and \( \pi_{10} = \text{Prob}(\text{interview}=\text{yes}|\text{record}=\text{no}) \) that determines the extent of bias when recipiency is used as a variable in a regression (section 2.5).

These three measures are related: the probability of disagreement = \( \pi_{01} + (1-\pi) \pi_{10} \), and net under-reporting = \( \pi \pi_{01} (1-\pi) \pi_{10} \). The probability of disagreement tends to be lower for programs with low true participation rates as long as \( \pi_{01} > \pi_{10} \); relatively high values of both \( \pi_{01} \) and \( \pi_{10} \) can lead to near-zero net under-reporting but imply significant biases in a regression context.

Focusing our attention first on reporting of receipt of a particular transfer program, among the full design studies, there does appear to be a tendency for respondents to underreport receipt, although there are also examples of overreporting recipiency status. For example, Oberheau and Ono (1975) report a low correspondence between administrative records and household report for receipt of AFDC (monthly and annual) and Food Stamps \( (\pi_{10}=.2, \pi_{01}=.3) \), but relatively low net rates of under- and over-reporting. In the study reported by Marquis and Moore (1990), respondents were asked to report recipiency status for eight months (in two successive waves of SIPP interviews). Although Marquis and Moore report a low error rate of approximately 1 percent to 2 percent (not shown in table), the error rate among true recipients is significant, in the direction of underreporting. For example, among those receiving AFDC, respondents failed to report receipt in 49 percent of the person-months. Underreporting rates were lowest among OASDI beneficiaries, for which approximately 5 percent of the person-months of recipiency were not reported by the household respondents. The mean rates of participation based on the two sources suggest little difference; absolute differences between the two sources differed by less than one percentage point for all income types. However, the rareness of some of these programs means that small absolute biases mask high rates of relative bias among true participants, ranging from +1 percent for OASDI recipiency to almost 40 percent for AFDC recipiency. In a follow-up study, Moore, Marquis, and Bogen (1996)

---

\(^{63}\) Oberheau and Ono’s sample is restricted to low-income households. This is likely to lead to a larger value of \( \pi_{01} \) than would be obtained in samples with the full range of household incomes. For example, \( \pi_{01} \) would be increased by mis-reporting other transfers, and these would be more common in low-income households.
compared underreporting rates of known recipients to overreporting rates for known non-recipients and found underreporting rates to be much higher than the rate of false positives by non-recipients. They also note that underreporting on the part of known recipients tends to be due to failure to ever report receipt of a particular type of income rather than failure to report specific months of receipt. 

In contrast, Yen and Nelson (1996) found a slight tendency among AFDC recipients to overreport receipt in any given month, such that estimates based on survey reports exceeded estimates based on records by approximately 1 percentage point. Oberheu and Ono (1975) also note a net overreporting for AFDC (annual) and Food Stamp recipiency (annual), 8 percent and 6 percent, respectively.

The studies vary in their conclusions with respect to the direction and magnitude of response error concerning the amount of the transfer, among those who report receiving it. Several studies report a significant underreporting of assistance amount (e.g. Livingston, 1969; Oberheu and Ono, 1975; Halsey, 1978) or significant differences between the survey and record reports (Grondin and Michaud, 1994). Other studies report little to no difference in the amount based on the survey and record reports. Hoaglin (1978) finds no difference in median estimates for welfare amounts and only small negative differences in the median estimates for monthly Social Security Income. Goodreau, Oberheu, and Vaughan (1984) found that 65 percent of the respondents accurately report the amount of AFDC support; the survey report accounted for 96 percent of the actual amount of support. Although Halsey (1978) reported a net bias in the reporting of unemployment insurance amount of -50 percent, Dibbs, Hale, Loverock, and Michaud (1995) conclude that the average household report of unemployment benefits differed from the average true value by approximately 5 percent ($300 on a base of $5600).

In general, studies that assess the accuracy of transfer data from household surveys do not provide analyses of how such errors affect the parameters of behavioral models. An exception is Bollinger and David (1997), who estimate a parsimonious model of response error from validation data and then combine this information into a model of Food Stamp participation using a broader sample. They find that estimated effects of wealth and predicted earnings are increased by such corrections, though they note that these results depend on the model of response error based on a relatively small validation sample (N=2685, but with only 181 participants). They also note that low income households are much more likely to mis-report Food Stamp receipt because they confuse Food Stamps with other transfers they receive; high-income respondents do not have other transfer programs to confuse Food Stamps with. Thus, while many examples of differential measurement error in survey reports of transfers are due to deliberate under-reporting, Bollinger and David’s example shows that differential errors may also occur inadvertantly.

Studies of receipt of transfer payments are often interested not only in which groups are receiving transfers and how much they receive at one point in time, but also in the duration of receipt, and so in the transitions into and out of recipiency. Marquis and Moore (1990) matched data from SIPP interviews to administrative records for major transfer programs. They find that the number of
transitions (those starting to receive benefits, and those whose benefits end) are overstated by interview respondents for some benefit programs and understated for others. A more consistent pattern is that such transitions are over-stated when one compares the last month of the reference period of one interview with the first month of the next—the so-called “seam”—and understated when one compares reports for two months collected in the same interview.64

Comparing the findings for transfers with those for earnings suggests several broad conclusions. First, there is evidence of under-reporting of transfers, in contrast to the approximately zero-mean errors we found for earnings, and such under-reporting seems more important for AFDC and other public assistance than for Social Security. This is quite consistent with comparisons of aggregate estimates based on survey reports to independent estimates of aggregate amounts received.65 Second, both non-reporting and under-reporting by those who report receiving positive transfer benefits contribute to this under-reporting, though it is hard to draw firm conclusions about the relative importance of these two sources of error. Third, accuracy of reports for individual transfers is reduced by mis-classification; i.e., respondents who report receiving a transfer, and may even report the amount correctly, but incorrectly identify the program that provided the benefit. Fourth, the focus on extent of under-reporting in most studies leaves us with very little evidence on the likely effects of errors in reporting transfers when benefits from individual programs are used as either dependent or explanatory variables in behavioral models.66

---

64 The finding of more transitions at the “seam” than at other points in a retrospective history pieced together from a series of interviews has been documented repeatedly (Moore and Kasprzyk, 1984; Burkhead and Coder, 1985; Hill 1987).

65 In 1990, CPS totals amounted to 97 percent of independently-estimated levels of Social Security and railroad retirement benefits, and 89 percent of Supplemental Security Income payments. In comparison, CPS captured only 72 percent of AFDC and 86 percent of other public assistance (U.S. Census Bureau, 1993, Table C-1).

66 Since benefits received depend in part on choices made by the recipient, analysts often use some sort of instrumental variable procedure to account for this endogeneity; for example, the level of AFDC benefits available in a state might be used as an instrument for the reported benefit level. While one might hope that instrumenting would undo the bias from measurement error as well, we have stressed that this hope depends on the reporting error being “classical”. Given that benefits are bounded (at zero) and zero benefits are in fact common, we suspect errors are likely to be mean-reverting. Particularly for programs such as AFDC where reporting seems least accurate, the effect of reporting error on the consistency of IV estimates deserves explicit discussion.
6.3 Assets

The literature on accuracy of reports of individual assets (and so, implicitly, of net worth) is similar in important ways to the literature on transfer income. Comparisons of aggregate values based on survey reports to independent estimates of these aggregates suggests that under-reporting is likely to be a problem (Curtin, Juster, and Morgan, 1989).67 The literature has therefore focused on the extent of such under-reporting, rather than on the variance of the error relative to the variance of the true (record) value, or the correlation between errors and true values.

A limited number of studies have focused on the assessment of measurement error related to the reporting of assets and only one of these, the study by Grondin and Michaud (1994), focuses specifically on interest and dividend income generated from asset ownership. Several studies conducted during the 1960s examine the extent to which respondents accurately reported savings account and stock ownership, comparing survey reports with financial institution reports for a sample of respondents known to own the particular asset of interest. As noted above, reverse record check studies by design limit the detection of response error to underreports. Hence, one should be cautious in drawing conclusions concerning the direction of response error based on these studies. As noted in Table 3, between 5 percent and almost 50 percent of respondents fail to report existence of a savings account; 30 percent of those who own stock failed to report ownership. The high rate of underreporting is also evident in the full design validation study reported by Grondin and Michaud (1994).

Among those who report ownership of a savings account or stocks, the findings are mixed with respect to the accuracy of account amounts. Maynes (1965) and Ferber, Forsythe, Guthrie, and Maynes (1969a) report a small amount of net bias for reports of savings account amounts (−5 percent and 0.1 percent, respectively), while Ferber, Forsythe, Guthrie and Maynes (1969b) report that 80

---

67 Curtin, Juster, and Morgan report that the 1983 Survey of Consumer Finances produces aggregate net worth estimates that are close to those based on external (flow-of-funds) sources. This “adding up,” however, reflects a balance between substantial discrepancies on particular wealth components (e.g., SCF shows “too little” liquid assets but “too much” housing), and a close look at these discrepancies suggests that the external totals are often not very accurate benchmarks for the survey data (e.g., because of difficulties in the flow-of-funds accounts in separating household and business asset holdings). However, alternative wealth surveys show substantially lower levels of net worth than does SCF (PSID and SIPP being roughly 80 and 60 percent of SCF, respectively). Juster, Smith, and Stafford (1999) report that wealth surveys conducted in the 1960s typically found about two thirds of the net wealth found in the external sources. CPS reports of interest and dividend income were 51 and 33 percent of NIPA totals (U.S. Census Bureau, 1993). Thus, comparisons with external totals suggest that under-reporting is likely to be the norm, although failure to sample the wealthiest households also contributes to these discrepancies.
percent of respondents are accurate in their reports of stock holdings. In contrast, Feber et al. (1969a) indicate that there is a large degree of response error, with only 40 percent of respondents reporting the account amount within 10 percent of the true value. Similarly, Lansing, Ginsburg, and Braaten (1961) indicate an absolute discrepancy of almost 50 percent between financial records and household survey respondents’ reports of saving account amounts, a discrepancy similar to that reported by Grondin and Michaud (1994).

A few studies attempt to validate survey responses to questions about the value of owner-occupied housing, a very important component of wealth for most households. Kish and Lansing (1954) find that owners’ estimates are close to appraisers’ on average (mean discrepancy = 4 percent) but the two estimates differ by 30 percent or more in a quarter of the cases. Scrutinizing cases with the largest discrepancies—which a typical survey, without validation data, could not do—they find that the largest discrepancies were due to coding errors (e.g., omitting a zero or misreading a lead digit in moving from the interview form to the data record). Rodgers and Herzog (1987) find that differences between household estimates of assessed value and property-tax records of assessed value are positively related to the record value. This contrasts with the negative correlation they find for other variables, and which is typically found in other studies.

Related perhaps to respondents’ difficulty in providing accurate responses to questions about asset holdings is the substantial level of item response—it is not uncommon for 30 percent of those who report owning an asset to either refuse to provide or claim to not know the value of the asset (Juster and Smith, 1997). In response, surveys have increasingly used “unfolding brackets”: questions of the form “would it be more or less than X”, where a “yes” (“no”) to the first such followup leads to a second with a higher (lower) value of X. Thus, respondents unwilling or unable to provide a dollar amount are induced to specify a range in which they believe the value of their asset holding lies. Since those who are initially unwilling or unable to give a dollar value for the asset tend to be those with higher true values, brackets help to reduce the under-reporting typically found by comparing asset levels as reported in household surveys to external (aggregate) values. For example, Juster and Smith (1997, Table 8) report that bracket based imputations produce 6-12 percent higher estimates of mean net worth than imputations not based on bracket information.

However, experiments in which the bracket boundaries are varied randomly find that the distribution of amounts that comes out of the brackets depends on the bracket boundaries themselves. For example, in one study the fraction of cases with savings accounts less than $10,000 was 49 percent with the first bracket question set X equal to $1,000 but only 37 percent when the first bracket question set X=$20,000. This, in turn, has led to several attempts to obtain “corrected” estimates by jointly modeling the determinants of the asset value and the effect of the (randomized) bracket boundaries (e.g., Hurd et al., 1997; Hurd and Rodgers, 1998). Both studies find responses are pulled toward the boundary in the first bracket question. Setting bracket boundaries with an eye toward

---

68 Hurst, Luoh, and Stafford (1998) attribute to Donald Trump the observation that those who know how much their assets are worth can’t be worth very much.
maximizing the fraction of the variance in the asset that can be accounted for by the categorical responses will tend to place the first bracket boundary toward the middle of the distribution of the asset in question (Heeringa, Hill, and Howell, 1995). Consequently, it is likely that the error induced by “anchoring” effects is likely to be mean-reverting in most applications. Lacking validation data, however, it is hard to say much about the effects of using bracket-based imputed values in regressions that use wealth as either dependent or explanatory variable.

6.4 Hours Worked

Obtaining validation data for workers' reports of how many hours they work per week has proved more difficult than obtaining earnings data. In general, the administrative records--income tax, unemployment insurance, and Social Security payroll tax--used in many of the studies in Table 1 include no comparable data on hours worked. The largest Federal establishment survey of payroll and hours collects hours only for production workers in manufacturing and non-supervisory workers in other industries. A Bureau of Labor Statistics study that considered obtaining hours information for all workers noted "Hours data are less available than total payroll for most categories of workers" (U.S. Bureau of Labor Statistics, 1983, p. 22).

While the number of empirical investigations concerning the quality of household reports of hours worked is limited, one finding consistently emerges. Regardless of whether the measure of interest is hours worked last week, annual work hours, usual hours worked, or hours associated with the previous or usual pay period, comparisons between company records and respondents' reports indicate that interview responses overestimate the number of hours worked. The findings from seven studies in which household reports of hours worked are compared to employer’s records are reported in Table 4; all of these studies were also represented in Table 1. Findings from three studies in which the quality of the reports of hours worked is compared to time-use diary estimates are also reported in Table 4.

Carstensen and Woltman (1979) compared reports of "usual" hours worked per week. They found that compared to company reports, estimates of the mean usual hours worked were significantly overreported by household respondents; 38.4 hours vs. 37.1 hours, respectively, a difference on average of 1.33 hours, or 3.6 percent of the usual hours worked. Similarly, Mellow and Sider (1983) report that the mean difference between the natural ln of worker reported hours and the natural ln of employer reported hours was .039. They also report that the measurement error has a non-trivial variance (.064) but do not compare that variance to that of either the interview or the record hours variable.

In their reanalysis of this same data, Angrist and Krueger (1999) report a variance of the difference in ln hours of .038. This compares to the variance in ln survey hours of .195 or a signal to total variance ratio of roughly .8. Again, mean reversion will tend to reduce the implied attenuation to less than the .2 this number suggests.
Duncan and Hill (1985) find that worker reports of hours worked in the previous year (from the first wave of the PSID Validation Study) exceed company reports by 90 hours per year, nearly 6 percent of mean hours. The average absolute error was 157 hours. Recall of hours worked two years ago were less accurate, as expected, with a mean absolute error of 211 hours. More readily related to the discussion of biases in section 2 is their finding that the ratio of error to record variance is .37. Bound, Brown, Duncan, and Rodgers (1989) also find hours are overreported in the second wave of the Validation Study, though the mean error for ln (annual hours) is only .012, which is not statistically significant. However, the variance of the error is about .6 of the variance of record ln hours. Once again, there is evidence of significant mean reversion (correlation between error and true hours of -.37). Rodgers, Brown, and Duncan (1993) consider various time intervals—hours worked in the previous year, hours worked in the previous pay period, and "usual" hours worked. They find the correlation between interview reports and company records is .61 to .66 for all three measures; and, for all three measures, the correlation between error and company records is -.31 to -.37. It is worth recalling that the PSID Validation Study obtained data from one manufacturing firm with few part-time workers and therefore, limited variation in hours per week, but (at least at the first wave) less than full-year employment for many workers. Moreover, hours data were unavailable for salaried workers. Barron, Berger, and Black (1997) report a correlation between employers' records and respondents' reports of hours last week, .769; but this correlation falls to .61 for ln (hours).

One might wonder whether, in the case of hours, the company reported values should be treated as "true." For those who are paid by the hour, accurate recording of hours is essential for correctly paying the worker, and for those who "punch a clock" the company presumably has at least accurate records of the worker's coming and going. For other workers, the link between hours worked and pay is much less tight, and as noted above some employers may not even keep records of their hours.

Three of the studies represented in Table 4 take a different approach to assessing the worker's report of hours worked. In addition to CPS-like questions on hours worked per week, the time use studies obtained time diaries from respondents. These diaries involved detailed reporting of activities at each time in the previous day. While only a few days' time diaries were collected from each respondent, when aggregated across respondents, work hours reported in the time diaries should add up to those reported in CPS-like questions. All three studies in Table 4 that used time use data (Stafford and Duncan, 1980; Hamermesh, 1990; Robinson and Bostrom, 1994) report that CPS-style questions lead to higher estimates of work time than are obtained from the time diaries. The discrepancies are, if anything, larger than those between worker and employer reports, and the gap between CPS-like questions and the time diary based estimates is growing over time.

Evidence on the importance of measurement error in interview-based measures of change in hours is available in the studies based on the PSID-VS. Duncan and Hill (1985) find that constructing the change in annual hours by differencing reports for two previous years asked in a single interview (the first PSID-VS wave) leads to a relatively noisy measure, with an error to true variance ratio of .8. While sizeable, errors in changes calculated in this way are likely to be reduced by a positive
correlation between the errors in the two years (Rodgers, Brown, and Duncan, 1993) report that correlation as .36 in the second PSID-VS). Bound, Brown, Duncan, and Rodgers (1989) calculate the change in ln hours as one would in a longitudinal survey, as the difference between the logarithm of 1986 hours (reported in 1987) and 1982 hours (reported in 1983). Whether measured by the error to true variance ratio, the correlation between interview and record values, or the regression of record value on interview, the change in hours data are slightly more reliable than the levels data.

Examination of a model with earnings as the left-hand-side variable and hours worked as one of the predictor variables indicates that the high correlation between the errors in reports of earnings and hours (ranging from .36 for annual measures to .54 for last pay period) seriously biases parameter estimates. For example, regressions of reported and company record ln annual earnings on record or reported ln hours, age, education, and tenure with the company provide a useful illustration of the consequences of measurement error. Based on respondent reports of earnings and hours, the coefficient for ln hours is less than 60 percent of the coefficient based on company records while the coefficient for age is 50 percent larger in the model based on respondent reports. In addition, the fit of the model based on respondent reports is less than half that of the fit based on company records (R² of .352 vs. .780).

The small number of studies validating worker reports of work hours against employer reports provide little guidance on the relationship between errors in reporting hours and other variables. Mellow and Sider's (1983) regression explaining the difference between the two sources indicates that professional and managerial workers were more likely to overestimate their hours, as were respondents with higher levels of education and nonwhite respondents. In contrast, female respondents tended to underreport usual hours worked.

In contrast to the findings with respect to annual earnings, we see both a bias in the population estimates as well as bias in the individual reports of hours worked in the direction of overreporting. This finding persists across different approaches to measuring hours worked, regardless if the respondent is asked to report on hours worked last week (CPS) or account for the weeks worked last year, which are then converted to total hours worked during the year (PSID). The consistent direction of misreporting coupled with what appears to be a trend toward increasing discrepancy over time suggests that (1) respondents misinterpret the question (monthly CPS); (2) incorrectly account for weeks worked (March CPS supplement and PSID); or (3) overreport as a result of social desirability bias in the direction of wanting to appear to be working more than is true. The monthly CPS questions concerning hours worked ask the respondent to report the total number of hours worked, not hours spent at the employer’s site or hours of paid work. One potential source of error may be a difference in the underlying concept of interest, with users of the CPS data examining hours of paid employment and respondents indicating total number of hours, regardless of location or pay. The approach used in the March CPS and PSID to obtaining hours worked requires that the respondent report the number of weeks worked in the previous year. The March CPS question even includes the word “about” suggesting that the respondent can provide a rough estimate of the number of weeks worked. Here we would speculate that once again, the bias is in the direction of errors of omission related to exceptions.
The CPS is collected each month from a probability sample of approximately 50,000 households; interviews are conducted during the week of the month containing the 19th day of the month and respondents are questioned concerning labor force status for the previous week, Sunday through Saturday, which includes the 12th of the month. In this way, the data are considered the respondent’s current employment status, with a fixed reference period for all respondents, regardless of which day of the week they are interviewed. The design is a rotating panel design in which households selected for participation are interviewed for four consecutive months, followed by eight months of no interviews, and then interviewed for the same four months one year later. In any one month, 1/8 of the sample is being interviewed for the first time, 1/8 for the second time, etc.

6.5 Unemployment

Concern about the reliability of survey reports relating to unemployment focuses on a number of distinct but related issues. One question is how accurate are reports of current labor force status, in which individuals are classified as employed, unemployed, or not in the labor force. A question issue is how errors in reporting labor force status in one month affect estimates of various labor force transitions (e.g., leaving unemployment by finding a job or leaving the labor force), and estimates of the duration of spells of unemployment that are calculated from such transitions. Other studies have focused on the accuracy of retrospective reports, including the number of spells of unemployment, the duration of such spells (including on-going spells) and the total length of time unemployed in a particular period. Unlike the variables considered in previous sections, there are no employer or administrative records that allow one to verify whether non-working individuals are unemployed or not in the labor force.

6.5.1 Current Labor Force Status, and Transitions to and from Unemployment

The most widely used data on current employment status come from the Current Population Survey, which asks a series of questions each month and on the basis of the responses classifies individuals as employed, unemployed (roughly, looking for work), or not in the labor force (not working and not seeking work). Correctly classifying individuals involves taking proper—according to official definitions—account of complications such as wanting to work but believing none is available, search for a new job while on paid vacation from another job, school teachers on summer vacation, etc. Concerned with the accuracy of these responses, CPS regularly re-interviews a subsample of its respondents, re-asking the standard questions (about the reference week covered by the original interview) and attempting to reconcile any differences that the re-interview uncovers.

---

69 The CPS is collected each month from a probability sample of approximately 50,000 households; interviews are conducted during the week of the month containing the 19th day of the month and respondents are questioned concerning labor force status for the previous week, Sunday through Saturday, which includes the 12th of the month. In this way, the data are considered the respondent’s current employment status, with a fixed reference period for all respondents, regardless of which day of the week they are interviewed. The design is a rotating panel design in which households selected for participation are interviewed for four consecutive months, followed by eight months of no interviews, and then interviewed for the same four months one year later. In any one month, 1/8 of the sample is being interviewed for the first time, 1/8 for the second time, etc.
Several of the studies in Table 5 report estimates of the probability that an individual initially classified as unemployed (or employed, or not in the labor force) will be judged as unemployed following the re-interview process. A consistent finding of these studies (Poterba and Summers, 1984, 1986; Abowd and Zellner, 1985; Chua and Fuller, 1987) is that 11-16 percent of those classified as unemployed are likely to be mis-classified, with most of the re-classifications being to not in the labor force rather than to employed.\footnote{Poterba and Summers and Abowd and Zellner take the reconciled status from the re-interview as the “true” status. Chua and Fuller note that initial reports on the re-interview survey are more consistent with the initial CPS interview for the fraction of the sample where reconciliation is carried out than on the fraction where it is not. This suggests that, contrary to instructions, those conducting the re-interviews are aware of the initial CPS response before the respondent has answered the initial re-interview status. Poterba and Summers speculate that knowing the initial report leads re-interviewers to minimize discrepancies (and hence the work required to reconcile them). This would imply the reconciled responses are biased toward the original reports, and so taking them as true leads to underestimate the extent of error in the regular CPS.}

A distinct but related problem with the classification of labor market status is that those in households that are interviewed by CPS for the first time are more likely to be classified as unemployed than they are in later months. There is also a weaker tendency for fewer of those in their sixth and seventh months to be counted as unemployed.\footnote{Bailar (1975) reports that in 1968-69, the number counted as unemployed was 20 percent higher for those in their first month than the average regardless of month. This fell to 9 percent in 1970-72 (Bailar, 1975) and 8 percent in 1974-83 (Solon, 1986). Over these same time periods, the number counted as unemployed is 5-7 percent lower in month 7 than overall, and 3-4 percent lower in month 6.} This pattern of “rotation group bias” is also present in the re-interview data (Bailar, 1975), which serves as a reminder that the re-interview data should not be regarded as error-free.

If those who are mis-classified in one month are correctly classified (or misclassified in a different way) in the next month, the number of transitions from one state to another will be exaggerated. For example, some of those who appear to move from unemployed to not in the labor force will in fact have been out of the labor force in both months. The extent to which classification error in one month biases estimates of transitions between statuses depends on whether the errors are persistent or independent from one month to the next. Lacking direct evidence on this score, analysts assume that the errors in one month are unrelated to errors in the next. On this assumption, a significant fraction of the apparent transitions—in particular, .10-.18 of the roughly .5 probability of leaving unemployment from one month to the next—appear to be due to errors in classifying workers
in each of the months; transitions from unemployment to not in the labor force are exaggerated more than are transitions from unemployment to employment.

While there has been significant effort devoted to gauging the likely effects of errors in measuring labor force status on transition rates, there is much less evidence on how such errors might affect analyses of the effects of various factors on such transitions. Poterba and Summers (1995) explore the consequences of errors in reporting employment status for estimates of the effects of unemployment insurance and welfare receipt on the probability of leaving unemployment. Initially, they model the reporting errors as fixed probabilities, independent of the explanatory variables. Correcting for reporting errors based on re-interview evidence has little effect on the estimated effects of unemployment insurance, but substantially increases the effect of welfare receipt on labor force withdrawal. They note, however, that previous work suggests reporting errors in one month are higher for those who were unemployed in the previous month. They present alternative estimates intended to capture this intuition, albeit informally. If these alternative estimates of the probability of classification error are correct, effects of unemployment insurance and welfare receipt on transitions out of unemployment are significantly underestimated due to such error.

6.5.2 Retrospective Unemployment Reports

A substantial number of studies have examined directly the quality of unemployment reports. These studies, reported in Table 5, encompass a variety of unemployment measures including annual number of person years of unemployment, weekly unemployment rate, occurrence and duration of specific unemployment spells, and total annual unemployment hours. Only one study reported in the literature, the PSID validation study (Duncan and Hill, 1985; Mathiowetz, 1986; Mathiowetz and Duncan, 1988), compares respondents’ reports with validation data; the majority of the studies reported in Table 5 rely on comparisons of estimates based on alternative study designs or examine the consistency in reports of unemployment duration across rounds of data collection. In general, the findings suggest that retrospective reports of unemployment by household respondents underestimate unemployment, regardless of the unemployment measure of interest.

Several of the studies reported in Table 5 compare unemployment statistics based on reports to the monthly Current Population Survey (CPS) to those obtained from the Work Experience Survey (WES), a set of questions included in the March Supplement to the CPS. The studies by Morganstern and Bartlett (1974), Horvath (1982), and Levine (1993) compare the contemporaneous rate of unemployment as produced by the monthly CPS to the rate resulting from retrospective reporting of unemployment during the previous calendar year. The measures of interest vary from study to study; Morganstern and Bartlett focus on annual number of person years of unemployment, Horvath on average estimates of weekly unemployment, and Levine on the unemployment rate. Regardless of the measure of interest, the empirical findings from the three studies indicate that when compared to the contemporaneous measure, retrospective reports of labor force status result in an underestimate of the unemployment rate. The rate of underreporting, depending upon both the measure of interest, the population, and the year, ranged from as low as 3 percent to as high as 25 percent. The discrepancy
between the retrospective WES and the contemporaneous reports is generally taken as evidence of recall error. Note, however, that the monthly status reports are based on a complex algorithm that combines answers to a series of questions, while the WES allows the respondent greater freedom in self-classifying.

Across the three studies, the underreporting rate is significant and appears to be related to demographic characteristics of the individual. For example, Morgenstern and Bartlett (1974) report discrepancy rates of 3 to 24 percent, with the highest discrepancy rates among women (22 percent for black women; 24 percent for white women). Levine compared the contemporaneous and retrospective reports by age, race, and gender. He found the contemporaneous rates to be substantially higher relative to the retrospective reports for teenagers, regardless of race or sex, and for women. Across all of the years of the study, 1970-1988, the retrospective reports for white males, ages 20 to 59, were almost identical to the contemporaneous reports.

One of the strengths of these three studies is the ability to examine the underreporting rates across many years of data and the impact of economic cycle on the quality of the retrospective reports of unemployment. The findings suggest a relationship between economic cycle and the quality of retrospective reports; Morgenstern’s and Bartlett’s (1974) analyses indicate that in years of high unemployment, retrospective reports of unemployment from the WES overstate the amount of unemployment. Horvath (1982) found that in periods of increasing unemployment, discrepancies between estimates of the average weekly unemployment rate based on concurrent and retrospective reports were smaller than during other economic periods. Levine’s (1993) findings were similar to those of Horvath; he found that underreporting declined during recessionary periods and increased during expansionary periods.

In contrast to the findings comparing the estimates of unemployment from CPS and the WES, Duncan and Hill (1985) found that the overall estimate of mean number of hours unemployed one and two years prior to the survey based on employee reports and company records did not differ significantly. However, micro-level discrepancies, reported as the average absolute difference between the two sources, were large relative to the average amount of unemployment in each year.

In addition to studies which examine rates of unemployment, person-years of unemployment, or annual hours of unemployment, several empirical investigations have focused on spell-level information, examining reports of the specific spell and duration of the spell. Using the same data as presented in Duncan and Hill (1985), Mathiowetz and Duncan (1988) found that at the spell level, respondents failed to report over 60 percent of the individual spells. Levine (1993) found that between 35 percent and 60 percent of persons failed to report an unemployment spell one year after the event. In both studies, failure to report a spell of unemployment was related, in part, to the length of the unemployment spell; short spells of unemployment were subject to higher rates of underreporting.
With respect to reporting the duration of a spell of unemployment, there is some evidence that the direction and magnitude of response error is a function of the length of the unemployment spell. For continuous spells of unemployment (that is, those that had begun in month \(x\) and which were ongoing at month \(x+1\)) Bowers and Horvath (1984) compared reports of spell duration to the actual amount of time that had elapsed between the two interviews. They found, on average, that the increase in the reported duration of the unemployment spell exceeded the actual elapsed time between interviews. Torelli and Trivelato (1989) used a similar approach for a quarterly survey and found that approximately 40 percent of the reported spell durations were consistent with the actual elapsed time and that the magnitude of response error was a function of the actual length of the spell. Specifically, they found that the longer the duration of unemployment, the greater the propensity to underreport the duration. Approximately one-third of the inconsistent reports was attributed to rounding by the authors. Poterba and Summers (1984) also find that the increase in spell length between interviews is smaller for those with longer durations of unemployment.

The findings suggest that, similar to other types of discrete behaviors and events, the reporting of unemployment is subject to deterioration over time. The passage of time alone however may not be the fundamental factor affecting the quality of the reports. Some evidence suggests that the complexity of the behavioral experience is a significant factor affecting the quality of retrospective reports. Both the micro-level comparisons as well as the comparisons of population estimates suggest that behavioral complexity interferes with the respondent’s ability to accurately report unemployment for distant recall periods. Hence we see greater underreporting among population subgroups who traditionally have looser ties to the labor force (teenagers, women). Although longer spells of unemployment were subject to lower levels of errors of omission, a finding that supports other empirical research with respect to the effects of salience, at least one study found that errors in reports of duration were negatively associated with the length of the spell. Whether this is indicative of an error in cognition or an indication of reluctance to report extremely long spells of unemployment (social desirability) is unresolved.

### 6.6 Industry and Occupation

The measures discussed thus far are ones in which discrepancies between the gold standard, whether administrative records or reports obtained from preferred designs, have been attributed to the response process. In that process, the respondent, the interviewer, and the question wording as well as the content of the questionnaire can all contribute to that which is often labeled “response” error. Evaluation of error associated with the measurement of industry and occupation must consider yet another factor which could contribute to the overall quality of a measure, the error potentially introduced through the coding process. The literature on response error, however, contains little discussion of the extent to which coding (as well as other post data collection processing) contributes to the overall error associated with a particular measure, or specifically with the classification of industry and occupation. Therefore, in the discussion that follows, the reader is cautioned that although disagreement between household reported industry and occupation and administrative
records is often classified as response error, coding/classification errors most likely contribute to the overall level of discrepancy.

Based on the small set of studies which have examined the quality of industry and occupation reports, the findings presented in Table 6 indicate that, in general, industry is reported more accurately than occupation. For both industry and occupation, not surprisingly, the agreement rate between employees’ and employers’ reports classified according to a single-digit coding scheme are higher than the resulting reports categorized according to the more detailed three-digit classification. Mellow and Sider (1983) report agreement rates between 84 percent and 92 percent for industry classification and between 58 percent and 81 percent for the classification of occupation (three-digit and one-digit classification schemes, respectively) in their Current Population Survey sample. Agreement rates are lower in the EOPP data, but Mellow and Sider indicate there is reason here to doubt the accuracy of the record report. Brown and Medoff (1996) compared industry classification of workers’ reports to the SIC codes for the employer, as listed by Dun and Bradstreet. Using fourteen industry groups, their comparison yielded an agreement rate of 79 percent. The findings from Mathiowetz (1992) are similar to those of Mellow and Sider, with occupational agreement rates of 52 percent to 76 percent, for three-digit and one-digit classifications, respectively. In the study by Mathiowetz, two sets of coders independently coded the reports of the employers and the employees while a third set of coders examined the two reports jointly to determine if the occupation could be considered the same occupation, that is, result in the same three-digit code. The direct comparison yielded an agreement rate of over 87 percent, suggesting that a significant proportion of the inconsistency in three-digit classification may be due to very subtle effects related to specific words used by the respondent to describe his or her occupation or used by the coder to classify the occupation.

For variables like industry or occupation with multiple classifications, the effect of measurement error on estimated parameters depends critically on the details of the discrepancies. For example, if those in high-wage industries mis-report themselves to be in other high-wage industries, the bias in estimating industry wage effects will be less than if the mis-reporters are spread randomly across the remaining categories. Angrist and Krueger (forthcoming, Table 11) calculate wage-weighted industry and occupation indices, based alternatively on worker and employer data, from Mellow’s CPS sample. In univariate regressions, the effect of the industry index is biased downward by 8 percent, and occupation by 16 percent; controlling for standard covariates like education, potential experience, race, and sex leads to estimated biases of 10 and 25 percent, respectively. Hence, the general finding that occupation is measured “less accurately” than industry does seem to translate into larger biases, at least when the relative size of the coefficients associated with the various categories are constrained in this way.

With respect to the reporting of occupations, the evidence of the deleterious effect associated with longer recall periods is mixed. Weiss, et al. (1961) report a decline in the agreement rate of occupational classification by employee's and employer's from 70 percent for the current occupation to 60 percent for occupations held more than four years prior to the date of the interview. Mathiowetz
(1992) found no effect on length of recall period in the agreement between household and employer reports of occupation. Agreement rates between the two data sources for occupations held one year prior to the interview were 49 percent (3 digit) and 74 percent (1 digit) compared to 52 percent and 76 percent, respectively.72

Given the difficulties in obtaining accurate measures of industry and occupation at one point in time, and the tendency of most workers to change industry and occupation only infrequently, there is general concern that measurement error will exaggerate the occurrence of changes in industry and occupation when estimates of such change are obtained by comparing reports of industry and occupation obtained at two points in time. The extent of such exaggeration depends on the extent to which the measurement errors are independent (for a given individual) over time, as well as the pattern of mis-classification (e.g., are those who in a high-wage industry but are mis-classified assigned to another high-wage industry). We have no direct evidence on the independence of such errors over time. Krueger and Summers (1988) assume an error rate for one-digit industries half as large as reported by Mellow and Sider (1983) (but with the same pattern of mis-classification as Mellow and Sider found), and assume such errors are independent over time.73 They find such a correction has a more important effect on estimated industry wage differentials estimated from two successive years of CPS data (in which true changes are relatively infrequent, and so the fraction due to classification errors is probably high) than in data from the Displaced Worker Survey (where true changes are more common, and so the fraction due to classification errors is lower).74 The CPS results point clearly to the value of evidence on pattern of errors in measuring changes in industry and occupation.

6.7 Tenure, Benefits, Union Coverage, Size of Establishment, and Training

In addition to questions concerning earnings, hours employed or unemployed, and industry and occupation, many labor-related studies query the respondent as to their employment tenure, union membership, establishment or firm size, and the nature of various employment-related benefits. Few

72Because the respondents in the study by Mathiowetz were older, with more tenure, than a nationally representative sample, these estimates should be seen as conservative estimates of the decline in the quality of reporting occupation associated with an increase in the length of the recall period.

73 Their taking half of Mellow and Sider’s rate is intended as a rough correction for the fact that the employer reports in Mellow and Sider’s include some errors, and for the fact that errors for an individual are probably not independent over time.

74 For example, the wage difference between workers in manufacturing and otherwise similar workers in retail and wholesale trade is .07 in the CPS before correcting for measurement error, and .23 after correcting. For displaced workers, the difference changes from .11 to .13.
Estimates of returns to tenure are about .002 higher (on a base of .01) in cross-section regressions when record rather than interview data are used. However, this difference is due to a correlation between tenure and errors in reporting earnings (Duncan and Hill, 1985).

Freeman also finds that, in two supplements to the May 1979 CPS (in which union coverage was asked in each), the responses given by workers are inconsistent 3.2 percent of the time. The inconsistencies were about equally distributed between those who said they were covered only on the first supplement and only on the second.
In a simple bivariate regression of wages on union coverage, random zero-mean errors in measuring coverage would lead to an estimate whose proportional bias is equal to the sum of the two mis-classification rates \( \text{Prob}(x=0|x^*=1) + \text{Prob}(x=1|x^*=0) \). This bias is .19 and .31 in Mellow’s CPS and EOPP data, respectively; .10 in Freeman’s and .12 in Card’s sample of Mellow and Sider’s CPS data. Freeman stresses the extent to which this bias is inflated in longitudinal analyses. If measurement errors are independent over time, and the misclassification rates sum to .10, the bias becomes 29 percent in a fixed-effect model if 19 percent of the sample reports changing union status (as is true over 1970-79 in PSID); with smaller fractions of the sample changing coverage status (as would be true in studying one-year changes) the bias would be larger still. Of course, if the errors are positively correlated over time, the bias due to measurement error in a fixed-effect model would be smaller than under independence.  

Card argues that, instead of treating the employer reports as “true,” one should treat both employer and worker reports as subject to measurement error. He finds that the estimated impact of union status on wages is very similar using either worker or employer report, whereas the latter should have a larger effect if only the worker reports were subject to error. Indeed, he argues that both the wage equations and the patterns of agreement across industry are consistent with the hypothesis that both worker and employer reports are equally prone to error, with error rates (independent of true union status) of 2.5 to 3.0 percent. His model makes the common assumption that the error in reporting union status is uncorrelated with the error term in the wage equation. This rules out a number of plausible scenarios; for example workers who are not aware of being covered by a union contract being those in weak unions which fail to deliver high wages.

There appears to be considerable disagreement on the accuracy of employee reports of various fringe benefits. Duncan and Hill (1985) also report high levels of agreement for reports of health insurance (less than 1 percent disagreement), dental benefits (5 percent disagreement, all underreporting by the respondent), life insurance (10 percent, all in the form of underreporting by the respondent), number of vacation days (less than 1 percent), and number of sick days (9 percent, split between over and under-reporting). In contrast, Barron, Berger and Black (1997) report disagreement rates of 35 percent and 25 percent (with respect to initial benefits) and 19 percent and 13 percent (with respect to benefits after two years) for sick pay and life insurance, respectively. Berger, Black, and Scott (1998) compare March CPS reports of employer-provided coverage to reports one or two months later in special CPS supplements. They find 11 percent of the reports are inconsistent, with lower overall coverage rates in the March surveys. Comparing employer and worker reports, they find that three fourths of the disagreements are workers who report they are eligible but whose employer reports them ineligible.

Indeed, Freeman reports that misclassification rates summing to .10 would predict more changes of reported union status due to error alone than one observes in his 1974-1975 CPS panel.
The focus of Mitchell’s (1988) research was the respondent’s knowledge of pension plan provisions. Using a match sample of household respondents and pension providers identified as part of the 1983 Survey of Consumer Finances, Mitchell finds pension misinformation as well as respondent’s inability to answer questions concerning pension benefits to be quite widespread. The highest rates of inaccuracy by household respondents concerned knowledge of early retirement provisions; one third of the respondents could not answer the questions and among those who did respond, less than one third understood (or more specifically, could accurately report) the requirements for early retirement benefits.

These errors are likely to be particularly damaging in structural models that relate retirement decisions to pension incentives. As Gustman and Steinmeier (1999) note, workers with defined benefit pension plans do much better if they leave the firm at the early retirement age rather than even one year earlier. Thus, mis-reporting the age of early retirement eligibility by even one year can make it look like an individual is retiring at precisely the age at which economic incentives suggest retirement should not occur, and lead researchers to severely underestimate the importance of pension incentives. Gustman and Steinmeier also note that workers may in fact base their behavior on their perceptions rather than “true” incentives; for such workers, survey responses may be a better approximation for the variable that motivates behavior than is the “true” variable as calculated from the pension plans.

Using Dun and Bradstreet data as the record for comparison, Brown and Medoff (1996) examined the quality of household reports of establishment and company size as well as age of firm (i.e., how long the firm had been in business). Correlations ranged from .56 (correlation between worker report and D&B report of age of firm) to .82 (for ln establishment size) and .86 (In company size). The authors note in their findings that potential inaccuracies in the Dun and Bradstreet records “probably understate the correlation between the worker reports and perfectly accurate measures of employer size.”

Only one study reported in Table 7 examines the accuracy of respondents’ reports of training hours. Barron, Berger, and Black (1997) compared workers’ and employers’ reports of hours of training for several different types of training: formal training, training by co-workers, and training related to others performing the job. The correlation between the two reports was highest for off-site, formal training (.457) and for total number of hours of training (.457) and lowest for informal training by managers (.176).

The empirical research concerning the quality of respondent’s reports of benefits, tenure, and industry characteristics is limited; in many cases we have only a single study to inform us as to the error properties of these measures. Although the findings suggest that the reporting of tenure and union coverage is highly correlated with administrative records, caution should be taken in drawing any conclusions from this limited literature. With respect to the reporting of fringe benefits, the findings are mixed. Based on the PSID-V, it appears that employees are well informed as to the
characteristics of benefits whereas the studies by Barron, Berger, and Black (1997) as well as Mitchell (1988) suggest high rates of inaccurate reporting.

6.8 Measurement Error in Household Reports of Health-Related Variables

While the examples discussed so far tend to be drawn from surveys that are most often used by economists, the empirical literature in several other substantive areas is rich with examples of the misreporting of autobiographical information. An important example is health, where work typically done by those in other fields provides evidence on the validity of health-related measures often used by economists and other social scientists.

6.8.1 Health Care Utilization, Health Insurance, and Expenditures

As previously noted, much of the early work with respect to the assessment of the quality of retrospective reporting by survey respondents focused on the reporting of health care utilization, usually as reverse record check studies in which respondents were sampled from those with known hospitalizations or visits to physician offices. The design of these studies makes them well suited for investigating errors of omissions; however, many of these studies are uninformative with respect to overreporting errors.

Table 8 presents the findings from a selection of studies assessing either the reporting of health care utilization, characteristics of health insurance, or health care expenditures. Once again, we find evidence that response errors appear to be a function of the nature of the response task facing the individual, the length of the recall period, and the salience of the information to be retrieved.

Two of the studies reported in Table 8 assess the quality of reports of hospitalizations. Cannell, Fisher, and Bakker (1965) describe a reverse record check study\textsuperscript{78} in which approximately 1500 respondents were asked to report on hospitalizations occurring during the previous 12 months. Overall, approximately 13 percent of hospitalizations were not reported. Response error, as measured by the percent of hospitalizations not reported by the respondent, increased as a function of the length of time between the date of the hospitalization and the date of the interview. For example, for hospitalizations occurring within 10 weeks of the interview, the underreporting rate was 3 percent whereas among hospitalizations occurring a year prior to the interview, 40 percent were unreported. The duration of the hospitalization was related to the rate of underreporting; 5 percent of longer hospital stays (e.g., those lasting 30 or more days) were unreported by the household respondent as compared to 26 percent of one-day stays.

\textsuperscript{78} The sample consisted of persons selected from hospital records as well as a supplementary sample of persons without hospitalizations, so as to blind the interviewers as to the purpose of the study.
Other studies have examined the quality of the reports related to utilization of office-based physician services. For example, Cannell and Fowler (1963) found that a significant proportion of office-based physician visits were unreported by the household respondent, even for recall periods as short as one week (15 percent unreported) and that the underreporting rate increased sharply with an increase in the reference period to two weeks (30 percent underreporting rate).

The Medical Economics Survey reported by Yaffe and Shapiro (1979) was designed to test the feasibility and effectiveness of several different survey design features to obtain information concerning health care utilization, expenditures, and health insurance coverage. The study included an assessment of face-to-face vs. telephone mode as well as monthly vs. bimonthly interviews over a six month data collection period. In addition to the monthly or bimonthly interview, respondents were asked to maintain a diary (after the initial interview) to serve as a record-keeping system and memory aid for subsequent interviews. Prior to each follow-up interview and at the end of the study period, a cumulative summary of previously reported information was mailed to each household. Respondents were asked to review the report and to make any necessary additions or corrections, including entries about bills received since the time of the last interview. All medical care providers identified by the respondent as having provided care for anyone in the family during the study period as well as providers identified as the usual source of care, were contacted after the household data collection.

Several of the design features, specifically, the multiple rounds of data collection, coupled with the relatively short reference period, the use of a household diary, and the use of a summary were all included so as to minimize response error. These design features may account, in part, for the higher levels of agreement reported in Table 8 for this study as compared to other studies. In addition, Yaffe and Shapiro only report agreement rates for population estimates, that is, \((Y_{hh}/Y_{med}) \times 100\), where \(Y_{hh}\) represents the population estimate based on the household report and \(Y_{med}\) represents the population estimate based on the medical records. The estimates are provided for the two distinct geographical areas studied. As can be seen from the table, agreement rates for utilization are quite high for the most salient events (and less frequent) such as hospitalizations and emergency room visits, with agreement rates in excess of 90 percent. Agreement rates were lowest for clinic visits, 39 to 54 percent. With respect to expenditures, we once again see a high level of agreement between the two data sources for hospitalizations (87 to 99 percent) and the lowest agreement rates for hospital clinic visits (31 to 38 percent).

Cohen and Carlson (1994), using data from the National Medical Expenditure Survey, also examined the quality of household reports of medical expenditures. The entries in Table 8 present the mean household estimate, the mean medical record estimate, the mean of the simple difference and the mean of the absolute difference between household and medical record reports of total expenditures for each of four categories of utilization. The sample sizes provided in the table represent the number of events on which the estimates are made; the percent indicates what proportion of all household events of that type are included in the analysis. Due to the design of the NMES (which included a medical record component for a sample of all households) as well as
provider nonresponse and inability to match events reported by the household with events abstracted from medical records, not all events reported by the household respondent were included in the analyses. In addition, the analysis is limited to those events for which there was expenditure data from both the household and medical record files. The comparison of the two data sources indicate that although the simple differences tend not to be statistically significant, the absolute differences clearly indicate significant disagreement between the two data sources.

Very few studies have examined the ability of household respondents to report detailed information concerning features of their health insurance. Knowledge of the existence of out-of-pocket payments and sources of premium payments was quite high (78 percent and 74 percent, respectively), but quite low with respect to amounts of out-of-pocket payments and amount of insurance premiums paid by others (less than 30 percent for each) (Walden, Horgan, and Cafferata, 1982). As we would expect, the majority of respondents were able to accurately report the standard major categories of coverage (hospital room, physician in-patient surgery, other in-patient physician services, and dental services). Knowledge of coverage associated with richer benefit plans was much lower, however, with less than one-third of the respondents correctly identifying whether or not their insurance covered outpatient mental health, inpatient mental health or nursing home services.

6.8.2 Health Conditions and Health/Functional Status

Measurement error in health surveys is not limited to the reporting of utilization, expenditures, and health insurance characteristics, but is also evident in the reporting of medical conditions as well as the reporting of health and functional status. Findings from a sampling of the literature which addresses the validity and reliability of self reports of health conditions and functional status are presented in Table 9.

In two reverse record check studies (National Center for Health Statistics, 1961 and 1967), respondents were asked to report on the prevalence of chronic conditions. The second study also included an experiment designed to address the difference in the quality of data obtained from free recall as opposed to recognition from a checklist of conditions. The findings from these studies suggest that underreporting is a function not only of the length of the recall period (measured as the time since the last physician visit related to the condition), but also of the response task. Questions which frame the task as one of recognition as opposed to free recall resulted in lower rates of underreporting. However, for both response tasks, the underreporting rate was quite high, ranging from 32 percent underreporting for the recognition task related to the events occurring within the previous two weeks to an underreporting rate of 84 percent for free recall of events occurring four or more months prior to the interview. The improved reporting related to the recognition task is predictable; the presence of a cue provides both additional context for the respondent to understand the goal of the questions and an additional means for accessing the associated network of memory. The study by Madow (1973) is a complete record check design, limited to respondents in a specific health plan. As can be seen from the table, almost half of the conditions recorded in the medical
records were not reported by the household respondent whereas over 40 percent of the conditions reported in the household interview were not identified in the medical record.

As part of the National Medical Care Expenditure Survey (NMES), Johnson and Sanchez (1993) examined the congruence between medical conditions as reported by the household respondent and medical conditions as reported by the medical care provider. These data are based on the same matched sample of household reported events and provider reported events used by Cohen and Carlson (1994) in their analyses of the quality of household reports of health care expenditures. Household reports reflect conditions associated with hospitalizations, visits to emergency rooms, outpatient departments, as well as office based physician visits. Household reported conditions, which reflect a mix of self and proxy collected information, were coded to three-digit level of detail by experienced coders using the International Classification of Diseases, Version 9 (ICD-9). ICD-9 condition codes were abstracted from the medical records, independent of the knowledge of the condition described by the respondent. Household reports of utilization were linked to the medical record abstracted records via a probabilistic match function. One of the variables used in the probabilistic match was a one-digit collapsed classification of the condition related to the utilization. As a result, the agreement rates—which indicate the percent of medical events reported by the household respondent for which the two condition codes (household based and medical record based) agree—are likely to be optimistic. At the three-digit level of detail, there is agreement between the condition codes as reported by the household and the medical condition recorded in the medical records for less than half of the medical events. As we would expect, grosser levels of aggregation result in higher rates of agreement.

While the lack of congruence between survey data and medical records is disturbing, we want to emphasize that this information alone tells us very little about the effect of this measurement error on parameters estimates. First, it seems plausible that reporting errors decline with the severity of the condition (severe arthritis is more likely to be reported that is mild arthritis). Second, it many cases researchers will be interested in modeling jointly effects of various conditions on outcomes. In such cases it is hard to say much about either the magnitude or the direction of the bias on a single coefficient, since the coefficient on any one condition will be biased not only by the under and over-reporting of that condition, but also by the under and over reporting of other conditions (see the discussion in 2.2).

Table 9 also examines the reliability, and to the extent possible, the validity of several measures of health and functional status. The measures examined include the Index of Activities of Daily Living (Katz, Ford, Moskowitz, Jacobsen, and Jaffe, 1963), the Sickness Impact Profile (Bergner, Bobbitt, Kressel, Pollard, Gilson, and Morris, 1976), and the SF-36 (Ware, Snow, Kosinski, and Gandek, 1993). In contrast to the validation studies presented earlier, no external measure of validity exists for the majority of the measures related to health or functional status. Rather, as with most psychometric scales, the interests lies in the reliability of the measure (that is, test-retest reliability or internal consistency) or the validity of the index, measured as the correlation or consistency with other subjective scales.
Despite its broad use, there has been little published with respect to the assessment of the validity or reliability of the Index of Activities of Daily Living, especially within the general population. Katz, Downs, Cash, and Grotz (1970) applied the Index of ADLs as well as other indexes to a sample of patients discharged from hospitals for the chronically ill and report a correlation between the index and a mobility scale and a confinement measure of .50 and .39, respectively. Most assessments of the Index of ADL have examined the predictive validity of the index with respect to independent living (e.g., Katz and Akpom, 1976) or length of hospitalization and discharge to home or death (e.g., Ashberg, 1987). These studies indicate relatively high levels of predictive validity.

Despite these findings, there is a growing body of literature that suggest that the measurement of functional limitations via the use of ADL scales is subject to substantial amounts of measurement error and that measurement error is a significant factor in the apparent improvement or decline in functional health observed in longitudinal data. For example, Mathiowetz and Lair (1994) found that conditions of the interview, characteristics of the interviewer, and type of respondent (self or proxy) were predictive of improvement in functional status over the 18 months of interest whereas the individual’s demographic characteristics and health status were indicative of decline in functional status. Rodgers and Miller (1997) examined the consistency with which respondents reported functional limitations, using alternative sets of question wording. Consistent with other findings in the literature, they found that minor differences in the wording of questions resulted in significant variation in the proportion of respondents identified as being limited in one or more functional activities, ranging from a low of 6 percent (based on a single question) to more than 25 percent of the respondents79 (based on a set of six to nine ADLS questions).

The Sickness Impact Profile (SIP) measures health status by assessing the way sickness changes daily activities and behavior and consists of 136 statements grouped into twelve categories of activities. The profile focuses on actual performance as opposed to capacity. Bergner, Bobbitt, Carter, and Gilson (1981) report on the reliability of the profile for both interviewer administered questionnaires and self-administered forms, with reliability higher for the interviewer administered form (.97) than for the self-administered form (.87). Internal consistency, as measured by Cronbach’s

79Rodgers and Miller’s study is based on the respondents to the first wave of the AHEAD study.
alpha was similarly lower for the self-administered form (.81) than for the interviewer-administered form (.94).

The SF-36 is a generic health status measure, one that is not specific to age, disease, or treatment, that focuses on health-related quality of life outcomes. The index covers eight areas of health: physical functioning, role limitations due to physical health problems, bodily pain, general health, vitality, social functioning, role limitations due to emotional problems, and mental health. The measure is designed for both interviewer administration as well as self-administration and both modes of data collection have been assessed with respect to validity and reliability. Reliability of the SF-36 has been assessed in numerous studies (see Ware et al., 1993 for summary of these studies); across the various scales of the SF-36 and across the various studies, the median of the reliability coefficients equals or exceeds .80 (Cronbach’s alpha). The findings from two of the more recent studies examining the SF-36 are reported in Table 9. McHorney, Ware, Lu and Sherbourne (1994) examined the internal consistency of the SF-36 among approximately 3500 patients with one or more chronic conditions; as can be seen from the table the coefficients range from .78 for general health to .90 for mental health. A self-administered version of the questionnaire study conducted among a nationally representative sample of noninstitutionalized adults found similarly high measures of internal consistency (McHorney, Kosinski, and Ware, 1994).

6.9 Education

Despite the importance of schooling as both an outcome and as an explanatory variable in economic models, relatively little effort has been devoted to assessing the accuracy of survey reports of years of schooling or similar measures of educational attainment. The literature which is available, however, illustrates a number of interesting issues that are potentially relevant for other variables as well.

Typically, these studies (summarized in Table 10) have two interview-based measures of education, each of which is plausibly measured with error. In assessing what we can learn from such data, recall that the OLS bias in estimating $\beta$ in the model $y = \beta x^* + \varepsilon$ when instead of $x^*$ we use $x_1 = x^* + \mu_1$ depends on $1 - \lambda_1 = 1 - (\sigma_{x^*x_1}/\sigma_{x^*_1})$. If we have another measure of $x^*$, $x_2 = x^* + \mu_2$ then

---

80 Cronbach's alpha provides an estimate of internal-consistency reliability based on the average inter-item correlation and the number of items in the scale, expressed as $k r/[1+(k-1)r]$ where $k$ equals the number of items in the scale and $r$ is the average correlation between items. The coefficient alpha will be higher (1) the more questions asked about the topic and (2) the higher the average correlation between the scores for all possible combinations of the entire set of questions. In most applied studies, the lowest acceptable level of internal consistency reliability is .70 for group data and .90 for individual-level analysis (Nunnally and Bernstein, 1994).
\[ \lambda_1 = \frac{\sigma_{x_1 x_2}}{\sigma_{x_2}^2} \] as long as \( \mu_2 \) is uncorrelated with \( x^* \) and with \( \mu_1 \). In other words, as long as the error in measuring \( x_2 \) is “classical” whether \( x_2 \) is itself a particularly reliable indicator of \( x^* \) is unimportant. If, in contrast, \( \mu_1 \) and \( \mu_2 \) are positively correlated, the covariance between the two measures of \( x^* \) will overstate \( \lambda_1 \), and holding that correlation constant, the larger the measurement error in \( x_2 \) the worse the overstatement will be.

An early study of the reliability of reported years of schooling is Siegal and Hodge’s (1968) analysis of 1960 Census data. Validation data came from the Post-Enumeration Survey (PES), a re-interview conducted to assess the accuracy of the original Census reports. They found that the Census reports and PES data on individual years of schooling are highly correlated. They also noted, however, that the variance of the Census report is slightly smaller than that of the PES education variable, which is inconsistent with the usual assumption that the Census report is equal to the true (PES) variable plus an uncorrelated measurement error. The discrepancy between the two reports was in fact negatively related to the PES value (\( r = -.20 \)). They argued that one should expect errors to be negatively related to true values for bounded variables, since for those with the highest (lowest) true level of education, errors must be negative (positive). Given that the variances of the Census and PES variable are essentially equal, the \( b_{\text{PES,Census}} = .93 \), so the bias due to errors in measuring education as an explanatory variable is small (as long as other explanatory variables are not highly correlated with education).

Siegal and Hodge (1968) recognized the possibility that the PES measure of education is also measured with error and considered several relatively elaborate models in which both years of schooling and income are mis-measured. These relied on rather arbitrary identifying assumptions, and Siegal and Hodge concluded “we have not been able to devise an entirely plausible solution.”

Bishop (1974) presents a comprehensive summary of the reliability of Census and CPS reports of education. Estimates of the correlation between Census and other measures of education center on .9, as do the alternative estimates of \( \lambda_1 \). Bishop notes that mean reversion would tend to reduce the bias caused by measurement error, while positive correlation in the errors would lead the values of \( \lambda_1 \) to be too high.

Bielby, Hauser, and Featherman (1977) compare Current Population Survey reports to subsequent interviews and re-interviews of the same households approximately six to seven months later as part of the Occupational Change in a Generation (OCG) study. Focusing on the sample of non-black males that participated in both the OCG interview and re-interviews, they find inter-correlations among the three measures of years of schooling of .80-.92. The OCG shows both lower correlation with the other two measures and higher variance, suggesting it is the least reliable of the three measures. A rather complicated measurement model—which allows errors to be correlated with true values for OCG and OCG-R but not CPS,\(^{81}\) and assumes errors in the three measures are

\(^{81}\) Bielby et al. argue that with true scores unobserved, the units of the “true” variable are arbitrary and regard the unit coefficient on the CPS measure as a normalization. Their estimates
suggest a slight positive correlation between error and true value for the two OCG measures.

If we take the estimates of the first three studies in Table 10 at face value, their implication is that biases in estimating the effect of education on other variables due to errors in measuring years of schooling are not likely to be large. There are, however, two important qualifications: (i) taking these estimates at face value means assuming that the errors in the alternative reports are (at least roughly) uncorrelated (ii) as noted in section 2, biases due to measurement error become more important if other (well-measured) explanatory variables are correlated with years of schooling.

A relatively extreme context for illustrating the latter point are recent “twin” studies that relate wage or earnings differences between twins to differences in their schooling. In effect, this strategy for estimating returns to education adds a set of dummy variables, one for each pair of twins, to a standard wage or earnings equation. Such between-twin differencing has much the same effect as the first-differencing in panel data—most of the variation in schooling is between rather than within twin pairs, and if reporting errors are not highly correlated the reliability of differences in education within twin pairs is likely to be lower than the reliability of reports of education in general.

Ashenfelter and Krueger (1994) obtained the usual information on wages and schooling in a sample of twins, and each sample member’s report of the years of schooling completed by his or her twin. This report of one’s twin’s schooling is highly correlated with the twin’s own report ($r=0.9$); assuming (as Ashenfelter and Krueger do) that errors in their own and twin reports are uncorrelated, this correlation is consistent with the reliability estimates in the earlier literature. However the correlation between twin 1’s report of own schooling minus twin 2's report of own schooling and twin 2's report of 1's schooling minus twin 1's report of 2's schooling is only $0.57$ in their sample of MZ (monozygotic, or “identical” twins) and $0.74$ in a small sample of DZ (dizygotic, or “fraternal” twins). This suggests, for the MZ twins, that estimates of returns to schooling based on differencing wages and schooling between twins are likely to underestimate the true returns by over 40 percent. IV estimates, using the difference in reports of twin’s schooling as an instrument for one’s own reports, are consistent with this calculation.$^{82}$

The assumption that reporting errors are uncorrelated with each other is subject to challenge on a number of grounds. First, one might anticipate that the error made, for example, by twin 1 in

$^{82}$ The IV estimate reproduces this calculation if the maintained assumption that the covariance between the difference in wages and the difference in years of schooling is the same using either measure of the difference in years of schooling. As Ashenfelter and Krueger note, this is approximately true in their data.
reporting own schooling would be positively related with the error in twin 2’s report of 1’s schooling, so that errors in the own- and cross-reports of the difference in schooling would be positively related. This would lead the covariance between differences in years of schooling based on own reports and on twin reports to be greater than the variance of the true difference, and the bias due to measurement error understated by the classical model. A second possibility is that errors in one twin’s report of own and twin’s schooling are positively related. This would imply that twin 1’s report of own schooling would be more highly correlated with his/her report of 2's schooling than with 2's report of own schooling, and the data support this conjecture. Ignoring such a correlation would lead the standard correction for the bias due to measurement error to be too large.

A solution to the second problem is to use one twin’s report of the difference in schooling as an instrument for the other’s report. This leads to a downward revision, as expected, in the estimated return to schooling. Behrman and Rosenzweig (1999), in contrast, find no evidence in their sample from the Minnesota Twin registry that errors in reports of own and twin’s schooling are correlated, and so find estimated returns to schooling are unaffected by allowing for such a correlation.

A subsequent paper by Rouse (1999), using four waves of twin surveys rather than the first wave used by Ashenfelter and Krueger, found somewhat different substantive results but quite similar conclusions as regards the importance of measurement error in the schooling variable.

Miller, Mulvey, and Martin (1995) conducted a similar analysis using a larger sample of Australian twins. Their findings differ from the U.S. twin studies in two respects. First, the correlation between the difference in own reports and the difference in twin reports of education is substantially lower, at least for MZ twins. Second, the variance of schooling using twin reports is

---

83 While we lack a firm understanding of the situations which lead to errors in reporting schooling, it seems reasonable that there would be certain situations in which errors are particularly frequent, and if there is an error it is particularly likely to go in one direction. For example, if one ends one’s schooling after a not-particularly-successful sophomore year of college, “true” years of schooling might be 17, with the most likely error reporting 18 instead. If twin 1 is in this situation, both twin 1 and twin 2 would be more likely to over-report schooling (by one year) than to make some other error.

84 Unlike other twin studies, Ashenfelter and Krueger (1994) found that a first-differenced specification (not corrected for measurement error) leads to larger estimates of the returns to schooling than is obtained without fixed (twin) effects; Rouse’s larger sample reaffirms the conventional wisdom in this regard.

85 Ashenfelter and Rouse (1998) use the first three waves of the twin survey; their correlations between own and twin reports are very similar to those from Rouse’s study which uses four.
lower than using own reports. This suggests that the twin reports are more accurate or the errors are more mean-reverting, neither of which seem likely on a priori grounds.

Kane, Rouse, and Staiger (1999) return to the “standard” framework for estimating wage equations, simple cross-sections with no (identifiable) twins. They focus instead on the assumption that the error in reporting years of schooling is unrelated to the true value. As noted above, for binary variables (e.g., has graduated from college vs. has not graduated), any error must be negatively related to the true value. The same sort of negative correlation is likely (though not inevitable) for bounded variables such as schooling.

Kane, Rouse, and Staiger (1999) analyze schooling as reported by respondents in the National Longitudinal Study of the Class of 1972, virtually all of whom graduate from high school. Their focus is on reports of education beyond high school, as reported by NLS72 respondents and as recorded in transcripts of all post-secondary schools they reported attending (which were collected as part of the NLS72 study). While one might be tempted to take the latter as an indicator of “true” schooling, internal evidence suggests this is unlikely: holding constant BA receipt or non-receipt according to the transcript data, those who self-report having one earn higher wages than those who do not (and, less surprisingly, holding constant self-reported BA status, those who have a BA according to the transcript data earn higher wages than those who do not).

This provides the basis for a method-of-moments estimation strategy that does not rely on the standard IV assumption that measurement errors are correlated with true values. Kane, Rouse, and Staiger (1999) do, however, maintain the standard assumption that errors in reporting schooling are uncorrelated with wages, with each other, and (in models with covariates) with the covariates. In the simplest case, with schooling a binary variable and no covariates, there are seven unknowns: the intercept and BA-premium in the ln-wage equation, the true probability of having a BA, and four parameters of the “measurement” model (which has each measure of schooling as a linear function [with intercept] of true schooling). There are also seven observable means or sample proportions: if we define a two-by-two table for combinations of self- and transcript-reported BA status, there is one mean ln wage in each of these cells and four (but only three independent) sample proportions. This equivalence provides the basis for jointly estimating wage equations and the measurement model by GMM. Kane, Rouse, and Staiger show how this intuition can be extended to many educational categories, and to include covariates (which lead to the model being over-identified).

Substantively, they find that most differences between self-reports and transcript data—and most of the error, according to their GMM estimate—occur where one or the other of the reports claims some college, but less than a BA degree. This means that the extent to which OLS under-estimates and traditional IV overstates the return to schooling is largest as a proportion of the true value for those reporting some college. According to their estimates, OLS is less than the GMM estimate of returns to some college and a BA by about .02 (on a base of .125 and .308, respectively) while IV over-estimates each return by about the same amount (Kane, Rouse, and Staiger, 1999, Table 6).
On balance, the studies in Table 10 support four general conclusions. First, evidence on the reliability of survey reports of educational attainment rely more on multiple measures, each of which is likely to contain non-negligible error, and less on direct validation evidence than is true for most of the other variables considered in this paper. Second, unless there is substantial positive correlation in these multiple measures, the bias due to errors in measuring years of schooling in traditional applications such as cross-sectional earnings functions is unlikely to be large. Third, while it is generally assumed that the errors are uncorrelated with each other and with the dependent variable (typically, ln wage or ln earnings), there is no direct evidence on this score. Most discussions in the literature treat positive correlations as the most likely alternative; if positive is more likely than negative, there is every reason to fear that positive is more likely than zero. Fourth, here as elsewhere, differencing (in this case, differences within twin pairs) greatly exacerbates the bias due to errors in measuring schooling, but the availability of reports of one’s twin’s schooling as well as one’s own provides some leverage for undoing such bias.

7.0 Conclusions

Empirical research in economics has increasingly used individual- or household-level data derived from surveys. Unlike aggregate data based on surveys where one might hope that the errors would “cancel out,” the move to micro data requires a continuous concern about measurement error as a likely source of bias. Some variables (transfer income, wealth holdings, medical care utilization and expenditures) are sufficiently difficult to measure that such concerns would arise even in estimating simple bivariate regressions; others (union coverage, schooling, and perhaps earnings) that seem to be reported with reasonable accuracy become candidates for concern when panel data are used in ways that effectively difference out much of the true variation while increasing the noise.

The impact of measurement error on parameter estimates depends on the magnitude of the error relative to the true variation, but more generally on the joint distribution of the measurement errors and the true variables. If we are going to use data on X and y in order to study the impact of X* on y*, in principle we need to know the entire data-generating mechanism; that is., f(y, X, y*, X*).

---

86 Our comments should not be taken to suggest with think aggregate data is without significant error. While response errors are presumably less important in aggregate data than they are in individual- or household-level survey data, there are certainly other important sources of error. Many aggregate series (e.g. unemployment rates) are based on survey data and, as such, are subject to sampling error. More fundamentally, much aggregate data is constructed using procedures that are likely to introduce systematic error into data series. Thus, for example the Department of Commerce’s Bureau of Economic Analysis (BEA) uses procedures to construct value added (Peterson, 1987) that, outside of manufacturing and a few other industries are likely to underestimate the growth in output and thus productivity (Griliches, 1994) and to create spurious correlations between input growth productivity growth (Waldmann, 1991). Any discussion of such issues is well beyond the scope of this chapter.
Standard methods for “correcting” for measurement error such as instrumental variables procedures typically involve strong assumptions regarding the nature of the data-generating mechanism (i.e. that errors are classical) that are rarely discussed or defended. Short of detailed knowledge of the data-generating mechanism, the theoretical literature suggests that when the correlations between our measures and our constructs is high and when our models are simple, we can be reasonably confident regarding the robustness, in qualitative terms, of our inferences. This is the situation where standard methods for correcting for measurement error have little effect on our estimates. In contrast to this, in situations were we have reason to believe that measurement error on key variables is sufficiently large as to have qualitative effects on our estimates, serious sensitivity analysis is in order.

Validation data has provided considerable evidence on the magnitude of measurement error. Gradually, the focus has shifted from the extent of under- or over-reporting (i.e., on the mean error) to the ratio of the variance of the reporting error to the variance of the true value, and more recently to consideration of whether errors are, as is so often assumed, uncorrelated with true values. Such evidence as we have suggests that errors are often negatively related to true values and, indeed, this must be so for binary variables. Fewer studies focus on the correlation between errors in measuring one variable and either measured or true values of other variables. The very limited evidence we have suggests that such correlations do not lead to appreciable or predictable biases except in contexts where variables are definitionally related (e.g., hours worked per week and earnings per hour defined as weekly earnings/weekly hours).

Despite the effort that has gone into validating various survey measures it is striking to us how little is known about the accuracy of much of the data that is routinely collected in household surveys. To take a simple example, there is no hard evidence on how reliably hourly earnings are reported for men and women paid by the hour. Nor is there much data on the accuracy with which individuals report wealth or consumption expenditures. In other contexts, such as for health conditions, we know something about the accuracy of such reports, but no virtually nothing about the impact that mis-reporting has on parameter estimates. Similarly, there are many studies of the accuracy of retrospective reporting of events, but few clues as to how the (often important) errors found in such studies will bias parameter estimates of event-history studies.

Increasing use of panel data has been accompanied with a heightened awareness of the tendency of such estimation to increase the importance of measurement error. The panel-data literature has benefitted from simple, intuitive results that alert analysts to situations where such errors are likely to be most harmful. Unfortunately, even the most rudimentary corrections for measurement error in such contexts depend on knowing the correlation between errors—for an individual’s wage over time, for twins’ reports of their education, etc.—and there is almost no direct evidence on such correlations. Obtaining validation data sufficient to calculate such correlations requires at least two rounds of survey data and either two rounds of validation data (e.g., the PSID Validation Study) or the good fortune to be able to obtain validation of two rounds of survey data in a single step (e.g., the matched CPS-SSA data, and matches of transfer program records to SIPP data). Hopefully, in the future, it will be possible to merge administrative data to existing panel data.
As with panel data, there is good reason to fear that parameter estimates in non-linear models are likely to be more sensitive to measurement error than those in simple (linear) models. Unfortunately, the analysis of non-linear models has proceeded on a case-by-case basis, and it has not highlighted any key feature of the error distribution for validation studies to assess. Thus, analysts must often choose between less ambitious linear models for which the consequences of measurement error is better understood and more elaborate models which may well be more vulnerable to such errors. At a minimum, assessment of the relative benefits of the two approaches needs to put greater weight on this vulnerability.

One reason for remaining gaps in our knowledge about the inaccuracies of survey data is that users of the data are rarely involved in the validation studies. As a result, it is natural for them to focus on the accuracy of the reports rather than the effect of inaccuracies on parameter estimates. Since different researchers are interested in different parameters, researchers conducting validation studies will never be able to satisfy all audiences. However, researchers can sometimes make their data publically available. It is interesting to note that both the CPS-SSA data by Bound and Krueger and the PSID-V data have been put to very good use by researchers outside the teams that originally developed the two data sets (e.g., Bollinger, 1998; Pischke, 1995; French, 1998; Brownstone and Valetta, 1996). In addition, there are clear payoffs to greater involvement of users in the design of validation studies.

While in general we believe that more effort devoted to collecting and analyzing validation data would significantly enhance the value of survey data, it is important to recognize the limitations of such initiatives. Those collecting validation data usually begin with the intention of obtaining “true” values against which the errors of survey reports can be assessed; more often than not we end up with the realization that the validation data are also imperfect. While much can still be learned from such data, particularly if one is confident the errors in the validation data are uncorrelated with those in the survey reports, this means replacing one assumption (e.g., errors are uncorrelated with true values) with another (e.g., errors in survey reports uncorrelated with errors in validation data).

Many of the validation studies reported in this chapter are based on small convenience samples (workers in a firm which cooperates by providing payroll records, households with accounts at cooperating financial institutions). The use of small samples means the reliability of the data is itself assessed with considerable sampling error. Moreover, the distribution of the variables of interest may well differ in the smaller validation sample and the large sample about which one wishes to make inferences (e.g., true wage variation will be smaller within one firm than in the economy). Even when validation data is provided for a sizeable share of a larger survey, concerns about representativeness are hard to dismiss (are those who under-report transfers less likely to cooperate in validating their responses?)

---

87 These comments echo somewhat similar comments often made by Griliches (e.g. 1984, 1994) that economists should become more involved in the generation of the data they use.
A final limitation of validation studies is that, even if the validation corresponds exactly to the “correct” answer to the survey question, it may not correspond to the “true” value of the variable in question. On the one hand, the construct we wish to test may be more subtle than questions that our surveys can ask. For example, earnings presumably depend on the interaction of years of schooling, school quality, and student effort that produce “education” or “learning”; the gap between “education” and “years of schooling” will remain no matter how successful we are in inducing individuals to accurately report their years of schooling. On the other hand, in some cases it may be the respondent’s perception of a variable rather than the “true” value of the variable that motivates behavior. Thus, for example, savings behavior of smokers may depend on their own estimate of their life-expectancy, not the Surgeon General’s.

It is widely recognized that survey data—and, indeed, other types of data—are often imperfect. Analyzing such data requires an understanding of their most significant shortcomings. Validation data are often imperfect, too. But they give important clues about these shortcomings—clues that would otherwise be unavailable—and suggest strategies for dealing with them. As econometricians create more complicated tools, understanding the effects of imperfect data on the performance of these tools becomes more important. Validation studies are an essential part of that enterprise.
References


Cannell, C. and F. Fowler (1963), "A Study of Reporting of Visits to Doctors in the National Health Survey" (Survey Research Center, Ann Arbor MI).


Duncan, G. and N. Mathiowetz. (1985), A Validation Study of Economic Survey Data (Survey Research Center, University of Michigan, Ann Arbor).


Griliches, Z. and V. Ringstad (1971), Economies of Scale and the Form of the Production Function: An Econometric Study of Norwegian Manufacturing Establishment Data ( North-Holland, Amsterdam)


Little, R. J. and D. B. Rubin (1987), Statistical Analysis with Missing Data (Wiley, New York)


McHorney, C., M. Kosinski, and J. Ware (1994), “Comparison of the Costs and Quality of Norms Collected by Mail versus Telephone Interview: Results from a National Study,” Medical Care 32: 551-567.


### Table 1. Assessment of Measurement Error: Earnings

<table>
<thead>
<tr>
<th>Reference</th>
<th>Variables of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keating, Paterson, and Stone (1950)</td>
<td>Weekly wages on jobs held in previous year (Survey of currently unemployed)</td>
<td>Employer’s records</td>
<td>r(interview, record) = .90 (men) and .93 (women)</td>
</tr>
</tbody>
</table>
| Miller and Paley (1958) | Annual earnings (Decennial census post-enumeration survey) | IRS tax forms | • Receipt of earnings/wages: under-reported at 2% to 6%  
  • Comparison of median income indicates small (1%) net bias; under-reporting for families (3%), over-reporting for unrelated individuals (4%) |
| Hardin and Hershey (1960) | Weekly earnings (salaried workers) | Employer’s records | r(interview, record) = .98 for men and .99 for women  
  • Those who had recently received a raise were more likely to under-report their earnings |
  Mean (employer report): $63.98  
  Mean (simple difference) $3.39  
  r(household, employer) = .95  
  • Difference higher for males and those with higher reported earnings and hours; and for older workers and those with more education. |
  Work history reports: mean error=46.67, s.d.=623.49  
  Broad questions: mean error=$38.57, s.d.  767.14  
  • Over 15% of responses misreported $1,000 or more  
  • Work history approach resulted in smaller response errors among those with some college education and for persons with average or above average earnings while the broad questions resulted in more accurate data among poor persons with no college education |
| Dreher (1977) | Monthly salary (nine $500 intervals) | Employer’s records | r(interview, record)=.91 |
| Carstensen and Woltman (1979) | Rate of pay, usual weekly earnings (CPS special supplement) | Employers’ report | • Those reporting pay per hour:  
  Mean (household): $4.21  
  Mean (employer): $4.44  
  Mean (difference): -$0.23 (s.e. = $0.02)  
  • Those reporting pay per week:  
  Mean (household): $203  
  Mean (employer): $217  
  Mean (difference): -$14 (s.e. = $2.70)  
  • Those reporting pay per month:  
  Mean (household): $1173  
  Mean (employer): $1068  
  Mean (difference): $104 (s.e. = $14.90)  
  • Those reporting pay per year:  
  Mean (household): $16,868  
  Mean (employer): $16,068  
  Mean (difference): $800 (s.e. = $403)  
  • When pay reported per hour, both self- and proxy reports have small mean error; when pay reported per week, self-reports have small mean error but proxy reports are 20% below true values. |
| Greenberg and Halsey (1983) | Quarterly earnings (Participants in Gary and Seattle-Denver income-maintenance experiments) | Employer reports to state unemployment-insurance agency | • Controls in S-D experiment slightly over-reported earnings whereas Gary controls significantly under-reported earnings (28, 37, and 36 percent for husbands, wives, and female “heads” respectively)  
  • Those eligible for experimental income-maintenance payments tended to under-report earnings (except for husbands in S-D).  
  • Earnings difference between experimentals and controls exaggerated by misreporting for all groups, ranging from 2-3 percent of earnings (husbands at both sites) to 16 percent (young non-heads in S-D). |
| Mellow and Sider (1983) | Wage per hour (CPS) | Employers’ Records | • ln (employer reported wage) - ln (worker reported wage):  
  mean = .048; variance = .167  
  • Regression with employer-worker wage difference as the dependent variable, no significant coefficients  
  • Wage equations based on employer vs. worker reported wages indicated no difference in structure of wage determination |
<table>
<thead>
<tr>
<th>Duncan and Hill¹ (1985)</th>
<th>Annual earnings, year t and t-1 (PSID Validation Study)</th>
<th>Employer’s Records</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• 1982 Earnings:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Annual Earnings:</td>
<td>Hourly Earnings</td>
</tr>
<tr>
<td></td>
<td>Mean (interview): $29,917</td>
<td>$16.31</td>
</tr>
<tr>
<td></td>
<td>Mean (record): 29,972</td>
<td>16.97</td>
</tr>
<tr>
<td></td>
<td>Mean (difference): -55</td>
<td>-.63</td>
</tr>
<tr>
<td></td>
<td>Mean (absolute difference): 2,313</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td>Error/Record variance ratio: .154</td>
<td>2.801</td>
</tr>
<tr>
<td></td>
<td>• 1981 Earnings:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Annual Earnings:</td>
<td>Hourly Earnings</td>
</tr>
<tr>
<td></td>
<td>Mean (interview): $29,579</td>
<td>$14.71</td>
</tr>
<tr>
<td></td>
<td>Mean (record): 29,873</td>
<td>15.39</td>
</tr>
<tr>
<td></td>
<td>Mean (difference): -294</td>
<td>-.66</td>
</tr>
<tr>
<td></td>
<td>Mean (absolute difference): 2,567</td>
<td>2.13</td>
</tr>
<tr>
<td></td>
<td>Error/Record variance ratio: .301</td>
<td>1.835</td>
</tr>
<tr>
<td></td>
<td>• 1982-1981 Change:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Annual Earnings:</td>
<td>Hourly Earnings</td>
</tr>
<tr>
<td></td>
<td>Mean (interview): $426</td>
<td>$1.61</td>
</tr>
<tr>
<td></td>
<td>Mean (record): 179</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>Mean (difference): 247</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>Mean (absolute difference): 2,477</td>
<td>2.82</td>
</tr>
<tr>
<td></td>
<td>Error/Record variance ratio: .501</td>
<td>2.920</td>
</tr>
<tr>
<td>Source</td>
<td>Description</td>
<td>Social Security Administration Records</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
</tbody>
</table>
ln interview earnings - ln record earnings: mean=.004, variance=.114  
• Women:  
annual earnings (household report): mean = $7,906  
ln interview earnings - ln record earnings: mean=-.017, variance=.051  
(above based on sample with record earnings below SS maximum)  
• 1977 ln (earnings):  
|                                |                                                                             | Women:                                    |                                                                            |
|                                |                                                                             | variance (interview)                      | .437 .666  
variance (record)              | .529 .625  
variance (difference)          | .116 .051  
r(interview, record)          | .884 .961  
r(error, record)           | -.420 -.028  
b(record on interview)      | .974 .962  
• 1977-1976 change in ln (earnings): | Men | Women |  
variance (interview)          | .186 .437  
variance (record)              | .223 .394  
variance (error)              | .121 .089  
r(interview, record)        | .707 .894  
r(error, record)          | -.481 -.123  
b(record on interview)      | .775 .848  
• Mismeasurement of earnings leads to little bias when CPS earnings on left-hand side of regression (errors weakly related to regressors)  
• Positive autocorrelation between errors in CPS reported earnings; coefficient of .40 for men and .10 for women |
| Coder (1992)                   | Sum of husband’s and wife’s annual wage and salary income (SIPP)            | Matched tax return information (from joint returns) | • For sample with unimputed SIPP earnings:  
mean (interview)              | $40,030  
mean (record)                | 42,060  
variance (interview)         | $787x10^6  
variance (record)            | $1446x10^6  
variance (error)             | $454x10^6  
r(interview, record)       | .834  
r(error, record)           | -.687  
b(record on interview)     | 1.130  
• Mean error larger (in absolute value) when SIPP earnings are partially of completely imputed.
Rogers, Brown, and Duncan\(^\text{1,3}\) (1993)

- Earnings and hourly wage (each measured three ways):
  - Annual
  - Most recent pay period
  - Usual

(PSID Validation Study)

<table>
<thead>
<tr>
<th>Employer’s Records</th>
<th>• Correlation between interview and record</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ln Earnings</td>
</tr>
<tr>
<td>Annual</td>
<td>.784</td>
</tr>
<tr>
<td>Most recent</td>
<td>.675</td>
</tr>
<tr>
<td>Usual</td>
<td>.456</td>
</tr>
</tbody>
</table>

- Correlation between error and record
  
<table>
<thead>
<tr>
<th>ln Earnings</th>
<th>ln Hourly wage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual</td>
<td>-.216</td>
</tr>
<tr>
<td>Most Recent</td>
<td>-.301</td>
</tr>
<tr>
<td>Usual</td>
<td>-.436</td>
</tr>
</tbody>
</table>

- Wage rates calculated from reported earnings and hours; variance of the errors can be decomposed into three parts: variance in errors in reported earnings, variance in errors in reported hours, and the covariance of those two reports. For annual wage rates, contribution due to error in annual earnings and annual hours are about equal (.93 and .80); errors are positively correlated \(r=.43\); covariance is negative \(-.74\); for wage rate based on most recent pay period, errors in reported earnings are about twice as important as errors in reported hours \(1.36 \text{ and } .62\); covariance again is negative \(-.98\). Based on usual pay the estimates are \(1.26, .32, \text{ and } -.58\).

- Mean error significantly different from zero (albeit small); significantly related to true values (negative), impact the magnitude of regression coefficients when wages are on the left hand side of the equation; and are correlated (weak, positive) across time
<table>
<thead>
<tr>
<th>Bound, Brown, Duncan, and Rodgers(^1,3) (1994)</th>
<th>Annual Earnings and wage per hour (PSID Validation Study)</th>
<th>Wage per hour for hourly workers only</th>
<th>Employer’s Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Mean differences between record and interview values of ln (earnings) small and statistically insignificant for both annual earnings and hourly wage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• 1986</td>
<td>ln earnings</td>
<td>ln wage/hour</td>
<td></td>
</tr>
<tr>
<td>variance (interview)</td>
<td>.0488</td>
<td>.0204</td>
<td></td>
</tr>
<tr>
<td>variance (record)</td>
<td>.0416</td>
<td>.0085</td>
<td></td>
</tr>
<tr>
<td>variance (difference)</td>
<td>.0108</td>
<td>.0121</td>
<td></td>
</tr>
<tr>
<td>r(interview, record)</td>
<td>.8862</td>
<td>.6350</td>
<td></td>
</tr>
<tr>
<td>r(error, record)</td>
<td>-.0785</td>
<td>-.0109</td>
<td></td>
</tr>
<tr>
<td>b(record on interview)</td>
<td>.8180</td>
<td>.4085</td>
<td></td>
</tr>
<tr>
<td>• 1986-1982 change in:</td>
<td>ln earnings</td>
<td>ln wage/hour</td>
<td></td>
</tr>
<tr>
<td>variance (interview)</td>
<td>.0365</td>
<td>.0433</td>
<td></td>
</tr>
<tr>
<td>variance (record)</td>
<td>.0357</td>
<td>.0112</td>
<td></td>
</tr>
<tr>
<td>variance (difference)</td>
<td>.0164</td>
<td>.0376</td>
<td></td>
</tr>
<tr>
<td>r(interview, record)</td>
<td>.7738</td>
<td>.3786</td>
<td></td>
</tr>
<tr>
<td>r(error, record)</td>
<td>-.3219</td>
<td>-.1404</td>
<td></td>
</tr>
<tr>
<td>b(record on interview)</td>
<td>.7657</td>
<td>.1930</td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>Type of Data</td>
<td>Source of Data</td>
<td>Notes</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Branden and Pergamit (1994)</td>
<td>Starting wages</td>
<td>Respondent’s reports in year $t$ compared to year $t+1$</td>
<td>- 42% of respondents report the same starting wage at the two points in time. Consistency related to the time unit used for reporting, with the highest rate of consistency among those reporting starting wage as an hourly or daily rate (47% and 52%, respectively); lowest among those reporting a biweekly wage rate (13% consistent)</td>
</tr>
<tr>
<td>Barron, Berger, and Black (1997) Table 5.1</td>
<td>Starting wages (Upjohn Institute Survey)</td>
<td>Employers’ records</td>
<td>- Mean (interview) = $8.84; Mean (record) = $9.95 difference in means not significant ($t=1.31$)</td>
</tr>
<tr>
<td>Bollinger (1998)</td>
<td>Annual earnings, previous calendar year (CPS)</td>
<td>Social Security Administration records</td>
<td>- Measurement error more severe (larger mean error for men, larger error variance for women) in single cross-section than in two-year panel</td>
</tr>
<tr>
<td>Angrist and Kruger$^3$ (1999)</td>
<td>Hourly wage</td>
<td>Employers’ records</td>
<td>- ln (employer reported wage) - ln (employee reported wage): mean=.017</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- variance (interview): .355</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>variance (record): .430</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>variance (difference): .238</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>r(interview, record): .650</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>r(error, record) : -.489</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>b(record on interview): .770</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Recoding lowest (highest) one percent of employee-reported wages to the 1st (99th) percentile value increased b(record on interview) to .88</td>
</tr>
</tbody>
</table>

**Notes:**
1. Sample limited to a single employer.
2. Small sample (n=300) in single geographic area; assessment of the accuracy of reports of annual earnings based on 173 persons for whom household reports could be linked to state employment records.
3. Respondent not asked to report hourly wage rate directly. Wages (Mellow and Sider; Angrist and Krueger ) or hourly earnings (Rogers, Brown, and Duncan and Bound, Brown, Duncan, and Rodgers) calculated from earnings divided by hours worked; error in reported hours therefore contributes to error in hourly wage rate.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Variables of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
</table>
| Bancroft (1940)   | Public relief                                              | Administrative records             | • Pr(interview=currently receiving|record=currently receiving) = .92  
• Pr(interview=never received|record=never received) = .84  
(this biased down because records miss receipt >2.5 years prior to interview) |
| David (1962)      | Public assistance                                           | Administrative records             | • 7 percent of recipients (according to records) reported not receiving public assistance.  
• Mean (interview) = $2,334  
Mean (record) = $2,758  
Mean (error) = -$434, or 18 percent of record  
rt(interview, record) = .30  
• Errors unrelated to respondent race, sex, age (but N=46) |
| Haber (1966)      | Social Security Income of “beneficiary unit” (couple or non-married individual) | Administrative records             | • Mean (interview) = $991  
Mean (record) = $1,052  
Mean (error) =-$51, o 5 percent of record (s.e. = $5 )  
• Under-reporting greatest for youngest (62-64) respondents, the institutionalized, and those with high (true) benefit levels. |
| Livingston (1969) | Several types of “public assistance” reported in special census | Administrative records in Dane County, WI | • 22% of known recipients failed to report receipt  
• Over 50% of those reporting receipt in census could not be matched to an administrative record  
• Among those who report assistance and receipt is corroborated in records: median reported in interview is 73 percent of median in records. For old age assistance and AFDC separately, corresponding ratios are 80 and 70 percent, respectively. |
<p>| Hu (1971)         | Cash and medical assistance                                | Administrative records             | • 27% of recipients failed to report assistance receipt |</p>
<table>
<thead>
<tr>
<th>Study</th>
<th>Program</th>
<th>Methodology</th>
<th>Findings</th>
</tr>
</thead>
</table>
| Oberheu and Ono (1975)        | AFDC participation “last month,” annual AFDC, and Food Stamp receipt “last month” among low-income households with children | Administrative records | - Reporting of AFDC recipient for last month:  
  \[ Pr(\text{interview}=\text{yes}|\text{record}=\text{yes})=0.68 \]  
  \[ Pr(\text{interview}=\text{no}|\text{record}=\text{no})=0.77 \]  
  Net under-reporting of receipt = 2%  
  Among those who correctly report having received benefits, mean under-report = $89/month  
- Similar findings for AFDC receipt for last year  
- Reporting of Food Stamp participation for last month:  
  \[ Pr(\text{interview}=\text{yes}|\text{record}=\text{yes})=0.70 \]  
  \[ Pr(\text{interview}=\text{no}|\text{record}=\text{no})=0.85 \]  
  Net over-reporting of receipt = 6% |
| Vaughan and Yuskavage (1976)  | Social Security Income (CPS)         | Administrative records | - Among cases where both record and interview showed a positive amount received:  
  54 percent of interviews exceed record amount  
  39 percent of interviews are less than record amount  
  7 percent of cases agree within $10  
  Mean error = $68 (5 percent of average benefit)  
  Mean absolute error = $225 (15 pct. of average benefit) |
| Halsey (1978)                 | AFDC, Unemployment Insurance         | Administrative records | - Among cases where record and/or interview showed a positive amount received:  
  Mean amount of AFCD under-reported by 25-30%  
  Mean amount of Unemployment Insurance under-reported by 50%  
- \( r(\text{interview}, \text{record}) \) are in .40-.60 range |
| Hoaglin (1978)                | Social Security Income, Supplementary Security Income, “welfare” | Administrative records | - Median response error is $0 for “welfare” (combines AFDC, general assistance and other programs), SSI, and unemployment insurance reports; slightly negative for reports of monthly Social Security amounts |
| Vaughan (1978)                | Social Security Income               | Administrative records | - \( Pr(\text{interview}=\text{yes}|\text{record}=\text{yes})=0.87 \)  
- Most of the remaining 13 percent appear to mis-report SSI income as Social Security rather than failing to report any transfer income at all. |
<table>
<thead>
<tr>
<th>Study</th>
<th>Income</th>
<th>Data Source</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klein and Vaughan (1980)</td>
<td>AFDC receipt</td>
<td>Administrative records</td>
<td>Pr(interview=yes</td>
</tr>
<tr>
<td>Goodreau, Oberheu, and Vaughan (1984) Tables 1,3,4</td>
<td>AFDC receipt</td>
<td>Administrative records in California, North Carolina, Pennsylvania, and Wisconsin</td>
<td>91% report receiving cash assistance, but only 78% correctly identify the payment of AFDC per se.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Amount last month (those receiving any cash assistance)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean (record): $286</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean (household report): $276</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Simple difference: $10 (3.5%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reporting error negatively related to record amount</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Among those receiving AFDC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>74 percent reported as AFDC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13 percent reported as other transfers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 percent under-reported by those reporting receipt</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9 percent received by those reporting no cash transfers</td>
</tr>
<tr>
<td>Marquis and Moore (1990)</td>
<td>AFDC, Food stamps, Unemployment Insurance Benefits, Workers Compensation, Social Security (OASDI), Supplemental Security Income and Veteran’s Pensions and Compensation as reported in SIPP</td>
<td>Administrative records in Florida, New York, Pennsylvania, and Wisconsin</td>
<td>Under-reporting by known recipients (Pr(interview=no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Under-reporting Known recipients Relative Net</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Program</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AFDC:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Unemployment insurance:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Food Stamps:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Supp. Security Income:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Veterans’ Benefits:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Social Security:</td>
</tr>
<tr>
<td>Grondin and Michaud (1994)</td>
<td>Unemployment benefits</td>
<td>Canadian tax returns</td>
<td>4-6% of reports on recipiency are in error</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Net under-reporting rate 3-4% for recipiency</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pr(interview=no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pr(interview=yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Discrepancy between record and interview report exceeds 5% for approximately a third of those with non-zero amounts for both</td>
</tr>
<tr>
<td>Authors</td>
<td>Programs</td>
<td>Data Source</td>
<td>Notes</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------------------------------------</td>
<td>------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Dibbs, Hale, Loverock, and Michaud (1995) ¹ | Unemployment benefits | Tax returns | • Mean (record) = $5600  
Mean (interview) = $5300  
Mean (error) = $300, or 5% |
| Moore, Marquis, and Bogen (1996) ¹ | AFDC, Food Stamps, Unemployment Insurance, and Supplemental Security Income as reported in SIPP; two experimental questionnaires | Administrative records in Wisconsin | • Under-reporting by known recipients  
(Pr(interview=no|record=yes)) and over-reporting by non-recipients (Pr(interview=yes)|Pr(record=no)), by program:  
|                                   |                                               |                                    | AFDC: Under-reporting 0.10–0.12 Non-recipients 0.03–0.04 |  
|                                   |                                               |                                    | Unemployment insurance 0.41–0.44 Non-recipients 0.01     |  
|                                   |                                               |                                    | Food Stamps 0.12–0.17 Non-recipients 0.02–0.03          |  
|                                   |                                               |                                    | Supp. Security Income 0.08–0.13 Non-recipients 0.03      |  
|                                   |                                               |                                    | • 70% to 80% report AFDC, Food Stamps, and SSI within 5% of record; 20% to 30% accurately report unemployment insurance |
| Yen and Nelson (1996) ¹          | AFDC                                          | Administrative records in Washington | • 93% of the 49,000 eligible person-months reported correctly.  
• Survey-based estimates of monthly participation exceeded record-based estimates of participation by approximately 1 percentage point |
| Bollinger and David (1997)       | Food Stamp participation (SIPP) individual records aggregated to the household level | Administrative records | • Pr(interview=yes|record=yes) = 0.88  
Pr (interview=yes|record=no)=0.003 |

Notes: ¹ Unpublished paper, reported in Moore, Stinson, and Welniak (1977)
<table>
<thead>
<tr>
<th>Reference</th>
<th>Variables of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
</table>
| Kish and Lansing (1954)         | House value (1950 Survey of Consumer Finances) | Appraisals (conducted specifically for validation) | • Mean value reported by owner = $9,560  
• Mean value reported by appraiser = $9,210  
• Difference = $350 (s.e. = 170)  
• Owner’s and appraiser’s estimates differed by at least 30 percent in 24 percent of the cases  
• Largest discrepancies ultimately traced to coding errors |
| Lansing, Ginsburg, and Braaten (1961) | Savings account ownership ³            | Financial institution records                                                      | • Ownership of savings account:  
  Pr(interview=yes|record=yes)=.75  
• Conflicting evidence on extent of under-reporting by those who report having accounts:  
  In one sample, mean(record)=$3310,  
  mean(error)=-500 or -14%; mean abs error=$1571  
  In second sample, mean error=+2% |
| Maynes (1965)                   | Savings account ownership ⁴            | Financial institution records                                                      | • Ownership of savings account:  
  Pr(interview=yes|record=yes)=.95  
• Of those who report an account and the amount in it:  
  Mean(record)=1827, mean error=-83 or 5 percent  
  Mean error=-1% for those who consult records, -10% for those who do not  
  Error negatively related to record amount  
  Savings over 9 months also under-reported, with errors negative related to record savings |
| Ferber (1966)                   | Savings account ownership              | Financial institution records                                                      | • Ownership of savings accounts: in three samples,  
  Pr(interview=yes|record=yes) = .81, .78, and .65.  
• Among those report an account and the amount in it,  
  mean errors = -20%, 0.3%, and 8% of record means |
<table>
<thead>
<tr>
<th>Study</th>
<th>Variable Description</th>
<th>Data Source</th>
<th>Observations</th>
</tr>
</thead>
</table>
| Ferber, Forsythe, Guthrie, and Maynes (1969a) | Savings account ownership                                            | Financial institution records | • 46 percent of known accounts not reported; 32 percent of families known to have at least one account reported not having any  
• Owners of larger accounts less likely to participate in survey, but more likely to report accounts if they participate  
• For accounts reported and matched to record data, mean error negligible (record=$3040, interview=$3042) but interview reports differ from record by ± 50 % for 28% of respondents.  
• Reporting error negatively related to record value |
| Ferber, Forsythe, Guthrie, and Maynes (1969b) | Stock ownership (in shares of particular cooperating firms)         | Financial institution records | • Ownership of stock (in a particular firm):  
  \[ \text{Pr}(\text{interview}=\text{yes}|\text{record}=\text{yes}) = .70 \]  
• Those who own more shares less likely to participate in survey; among participants, reporting owning (any of) the stock not monotonically related to shares actually owned  
• For stocks reported and matched to record data, mean error negligible (record=63.8 shares, interview=63.9); interview reports differ from record by ± 50 % for 13% of respondents.  
• Those with largest holdings tend to under-report, but otherwise relationship between error and record values is irregular |
| Rodgers and Herzog (1987)    | Assessed value of house (Study of Michigan Generations)             | Property tax records     | • \( r(\text{error, record}) = .242 \) (s.e. = .138)  
• error uncorrelated with age, education, marital status, or race; correlation with income = .224 (s.e. = .123) |
<table>
<thead>
<tr>
<th>Grondin and Michaud (1994)</th>
<th>Asset ownership; interest and dividend income</th>
<th>Canadian tax returns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 22% and 11% of survey respondents (in two studies, paper-and-pencil and computer assisted, resp.) misreport whether income and dividend income was received
- Net under-reporting rate of 19 and 6 percentage points, respectively
- \( \Pr(\text{interview=yes} | \text{record=yes}) = 0.58 \) and 0.78, resp.
- \( \Pr(\text{interview=no} | \text{record=no}) = 0.98 \) and 0.96, resp.
- Of those with positive amount of income and dividend income in both interview and record, approx. 70 percent agree within 5 percent on the amount.

Notes:
1. Most validation studies involving assets have focused on the respondent’s ability to report the ownership of the asset and the amount of the asset rather than the income generated from the asset.
2. Reported in Moore, Stinson, and Welniak (1997)
3. Two samples, one limited to accounts \( \geq \$500 \), one limited to accounts \( \geq \$1000 \).
4. Sampling rate higher for large accounts; accounts < 10 guilders excluded. Reported statistics unweighted. Data from Netherlands.
## Table 4. Assessment of Measurement Error: Hours Worked

<table>
<thead>
<tr>
<th>Reference</th>
<th>Variables of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
</table>
| Carstensen and Woltman (1979) | Usual hours worked per week (CPS special supplement)   | Employers’ records                       | • Mean (household): 38.43  
  Mean (employer): 37.10  
  Mean (difference): 1.33 (s.e.=.10)  
  • Compared to company records, estimates of the mean “usual hours worked” significantly over-reported by household respondents |
| Stafford and Duncan (1980)    | Average work week (Time Use Study) and Hours worked last week (CPS) | Time-diary reports of various work activities | • Mean (average hours/week): 41.8  
  Mean (time-diary reports): 36.8 [37.5 eliminating 11 outliers]  
  • Change in mean work hours per week between 1965 and 1975 larger in time-use data than CPS reports of hours worked last week:  
    | CPS Hours                   | Time Diary                      | Married men   | -1.3 | -3.4 |
|                               |                          |                           | Married women | -0.5 | -7.8 |
|                               |                          |                           | [Time-use means exclude those working < 10 hours/week]                                                                                   |
| Mellow and Sider (1983)       | Hours worked (CPS)                                                 | Employers’ records                       | • In (worker report) - ln (employer report):  
  mean=.039, variance =.064  
  • Regression model predicting difference indicates that professional and managerial workers report more hours than their employers; over-reports also associated with educated and nonwhite employees, while females tend to under-report hours |
<table>
<thead>
<tr>
<th>Study</th>
<th>Source of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duncan and Hill(^2) (1985)</td>
<td>Employer’s records</td>
</tr>
</tbody>
</table>
| Annual hours worked, year \(t\) and \(t-1\) hourly workers only (PSID Validation Study) | • Mean 1982 annual hours (interview): 1,693 Mean 1982 annual hours (record): 1,603 Mean difference: 90 (\(p<.01\))  
  • Mean 1981 annual hours (interview): 1,880 Mean 1981 annual hours (record) 1,771 Mean difference: 115 (\(p<.01\))  
  • Mean 1982-1981 simple change (interview): -185 Mean 1982-1981 simple change (record) -167 Mean difference: -17  
  • Mean absolute \(|1982-1981|\) change (interview): 357 Mean absolute \(|1982-1981|\) change (record) 286 Mean difference: 70  
  • Significant over-reporting of hours worked for both years with an average absolute error of \(\approx 10\%\)  
  • Error-to-record variance ratio: In 1982 annual hours: .366 |
| Hamermesh (1990)                           | Time diary data                                                                |
| Hours worked last week (Time use studies)   | • Average hours worked from CPS-like question on hours worked last week exceed time-diary estimates by 1.5 hours in 1975 and 3.6 hours in 1981 |
| Rogers, Brown, and Duncan\(^2\) (1993)     | Employer’s records                                                            |
| Hours worked:                                | • Correlation between self-report and company records: .66, .66, and .61 for annual, most recent, and usual pay periods, respectively, after deleting outliers. Relative ranking sensitive to this decision and so unclear, overall  
  • Correlation between error and company records: -.31, -.36, and -.37 for annual, most recent, and usual pay periods, resp.  
  • Weak positive correlation (.061, not significant) of errors over time (1986 and 1982). |
| Robinson and Bostrom (1994)                 | Time diary data                                                                |
| Hours worked last week (Time Use Study)     | • Hours worked last week exceed time-diary estimates of hours worked per week by 1 hour in 1965, by 4 hours in 1975, and by 7 hours in 1985 |
| Bound, Brown, and Duncan, and Rodgers\(^2\) (1994) | Annual hours worked (PSID Validation Study) | Employer’s reports | • Mean interview reports of ln hours insignificantly higher than record values.  
  • ln annual hours  
    - variance (interview)  .0180  
    - variance (record)  .0174  
    - variance (error)  .0104  
    - r(interview, record) = .7033  
    - r(error, record) = -.3701  
  • ln annual hours 1986-1982 change  
    - variance (interview)  .0620  
    - variance (record)  .0529  
    - variance (error)  .0237  
    - r(interview, record) = .7962  
    - r(error, record) = -.2061  
  • b(record on interview) = .6828  
  • Correlation between 1986 and 1982 errors positive but very small (.064) |
|---|---|---|---|
| Barron, Berger, and Black (1997) Table 5.1 | Hours worked per week (Upjohn Institute Survey) | Employers’ records | • Mean (interview) = 38.5, Mean (record) = 37.0; difference in means statistically significant (t=3.95) for ln(hours), difference in means = .031  
  • r(interview, record) = .769  
  for ln (hours), r(interview, record) = .61 |
| Angrist and Krueger\(^1\) (1999) Table 10 | Hours Worked (CPS) | Employers’ Records | • ln(employee-reported hours) - ln (employer reported hours): mean = .043  
  • variance (interview) = .195  
  • variance (record) = .182  
  • variance (difference) = .038  
  • r(interview, record) = .780  
  • r(error, record) = -.149  
  • b(record on interview) = .870  
  • Recoding lowest (highest) one percent of employee-reported wages to the 1st (99th) percentile value increased b(record on interview) to .91 |

Notes:
1. It is unclear from the empirical findings as to the time reference used for the reporting of hours worked.
2. Sample limited to a single employer. Hours worked calculated from the respondent’s account of each week of the year (working, sick or annual leave, etc.).
<table>
<thead>
<tr>
<th>Name (Year of Study)</th>
<th>Variable of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
</table>
| Morganstern and Bartlett¹    | Annual number of person years of unemployment, 1964-71 (CPS; WES from March Supplement) | None; comparison of estimate based on annual report (WES) vs. recall for previous week (CPS) | • Average percentage discrepancies between CPS and WES (standard errors in parentheses):  
  White males: 3.25 (6.23)  
  Black males: 4.32 (6.45)  
  White females: 23.95 (6.45)  
  Black females: 21.56 (7.43)  
  • WES (annual recall) tends to underestimate unemployment, with the greatest discrepancy for women and youths.  
  • In high unemployment years, tendency for WES to overstate the amount of unemployment by 4 percentage points; independent of the overall unemployment rate, tendency for the WES to understate the CPS rate for women by an average of 19 percentage points  
  • Some indication of overreporting of unemployment among white males, 25 and older and white females, 45 and older |
| Horvath¹                    | Average estimate of weekly unemployment (CPS; WES from March Supplement)               | None; comparison of annual unemployment data with average computed from monthly CPS | • Underestimate of unemployment based on annual WES measure ranged from about 9 to 25 percent; underestimate smallest for periods of increasing unemployment  
  • Unemployment during the first six months of the year less likely to be reported in WES than unemployment in second six months of the year |
| Bowers and Horvath\(^{1,2}\) (1984) | Duration of unemployment spell (CPS) | None; reporting of continuous unemployment spell one month later compared to report at time t | • Approximately 25% of respondents consistent in their report of unemployment duration  
  • Percent consistent in reports of unemployment duration as a function of spell duration:  
    < 5 weeks: 32.8% - 40.0%  
    5-10 weeks: 23.3%-28.3%  
    11-14 weeks: 16.0%-21.2%  
    15-26 weeks: 18.6%-29.8%  
    27-51 weeks: 20.0%-23.1%  
    52 weeks: 0.0%-10.7%  
    53-99 weeks: 8.7%-18.6%  
  • The longer the spell reported at time t, the smaller the increase in reported duration one month later |
|---|---|---|---|
| Poterba and Summers (1984) Table 1 | Employment status (CPS, May 1976) | CPS Reinterview Survey (after reconciliation with initial reports) | • Pr(CPS report|Re-interview Status), May 1976  
  Reinterview after Initial CPS Interview  
  Reconciliation Employed | Employed | .860 | .104 | Not in LF | .005 | .003 | .992 |
| Poterba and Summers (1984) Table 2 | Unemployment duration (CPS, June 1996) | Consistency with May 1976 interview | • Only 32 percent of June duration reports were “consistent” with (i.e., 3-5 weeks greater than) May report. Inconsistent reports about evenly divided between those with June-May difference greater than 5 weeks and those less than 3 weeks  
  • Difference between reports tended to be too large for those who reported being unemployed <20 weeks in May, and too small for those unemployed >20 weeks in May. |
<table>
<thead>
<tr>
<th>Study Authors (Year)</th>
<th>Description</th>
<th>Data Source</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abowd and Zellner (1985) Tables 6,7</td>
<td>Employment status (CPS, 1987-1982)</td>
<td>CPS Reinterview Survey (after reconciliation with initial reports)</td>
<td>• Pr(CPS report</td>
</tr>
<tr>
<td>Akerlof and Yellen (1985)</td>
<td>Average estimate of weekly unemployment (CPS; WES from March Supplement)</td>
<td>None; comparison of annual unemployment data with average computed from monthly CPS</td>
<td>• Previous-year unemployment reports from the WES average 90 percent of those obtained from the monthly CPS for the same calendar year. • WES-CPS difference has grown more negative over time • Under-reporting on WES most severe for those under 25 and for women 25-54.</td>
</tr>
<tr>
<td>Duncan and Hill (1985)</td>
<td>Unemployment hours, year t (1982) and t-1 (PSID Validation Study)</td>
<td>Employer’s records</td>
<td>• Annual unemployment hours 1982 1981 1982-1981 Mean (interview) 169 39 131 Mean (record) 189 63 126 Mean (error) -11 -16 5 Mean (absolute error) 52 45 77 • Mean of the difference between interview and record data not significantly different from zero in either year. • Average absolute difference was large relative to average amount of unemployment in each year-- about one-third the mean unemployment for 1982 and two-thirds for year 1981 (one- and two-year recall, resp.)</td>
</tr>
<tr>
<td>Study</td>
<td>Sample Period</td>
<td>Data Source</td>
<td>Methodology</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Poterba and Summers (1986) Tables II and V</td>
<td>Employment status (CPS, 1977-1982)</td>
<td>CPS Reinterview Survey (after reconciliation with initial reports)</td>
<td>Pr(CPS report</td>
</tr>
<tr>
<td>Chua and Fuller (1987) Tables 1-3 and 6-7</td>
<td>Employment status (CPS, 1976-1978)</td>
<td>CPS Reinterview Survey (not reconciled with initial reports)</td>
<td>Pr(CPS report</td>
</tr>
<tr>
<td>Mathiowetz and Duncan (1988); Mathiowetz (1986)</td>
<td>Unemployment spells (PSID Validation Study)</td>
<td>Employer’s records</td>
<td>Overall, 66% of spells unreported. • Accurate reporting of spells associated with the amount of unemployment in a given month and the temporal complexity of the spell.</td>
</tr>
<tr>
<td>Torelli and Trivellato (1989)</td>
<td>Unemployment duration (Youth 14-29 in Italy’s quarterly labor force survey)</td>
<td>None; Consistency in reporting duration of unemployment spell between quarterly surveys and actual elapsed duration</td>
<td>Pr(CPS report</td>
</tr>
</tbody>
</table>
| Levine\(^1\) (1993) | Unemployment rate (CPS and WES from March Supplement) | None; comparison of contemporaneous rate and one year retrospective recall. | • Unemployment rate under-reported by 7% to 24% when comparing retrospective rate to contemporaneous rate.  
• 35% to 60% of persons failed to report unemployment one year after the event. Misreporting rate related to length of unemployment spell.  
• Error correlated with economic cycle—less under-reporting during recessionary periods, greater under-reporting during expansionary periods. |

**Notes:**
1. Estimates based on the monthly CPS unemployment rate used as the “gold” standard for comparison. Comparison involves the use of different questionnaires and different questions to obtain measure of unemployment.
2. The authors note that the findings may, in part, reflect rounding by those with very long (> 6 months) unemployment spells.
3. Sample limited to a single employer. Hours unemployed calculated from respondents accounting for each week of the year (working, sick, annual leave, etc.).
4. Sample limited to a single employer. Spell-level information obtained by asking the respondent to report the month in which he or she was unemployed for at least part of the month. Company plagued by sporadic unemployment during years of interest.
5. Sample limited to those ages 14-29 living in the Lombardy region of Italy.
<table>
<thead>
<tr>
<th>Name (Year of Study)</th>
<th>Variable of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
</table>
| Weiss, Dawis, England and Lofquist (1961) | Occupation classification: 3-digit level | Employers’ responses to independent questionnaire | • Self-reports agreed with company reports for 67% of the jobs reported during the past five years  
• Agreement rate for current occupation: 70%; for occupations more than 4 years ago, agreement rate drops to 60%  
• Agreement rate higher for older workers, but not related to age, education, or broad occupation |
| Dreher (1977) | Tenure (nine 5-year intervals) | Employer’s records | • $r_{(interview, record)}=.97$ |
| Mellow and Sider (1983) | Industry classification: 1 and 3 digit level (CPS and Employment Opportunity Pilot Project (EOPP)) | Employers’ reports | • CPS (N=4523)  
• One digit agreement rate: 92.3%  
• Three digit agreement rate: 84.1%  
• Agreement rates only slightly higher for self- than for proxy reports  
• EOPP (N=3327)  
• One digit agreement rate: 87.5%  
• Three digit agreement rate: 71.1% |
| Mellow and Sider (1983) | Occupation classification: 1 and 3 digit level (CPS) | Employers’ records | • One digit agreement rate: 81.0%  
• Three digit agreement rate: 57.6%  
• Agreement rates only slightly higher for self- than for proxy reports |
<table>
<thead>
<tr>
<th>Study</th>
<th>Classification</th>
<th>Source</th>
<th>Agreement Rates</th>
</tr>
</thead>
</table>
| Mathiowetz\(^2\) (1992)                    | Occupation classification: 1 and 3 digit level; direct comparison (PSID Validation Study) | Employer’s records | • One digit agreement rate: 75.7%  
• Three digit agreement rate: 51.8%  
• Direct comparison (coder looks at worker and employer description at same time) agreement rate: 87.3% |
| Brown and Medoff\(^3\) (1996)              | Industry classification: 14 industry groups | Dun and Bradstreet | • Workers’ reports and D&B SIC code agreed 79% of the time                      |

Notes:
1. Sample limited to the first 325 persons of the Work Adjustment Project (Minneapolis-St. Paul metropolitan area) for whom both interview and employer work histories were obtained.
2. Sample limited to a single employer.
3. Sample limited to those respondents for whom respondents’ reports of employer could be matched to D&B files. Successfully matched employers tended to be larger and in business longer than employers in the overall sample.
<table>
<thead>
<tr>
<th>Name (Year of Study)</th>
<th>Variable of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
</table>
| Weiss, Dawis, England and Lofquist<sup>1</sup> (1961) | Starting date | Employers’ responses to independent questionnaire | • Reported starting date agreed with company records (within one month) for 71% of the jobs in past 5 years  
• Validity significantly declines as a function of the length of time between start date and date of interview |
| Mellow and Sider (1983) | Coverage under union contract (CPS and Employment Opportunity Pilot Project (EOPP))<sup>2</sup> | Employers’ reports | • CPS (N=4523) sample proportions  
Employer Report  
Worker Report Covered .235 Not Covered .041  
Worker Report Covered .362 Not Covered .098 |
| Duncan and Hill<sup>3</sup> (1985) | Coverage under union contract (PSID Validation Study) | Employer’s records | • Less than 1% disagreement, all in the direction of workers claiming coverage when employer did not |
| Duncan and Hill<sup>3</sup> (1985) | Union membership (PSID Validation Study) | Employer’s records | • Less than 1% disagreement |
| Duncan and Hill<sup>3</sup> (1985) | Health insurance, dental benefits, life insurance (PSID Validation Study) | Employer’s records | • Health Insurance: < 1% disagreement  
• Dental Benefits: 5% disagreement (workers claim no benefits when employer indicates benefit)  
• Life Insurance: 10% disagreement (workers claim no benefits when employer indicates benefit) |
<table>
<thead>
<tr>
<th>Source</th>
<th>Variable Description</th>
<th>Source Type</th>
<th>Agreement/Correlation Details</th>
</tr>
</thead>
</table>
| Duncan and Hill (1985)         | Paid time off for vacation days, sick days               | Employer’s records   | • Vacation: <1% disagreement  
• Sick Leave: 9% disagreement (3% claiming benefit when company record indicates no benefit; 6% employer claims benefit and employee reports no benefit) |
<p>| Duncan and Mathiowetz (1985)   | Tenure                                                    | Employer’s records   | • 90% of respondents report hire date within one year of date recorded by company                                                                                   |
| Bound, Brown, Duncan, and      | Tenure                                                    | Employer’s records   | • Correlation between worker and employer reports of tenure with employer = .99 in both 1982 and 1986.                                                                 |
| Rodgers (1994)                 |                                                           |                      |                                                                                                                                                                |
| Brown and Medoff (1996)        | Establishment and company size                           | Dun &amp; Bradstreet     | • Correlation between worker report and D&amp;B value: .82 (ln establishment size) and .86 (ln company size)                                                        |
| Brown and Medoff (1996)        | Age of firm                                              | Dun &amp; Bradstreet     | • Correlation between worker report and D&amp;B value: .56 (years firm in business) and .50 (ln years firm in business)                                                 |
| Barron, Berger, and Black      | Union coverage                                           | Employers’ reports   | • Correlation between worker and employer report = .689                                                                                                        |
| (1997) Table 5.1               |                                                           |                      |                                                                                                                                                                |
| Barron, Berger, and Black      | Eligibility for health and life insurance, and retirement| Employers’ reports   | • Correlation between worker report and employer record:                                                                                                         |
| (1997) Table 5.1               |   plan                                                   |                      | When first hired   | After two years with firm                                                                                                                                           |
|                                |                                                           |                      | Health insurance   | .590          | .469             |
|                                |                                                           |                      | Life insurance     | .516          | .508             |
|                                |                                                           |                      | Retirement plan    | .312          | .327             |
| Barron, Berger, and Black      | Eligibility for paid vacation and sick pay               | Employers’ reports   | • Correlation between worker report and employer record:                                                                                                         |
| (1997) Table 5.1               |                                                           |                      | When first hired   | After two years with firm                                                                                                                                           |
|                                |                                                           |                      | Paid vacation      | .247          | .490             |
|                                |                                                           |                      | Sick pay           | .294          | .428             |
| Barron, Berger, and Black      | Hours of training                                        | Employers’ reports   | • Correlation between workers’ reports and employers for various types of training:                                                                          |
| (1997) Table 5.1               |                                                           |                      | On-site formal training | .398          |
|                                |                                                           |                      | Off-site formal training | .457          |
|                                |                                                           |                      | Informal, managerial | .176          |
|                                |                                                           |                      | Informal, coworker  | .379          |
|                                |                                                           |                      | Total training      | .475          |</p>
<table>
<thead>
<tr>
<th>Covered by employer-provided health insurance (March CPS)</th>
<th>Covered by employer-provided health insurance (April/May CPS Supplements)</th>
<th>Covered by employer-provided health insurance (March 1988 CPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• March 1988 vs May 1988 CPS (N=10,070)</td>
<td>• March 1993 vs April 1993 CPS (N=11,603)</td>
<td></td>
</tr>
<tr>
<td>May 1988 CPS</td>
<td>Not Covered</td>
<td>Covered</td>
</tr>
<tr>
<td>Not Covered</td>
<td>.094</td>
<td>.034</td>
</tr>
<tr>
<td>Covered</td>
<td>.073</td>
<td>.799</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eligible for health insurance (Upjohn Institute Survey)</th>
<th>Employers’ reports</th>
<th>• Employee vs Employer Reports (N=257)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worker Report</td>
<td>May 1988 CPS</td>
<td>Not Covered</td>
</tr>
<tr>
<td>Ineligible</td>
<td>.459</td>
<td>.156</td>
</tr>
<tr>
<td>Eligible</td>
<td>.054</td>
<td>.331</td>
</tr>
</tbody>
</table>

**Notes:**
1. Sample limited to the first 325 persons of the Work Adjustment Project (Minneapolis-St. Paul metropolitan area) for whom both interview and employer work histories were obtained.
2. In EOPP data, union coverage coded as yes if employer reports majority of (non-supervisory) workers are covered.
3. Sample limited to a single company.
4. Sample limited to those respondents for whom respondents’ reports of employer could be matched to D&B files. Successfully matched employers tended to be larger and in business longer than employers in the overall sample. The authors note that due to potential inaccuracies in D&B counts of employer size, correlations listed above “probably understate the correlation between worker reports and perfectly accurate measures of employer size” (p. 280).
Table 8. Assessment of Measurement Error: Health Care Utilization, Expenditures, and Insurance

<table>
<thead>
<tr>
<th>Reference</th>
<th>Variables of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cannell and Fowler (1963)</td>
<td>Physician visits</td>
<td>Physician records</td>
<td>• Errors of omission increase as a function of the length of recall: 15% unreported for one-week recall; 30% unreported for two-week recall</td>
</tr>
<tr>
<td>Cannell, Fisher, and Bakker (1965)</td>
<td>Hospitalization utilization</td>
<td>Hospital records</td>
<td>• Overall, 12 percent of hospitalizations not reported for one year recall period. Errors of omission related to: • length of the recall period, ranging from 3% for hospitalizations within 10 weeks of interview to 40% not reported for those occurring 52 weeks prior to interview. • length of hospital stay, with longer stays (30+ days) subject to lower rates of omissions (~5%) than shorter stays (26% underreporting for stays of 1 day). • perceived threat of the condition associated with the stay; 10% rate of omission for conditions judged to not be threatening to 21% of those judged most threatening. • Sample size of 1,505 persons with 1,833 hospital discharges during the past year.</td>
</tr>
<tr>
<td>Yaffe and Shapiro (1979)</td>
<td>Hospitalizations, physician visits, dental visits, prescription medicines</td>
<td>Provider records (Note: agreement rates are for population estimates and not at the person level; data provided for two separate geographical areas)</td>
<td>• Agreement rates for utilization: Office-based physician visits: 72%-83% Clinic visits: 39%-54% Emergency Room: 94%-96% Dental visits: 82%-86% Prescribed medicines: 61%-75% Hospitalizations: 94%-97% • Agreement rates for expenditures: Office-based physician visits: 68%-78% Clinic visits: 31%-38% Emergency Room: 65%-90% Dental visits: 89%-99% Prescribed medications: 65%-77% Hospitalizations: 87%-99% • Based on completed interviews with 802 families with information for 2,300 persons</td>
</tr>
<tr>
<td>Walden, Horgan, and Cafferata (1982)</td>
<td>Health insurance characteristics</td>
<td>Health insurance plan information collected from insurance provider</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>----------------------------------</td>
<td>---------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percent for whom agreement between two sources of information:</td>
<td>Existence of out-of-pocket payments: 78%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Amount of out-of-pocket payments: 32%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sources of premium payments: 74%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Amount of premium paid by others: 28%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Agreement on coverage characteristics:</td>
<td>Semi-private hospital room: 86%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physician in-patient surgery: 88%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other in-patient physician: 80%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maternity: 55%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eye exam for glasses: 73%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routine dental care: 78%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Orthodontia: 69%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ambulatory x-rays; diagnostic tests: 70%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ambulatory physician: 54%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ambulatory prescriptions: 54%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Outpatient mental health care: 29%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inpatient mental health care: 32%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nursing home/similar facility: 33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Based on data for 20,001 individuals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohen and Carlson (1994)</td>
<td>Health care expenditures</td>
<td>Medical records</td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------------------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Expenditures by type of utilization (standard errors in parentheses):</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inpatient hospital stays (n=1,050; 19% of hospitalizations)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (household): $5228 (630)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (record): $4975 (451)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (Simple diff): $252 (451)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (absolute difference): $847 (210)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Emergency Room visits (n=1,765; 21% of emergency visits):</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (household): $155 (10.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (record): $153 (7.1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (Simple diff): $2 (8.9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (absolute difference): $59 (9.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Outpatient Department visits (n=2,609; 13% of outpatient visits):</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (household): $227 (15.4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (record): $238 (15.2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (Simple diff): $10 (17.1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (absolute difference): $132 (17.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medical provider contacts (n=17,169; 11% of medical provider visits)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (household): $47 (1.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (record): $53 (1.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (Simple diff): $5.5 (1.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (absolute difference): $23 (1.5)</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Variables of Interest</td>
<td>Validation Source</td>
<td>Findings</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------</td>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>National Center for Health Statistics (1961)</td>
<td>Chronic conditions</td>
<td>Medical records</td>
<td>• Correspondence between medical records and household reports of chronic conditions ranged from 20% to 44%. Higher rates of correspondence between two data sources associated with: &lt;br&gt; • use of a checklist rather than free recall &lt;br&gt; • higher number of physician visits during the past year associated with the condition &lt;br&gt; • most recent physician service within the past two weeks. &lt;br&gt; • No differences in the quality of the report by self/proxy status, age, gender, or race. &lt;br&gt; • Sample of approximately 1400 families; medical records indicate 4,648 chronic conditions among respondents</td>
</tr>
<tr>
<td>National Center for Health Statistics (1967)</td>
<td>Chronic conditions</td>
<td>Medical records</td>
<td>• Errors of omission as a function of time since last visit and response task: &lt;br&gt; Time            Recall           Recognition &lt;br&gt; &lt; 2 wks          58%                32%   &lt;br&gt; 2wk - 4mo.      79%                51%   &lt;br&gt; &gt; 4 mo.           84%                66%</td>
</tr>
<tr>
<td>Katz, Downs, Cash, and Grotz (1970)</td>
<td>Index of Activities of Daily Living; study of 270 patients at discharge</td>
<td>Correlation between index scores and other assessment scales</td>
<td>• Correlation coefficient of .50 with mobility scale and .39 with house confinement measure</td>
</tr>
<tr>
<td>Madow (1973)</td>
<td>Chronic conditions</td>
<td>Medical records</td>
<td>• 46.8% of conditions recorded in medical records unreported in household interview (underreporting) while 40.4% of household reported conditions were not listed in medical record (overreporting). &lt;br&gt; • Interviews with approximately 5,000 persons with over 15,000 conditions obtained from the two data sources</td>
</tr>
<tr>
<td>Bergner, Bobbitt, Carter, and Gilson (1981)</td>
<td>Sickness Impact Profile (SIP); various trials of different samples</td>
<td>Test-retest reliability; internal consistency</td>
<td>• Test-retest reliability scores ranged from .97 for interviewer-administered to .87 for self-administered forms. Alpha coefficients for 136-item version=.94 (interviewer administered) and .81 (self administered)</td>
</tr>
<tr>
<td>Study</td>
<td>Health Conditions</td>
<td>Medical Records</td>
<td>Internal Consistency Reliability (Chronbach’s $\alpha$)</td>
</tr>
<tr>
<td>-------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>Johnson and Sanchez (1993)</td>
<td>3-digit condition classification</td>
<td>40.4%</td>
<td>131-summary grouping 54.4%</td>
</tr>
<tr>
<td>McHorney, Ware, Lu, and Sherbourne (1994)</td>
<td>Eight scales from the SF-36 instrument; study conducted among 3,443 patients with one or more chronic conditions</td>
<td>Physical Functioning .93</td>
<td>Role Physical .84</td>
</tr>
<tr>
<td>McHorney, Kosinski, and Ware (1994)</td>
<td>Eight scales from the SF-36 instrument; study conducted among national sample of adults, n=1692; self administered</td>
<td>Physical Functioning .94</td>
<td>Role Physical .89</td>
</tr>
</tbody>
</table>
Table 10. Assessment of Measurement Error: Education

<table>
<thead>
<tr>
<th>Reference</th>
<th>Variables of Interest</th>
<th>Validation Source</th>
<th>Findings</th>
</tr>
</thead>
</table>
| Siegal and Hodge (1968) Table 2.1 and Figure 2.1 | Years of schooling 1960 Census | Post-enumeration Survey (PES) | • variance (interview) = 12.82  
• variance (record) = 13.03  
• r(interview, record) = .933  
• r(error, record) = -.205 |
| Bishop (1974) Table 1 | Years of Schooling, Census, 1950-70 | Current Population Survey (CPS)  
Post-enumeration Survey (PES)  
Census Reinterview Survey (CRS) | • 1970 Census  
r(Census, CPS) = .88 (males), .88 (females)  
b(CPS, Census) = .88 (males), .86 (females)  
b(Census, CPS) = .89 (males), .89 (females)  
• 1960 Census  
r(Census, PES) = .93  
r(Census, CRS) = .91  
b(PES, Census) = .94  
b(CRS, Census) = .91  
b(Census, PES) = .93  
b(Census, CRS) = .92  
• 1950 Census  
r(Census, PES) = .86  
b(PES, Census) = .85  
b(Census, PES) = .87 |
| Bielby, Hauser, and Featherman (1977) Tables 2 and 3 | Years of schooling March 1973 CPS | 1973 Occupation Changes in a Generation (OCG) and OCG re-interview (OCG-R) | • Means and variances:  
CPS 12.18 8.24  
OCG 11.98 11.70  
OCG-R 12.12 8.58  
r(CPS, OCG) = .801  
r(CPS, OCG-R) = .921  
r(OCG, OCG-R) = .838 |
<table>
<thead>
<tr>
<th>Source Descriptions</th>
<th>Description</th>
<th>Source Descriptions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ashenfelter and Krueger (1994) Tables 1, 2, and 8</td>
<td>Years of schooling, twins attending twin festival</td>
<td>Twin’s report</td>
<td>• MZ twins DZ twins variance (own report) 4.67 4.04 variance (twin report) 4.58 4.28 r(own report, twin report) .90 .91 r(Δown report, Δtwin report) .57 .74</td>
</tr>
<tr>
<td>Kane, Rouse, and Staiger (1999) Appendix Table 1</td>
<td>Years of schooling, National Longitudinal Class of 1972</td>
<td>Transcript data</td>
<td>• Sample proportions:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller, Mulvey, and Martin (1995) Tables 1 and 2</td>
<td>Years of schooling, Australian Twin Register</td>
<td>Twin’s report</td>
<td>• MZ twins DZ twins variance (own report) 6.25 5.86 variance (twin report) 5.39 4.72 r(own report, twin report) .88 .82 r(Δown report, Δtwin report) .36 .60</td>
</tr>
<tr>
<td>Rouse (1999) Table 1; appendix tables 1a and 1b</td>
<td>Years of schooling, twins attending twin festival</td>
<td>Twin’s report</td>
<td>• MZ twins variance (own report) 4.24 variance (twin report) 4.28 r(own report, twin report) .92 r(Δown report, Δtwin report) .62</td>
</tr>
</tbody>
</table>