Abstract—We propose a new methodology that does not assume a prior specification of the statistical properties of the measurement errors and treats all sources as noisy measures of some underlying true value. The unobservable true value can be represented as a weighted average of all available measures, using weights that must be specified a priori unless there has been a truth audit. The Census Bureau’s Survey of Income and Program Participation (SIPP) survey jobs are linked to Social Security Administration earnings data, creating two potential annual earnings observations. The reliability statistics for both sources are quite similar except for cases where the SIPP used imputations for some missing monthly earnings reports.

I. Introduction

In the large and long literature on measurement error, most studies begin with data that are believed to contain errors and look for a way to quantify those errors. The goal is to account for and remove the effects of those errors. Defining data errors fundamentally requires some measure of truth: an objective standard by which the accuracy of the data can be judged. A common approach to this measurement error problem is to find a second data source that contains the “truth” and define errors in the first data set as the difference between the two sources. We believe, however, that the assumption that some data contain errors while other data do not is fundamentally flawed. While the error-generating process may be different in the two sources, no source is likely to be completely error free.

In this paper, we expand the measurement error literature in two ways. Our first contribution is to extend the methodology by showing that defining truth with respect to an observed quantity requires a researcher to place priors on which source of data is the most reliable. These priors define the measurement error and its properties. After recognizing this dependence on prior beliefs, we relax the assumption that one source of data is truth. We show how different priors lead to different measures of truth and, subsequently, different amounts of error. We then specify and estimate a multivariate linear mixed-effects model (MLMM) in the spirit of Abowd and Card (1989) and consider the level of error in both the fixed effects (the relationship between the measure and the observable characteristics of respondents) and the random effects (relationship between the measure and the unobservable characteristics of respondents and their employers), showing how different priors about the truth lead to different conclusions. Our approach includes the special case of declaring one source to be truth and can be used with complete or incomplete data from any number of sources.

Our second contribution is to implement our method using job-level annual earnings from two sources: the Census Bureau’s Survey of Income and Program Participation (SIPP) and the Internal Revenue Service (IRS)/Social Security Administration (SSA) W-2 forms. We use the largest nationally representative sample to date—five SIPP panels—matched at the job level to administrative data from the SSA’s Detailed Earnings Record (DER) to study the measurement error in earnings for these important national data. Our comprehensive data allow us to combine the early focus in the survey measurement error literature on employer versus employee reports, which were inherently job-level studies, with the later focus on comprehensive samples of survey respondents regularly used by empirical researchers, which were conducted at the person level because the particular survey studied, the Current Population Survey, collected complete data only at the person level.1

1 When both job-level and person-level earnings data are available, as in the SIPP, many studies use the job-level earnings measure, in combination with the survey hours measures, to construct outcomes and control variables of interest. See, for example, Gottschalk and Moffitt (1999), which uses the SIPP, CPS, and PSID in exactly this way, to study the effects of employer transitions.
We use the internal, confidential SIPP data that have uncapped earnings and a carefully edited job history intended to accurately track movement across jobs over the course of a year. The confidential administrative data also have uncapped earnings, measured separately for each employer. Among validation studies done with survey data, our analysis uses the most complete set of inputs to date. We analyze a range of priors about the reliability of each data source. We report the resulting measures of error for both the fixed and random components of a multivariate earnings equation estimated by residual maximum likelihood (REML).

The paper proceeds as follows. In section II, we review past findings from studies of measurement error in survey earnings and discuss how these relate to our study. In section III, we lay out our methodological framework for identifying and quantifying errors in earnings. We give a brief overview of our data and the linking process between the two sources in section IV. In section V, we present our results quantifying error in earnings in the SIPP and W-2 data. We conclude in section VI with thoughts on how both producers and users of public use data might take account of measurement error.

II. Background

Early studies of measurement error, beginning with Fuller (1987), defined observed quantities as the sum of unobserved true values and error. Hence, a variable \( Y_t \) is decomposed into \( y_t + u_t \), the sum the of true value, \( y_t \), and measurement error, \( u_t \). Assuming that the \( y_t \) and \( u_t \) are uncorrelated, one can calculate a reliability ratio that gives the percentage of total variance that is true variance:

\[
\kappa_{yy} = \frac{\text{Cov}[y, Y]}{\text{Var}[Y]} = \frac{\sigma_{yy}}{\sigma_{yy} + \sigma_{uu}}.
\]

This statistic is important because when a second variable \( A_t \) is a function of the true value \( y_t \) plus some error,

\[
A_t = \beta_0 y_t + e_t,
\]

but is regressed on \( Y_t \),

\[
A_t = \hat{\beta} Y_t + e_t = \hat{\beta}(y_t + u_t) + e_t,
\]

the reliability ratio defines the ratio of \( \hat{\beta} \) to \( \beta \). This is evident from the formula for the expected value of \( \hat{\beta} \), which, when \( \sigma_{yu} = 0 \) and \( \sigma_{ue} = 0 \), is given by

\[
E[\hat{\beta}] = \frac{\text{Cov}[A, Y]}{\text{Var}[Y]} = \frac{\beta \sigma_{yy}}{\left(\sigma_{yy} + \sigma_{uu}\right)} = \beta \kappa_{yy}.
\]

Angrist and Krueger (1999) define a reliability ratio for first differenced quantities \( \Delta Y = (Y_t - Y_{t-1}) = (y_t - y_{t-1}) + (u_t - u_{t-1}) \) as

\[
\kappa_{\Delta y \Delta y} = \frac{\sigma_{yy}}{\left(\sigma_{yy} + \sigma_{uu}\right) \left(\frac{1 - \tau}{1 - \rho}\right)},
\]

where \( \tau \) is the autocorrelation coefficient of the measurement error and \( \rho \) is the autocorrelation coefficient of \( y_t \). If \( \rho > \tau \), then \( \left(\frac{1 - \tau}{1 - \rho}\right) \) is greater than 1 and the reliability ratio declines relative to the ratio for levels of \( y_t \).

Finally if \( Y_t \) is used as a dependent variable and regressed on \( x_t \) and the covariance of \( y_t \) and \( u_t \) is not 0, then using \( u_t = \delta y_t + v_t \), where \( \delta \) is called the attenuation bias, Bound et al. (1994) show that the coefficient on \( x_t \) in the regression

\[
Y_t = (1 + \delta) y_t + v_t = x_\beta + \epsilon
\]

will be biased since

\[
E[\hat{\beta}/\beta] = \frac{1}{1 + \delta}.
\]

Hence, only if the variance and structure of the measurement error are known can unbiased estimators of \( \beta \) be obtained. Those studying measurement error have focused on estimating \( \kappa_{yy} \) and testing whether the assumptions of classical measurement error were violated. Studies that obtain a second report for the measured variable \( Y_t \) in order to calculate \( \sigma_{uu} \) and \( \sigma_{yy} \) have been called validation studies (as in, for example, Bound, Brown, & Mathiowetz, 2001). Without exception, the second report is treated as \( y_t \) (i.e., truth) and the measurement errors are calculated as \( u_t = Y_t - \hat{y}_t \) (first report) − \( \hat{y}_t \) (second report). The properties of these errors are then investigated. Researchers have often concluded that the assumptions of classical measurement errors are violated and that the errors are correlated with the true values, i.e., \( \sigma_{yu} \neq 0 \). However, they acknowledge that their models are driven by the assumption that they obtained a true measure of \( y_t \). Without this assumption, there would be no way to determine the relation between the errors and the true values. This assumption is fundamentally untestable and is justified solely by the authors’ knowledge of the quality of the secondary data source. We briefly discuss the history of this interpretation in the context of earnings validation studies next.

Some prominent examples of validation studies are Mellow and Sider (1983), Duncan and Hill (1985), Bound et al. (1994), Pischke (1995), Bound and Krueger (1991), Kapteyn and Ypma (2007), Gottschalk and Huynh (2010), and Meijer, Rohwedder, and Wansbeck (2012). The first four papers are similar to ours in that they estimate measurement error at the job level. Mellow and Sider (1983) used a special supplement to the January 1977 Current Population Survey (CPS) that obtained employment information from both workers and employers. Looking at matched pairs with both employer and employee wage reports, they found that employer-reported wages exceeded worker reports by 4.8% on average. Although they did not calculate a reliability ratio per se, they did test the sensitivity of statistical models and concluded that wage regressions were generally not that sensitive to the source of information: worker versus employer. In particular, the returns to education and experience were very similar across their different regression equations. This result is consistent with a relatively high reliability ratio.

\[2\text{Bound et al. (1994) also derive a formula for the reliability ratio when } y_t \text{ and } u_t \text{ are correlated.}\]
Like Mellow and Sider, Duncan and Hill (1985), Bound et al. (1994), and Pischke (1995) examine two wage reports for each job they observe. However, they do not use a representative sample of people in the United States but rather sample workers from a large anonymous manufacturing company. Workers at the company were interviewed using a Panel Study of Income Dynamics (PSID) survey instrument, and then information for these workers was obtained from company records. The authors treated the company reports of annual earnings as measures of true earnings values and considered any differences between worker and employer reports to be errors on the part of the workers, justifying this decision as due to their confidence in the accuracy of the company’s records.

The survey was carried out in two waves, and the authors reported a ratio of noise to total variance \( \left( \frac{\sigma_{\text{ur}}}{\sigma_{\text{uu}} + \sigma_{\text{yr}} + \sigma_{\text{ry}} + \sigma_{\text{rr}}} \right) \) of 0.302 for annual earnings in 1986 and 0.151 for annual earnings in 1982. They found evidence that errors in earnings were correlated with the true levels of earnings and reported noise-to-variance ratios that took account of this covariance as 0.239 in 1986 and 0.076 in 1982. They estimated the proportional attenuation bias when using earnings as a dependent variable, \( \delta \), as \(-0.172\) for 1986 and \(-0.104\) for 1982. Finally, the authors compared earnings equations using the two data sources and found that relative to the employer provided data, employee interview data overstated the return to education by 40% and understated the return to tenure by 20%.

Bound and Krueger (1991) depart from the job-level method and compare total earnings in a given year from two sources. Like our study, they use survey and administrative data, linking March 1977 and 1978 CPS respondents to Social Security earnings records. They reported large negative correlations between measurement error and true earnings for both CPS reference years (1976 at \(-0.46\) and 1977 at \(-0.42\)). They reported reliability ratios that did and did not take account of these correlations as 1.016 and 0.844, respectively, for 1976 and 0.974 and 0.819 for 1977.

Bound et al. (2001, p. 3832) summarized the general approach of all of these studies as: “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).”

Gottschalk and Huyhn (2010) also studied person-level earnings using matched SIPP and SSA data to conduct an extensive investigation of the effects of measurement error on earnings inequality and mobility. These are the same data sources that we use, but they use only the 1996 SIPP panel.

Initially these authors do not label the SSA data as truth and instead describe their study as quantifying differences between the two sources. Nonetheless, their ability to study the impact of SIPP error on mobility and inequality measures hinges on a definition of error that requires them to declare one source as truth. Using the DER as truth, the authors estimate a traditional reliability ratio of 0.67 for the SIPP. However, they conclude that measurement error is mean reverting and show that in their framework, this type of error partially offsets the bias in estimates of inequality in the SIPP. They also conclude that measurement error is correlated over time, and this diminishes the attenuation bias in the correlation of earnings and lessens the impact of measurement error on estimates of earnings mobility in the SIPP.

Kapteyn and Ypma (2007) use person-level linked survey and administrative data from Sweden. Like Gottschalk and Huyhn and our paper, they do not declare either source to be true earnings, which formally is a latent variable in their mixture of normals model. They use a prior specification of the effects of misclassification errors and different survey measurement errors to identify the posterior probabilities associated with the administrative and survey measures being true. Meijer et al. (2012) show that the Kapteyn and Ypma model is a special case of a mixture factor model with a specification very similar to ours. Their generalized model also identifies the marginal prior probabilities, which are treated as hyperparameters, and the conditional posterior probabilities by making functional form assumptions about the relation between the survey and administrative measures. The relation between the identification strategy in these mixture factor models and our multivariate linear mixed-effects model is discussed in section IIIB.

A final related study by Roemer (2002) uses matched CPS, SIPP, and DER data to study the distribution of annual earnings. Rather than focus on reliability statistics and regression-coefficient comparisons, Roemer compares the percentiles of the annual earnings distributions from the three sources. Treating the DER as truth, he concludes that both the SIPP and CPS estimate a person’s percentile rank more accurately than the dollar amount of earnings. In his analysis, the SIPP displays a shortage of high-earning workers compared to the DER.

### III. Statistical Model

#### A. Multivariate Linear Mixed-Effects Model

In this section we lay out the general statistical model that we employ to estimate the fixed and random components of our joint model for SIPP and DER earnings outcomes observed on the same matched job. The underlying specification is a multivariate linear mixed-effects model (MLMM). The advantage of the MLMM framework is that it shows with full generality how to accommodate two or more matched observations of earnings on the same job, how to
vary the prior assumptions about which measure is “true” systemically, and how to use external audit information, if available, to update the posterior distribution over which value is “true.”

The outcome under study, which is the dependent variable in the model, is \( y_{ist} \), a \( 1 \times Q \) vector of measures of log earnings for individual \( i = 1, \ldots, I \) in sequential job spell \( s \in \{1, \ldots, S\} \), and time period \( t \in \{1, \ldots, T\} \), where \( Q \) is the total number of sources of earnings reports. The indices \( s \) and \( t \) are always ordered sequentially, but not every \( i \) has values for every level of \( s \) and \( t \). Define the vectors \( x_{ist} \), the \( 1 \times K \) design of the fixed effects associated with sex, race, education, experience, and so on; \( d_i \) the \( 1 \times I \) design of the random effects associated with person \( i \); and \( f_{ist} \) the \( 1 \times J \) design of the random effects associated with the employer of \( i \) in job spell \( s \) during period \( t \). The full model is

\[
y_{ist} = x_{ist}B + d_i\Theta + f_{ist}\Psi + \eta_{ist},
\]

(1)

where \( B \) is the \( K \times Q \) matrix of fixed-effect coefficients, \( \Theta \) the \( I \times Q \) matrix of random person effects with \( i \)th row \( \theta_i \), the \( 1 \times Q \) vector of random person effects for individual \( i \); \( \Psi \) is the \( J \times Q \) matrix of random employer effects with \( j \)th row \( \psi_j \), the \( 1 \times Q \) vector of random employer effects for each employer \( j \), and \( \eta_{ist} \) is the \( 1 \times Q \) residual vector for individual \( i \) in job spell \( s \) for period \( t \).

Stacking the observation over \( i, s, t \), the model becomes

\[
Y = XB + ZU + H,
\]

(2)

where \( Y \) is the \( N \times Q \) matrix of dependent variables with \( N \) equal to the total number of person, job spell, and years in the data; \( X \) is the \( N \times K \) design matrix for all fixed effects; \( Z \equiv [D F] \) is the \( N \times (I + J) \) design matrix for the combined random effects; \( U \equiv [\Theta^T \Psi^T]^T \) is an \( (I + J) \times Q \) matrix of random effects; and \( H \) is an \( N \times Q \) matrix of residuals. Equation (2) is a multivariate linear mixed-effects model represented in canonical form. By construction, every column of \( Y \) can be represented as a single linear mixed-effects model, also represented in canonical form, for dependant variable \( Y_{(q)} \), where the subscript \((q)\) denotes selection of the indicated column from the associated matrix. For example, the \( q \)th column has the form

\[
Y_{(q)} = XB_{(q)} + ZU_{(q)} + H_{(q)}
\]

(3)

for \( q = 1, \ldots, Q \).

Parameterize the stochastic structure of equation (2) as follows:

\[
E[\eta_{ist}|X,Z] = 0,
\]

\[
\text{Var}[\eta_{ist}|X,Z] = \Sigma_0, \quad Q \times Q,
\]

\[
\text{Cov}[\eta_{ist},\eta_{ist'}|X,Z] = 0, \quad i' \neq i, \forall s, s', t, t',
\]

\[
\text{Cov}[\eta_{ist},\eta_{ist'}|X,Z] = \Sigma_{t-t'}, \quad Q \times Q.
\]

where \( \Sigma_{t-t'} \) is the \( Q \times Q \) autocovariance matrix of \( \eta_{ist} \) at lag \( |t - t'| \). We use ASREML (Gilmour et al. 2009) to fit equation (2) using the residual maximum likelihood method (REML) and assuming that vec \((U)\) and vec \((H)\) have independent joint normal distributions with zero means and the covariance structure specified above. This special-purpose software may not be as familiar to economists as it is to biostatisticians; however, all of the parameters that we specify in the MLMM are identified using conventional methods—proof that the residual likelihood function has a well-defined maximum for the given specification. The ASREML software checks these identification conditions (or estimability conditions, as they are known in biostatistics) and notes violations at the optimum, so that only estimates of identifiable parameters, for both fixed and random effects, are reported. We do not elaborate on the estimation because ASREML produces the REML estimates of all parameters and their estimated covariance matrix taking account of the full stochastic structure of the specified model. Although they are identifiable and calculated by the software, we make no use of the estimated person and employer effects, \( \hat{U} = [\Theta^T \Psi^T]^T \) in our notation.

To simplify the exposition of the results, we note here the exact formulas for all of the variance components for the case \( Q = 2 \) in which the SIPP value is listed first \((q = 1)\) and the DER value is listed second \((q = 2)\). Then we have

\[
\Sigma_0 = \begin{bmatrix}
\sigma_c^2 + \sigma_1^2 & \sigma_1^2 \\
\sigma_1^2 & \sigma_1^2 + \sigma_2^2 \\
\end{bmatrix},
\]

(4)

\[
\Sigma_{t-t'} = \begin{bmatrix}
\rho_c|t-t'|\sigma_c^2 & \rho_c|t-t'|\sigma_1^2 \\
\rho_c|t-t'|\sigma_1^2 & \rho_c|t-t'|\sigma_2^2 \\
\end{bmatrix},
\]

(5)

\[
G^{(0)} = \begin{bmatrix}
\sigma_{\theta_1}^2 & \rho_0\sigma_{\theta_1}\sigma_{\theta_2} \\
\rho_0\sigma_{\theta_1}\sigma_{\theta_2} & \sigma_{\theta_2}^2 \\
\end{bmatrix},
\]

(6)

\[
G^{(\psi)} = \begin{bmatrix}
\sigma_{\psi_1}^2 & \rho_\psi\sigma_{\psi_1}\sigma_{\psi_2} \\
\rho_\psi\sigma_{\psi_1}\sigma_{\psi_2} & \sigma_{\psi_2}^2 \\
\end{bmatrix},
\]

(7)

We require the assumption that \( G^{(0)}, G^{(\psi)}, \) and \( \Sigma_{t-t'} \) for all \( t' \) are consistently estimated when either \( U \) or \( H \) is nonnormal. This is reasonable because the estimation is based on minimizing an objective function that depends on only the first two moments of vec \((Y)\), albeit with a form that is based on the multivariate normal distribution. Departures from normal are likely to affect the covariance matrix in the asymptotic distribution of the unique parameters of \( G^{(0)}, G^{(\psi)}, \) and \( \Sigma_{t-t'} \), which we have not investigated.
B. Defining True Values, Associated Measurement Error, and Reliability Statistics for the Random Effects

Signal, measurement error, and truth audits. Let \( \omega \equiv (\omega_1, \ldots, \omega_Q)^T \) be a \( Q \times 1 \) vector where \( t^T \omega = \sum \omega_i \leq 1 \), and \( t \) is always a conformable column vector of 1s. The elements of the vector \( \omega \) correspond to the prior probabilities or weights associated with each of the elements being the correct or true value. The signal is, then, the expected true value, where the expectation is taken over the prior probabilities \( \omega \),

\[
\text{Sig}(y_{ist}) \equiv y_{ist} \omega,
\]

and the measurement error is the deviation of each measure from the signal component:

\[
\text{ME}(y_{ist}) \equiv y_{ist} - \text{Sig}(y_{ist}) = y_{ist} - y_{ist} \omega^T = y_{ist} [I - \omega^T].
\]

Hence, the weight vector \( \omega = (0, 1)^T \) when \( Q = 2 \) corresponds to declaring that the second measure (in our case DER) is correct. This is precisely the assumption in historical administrative record measurement error studies. We consider alternatives—in particular, \( \omega = (0.5, 0.5) \), the case where either measure is equally likely to be the truth, and \( \omega = (0.1, 0.9) \), the case where the DER is much more likely to be true. Neither of these definitions imposes multivariate normality on the signal or measurement error.

In a full Bayesian analysis, one would model \( \omega_{ist} \) as a latent data quality indicator whose prior expectation is the \( \omega \) in the preceding paragraph. Let \( g_{ist} \) represent the \( 1 \times L \) vector associated with the design of additional information used to determine the correctness of the measurement, and let \( w_{ist} \) represent the \( 1 \times Q \) outcome vector from the data audit whose elements are all zero except for a single column, \( q \), coded 1, which indicates that for the earnings outcome of person \( i \), at job spell \( s \), in time period \( t \), \( q \) was the correct measurement. Then the additional equation system that models \( w_{ist} \) as a function of \( x_{ist}, d_{ist}, f_{ist}, \) and \( g_{ist} \) provides the framework for computing a posterior estimate of \( \omega_{ist} \) that would replace the prior estimate used in equations (8) and (9).

It is also worth noting that our method generalizes to the case where there are not the same number of measurements for each \( i, s, t - \text{tuple} \). We implement this feature in our estimation, treating missing values of one or the other measure for a particular year \( t \) as ignorable (Rubin, 1976).

Identification of measurement error. No truth audit was conducted as a part of our SIPP-to-DER matching exercise. Therefore, we have no outcome data corresponding to \( w_{ist} \), even for a subsample. Hence, the prior \( \omega \) must equal the posterior \( \omega \), which is why we form the signal and measurement error components as we do. This assumption merely formalizes what has been the practice for more than three decades while also showing exactly what information would be required to eliminate the use of purely prior information about the correctness of one of the outcomes in measurement error studies.

Because the Kapteyn and Ypma (2007) and Meijer et al. (2012) models are facially similar to our model, we discuss them now in the context of the assumptions that identify measurement error. In our model and in theirs, if any measure is declared to be true, \( w_{isq} = 1 \) for some \( q \) in our notation, then measurement errors are observed and can be calculated as \( y_{isq}^\prime - y_{isq} \) for all \( q' \neq q \), again in our notation. In our model, without an audit there is no measurement on truth. We have not formalized the judgments that analysts might make based on the observed data to infer truth in the absence of an audit. This is the critical distinction between our approach and theirs.

Kapteyn and Ypma (2007), explicitly model the analyst’s behavior. If the survey and administrative measures match, then the analyst declares them both true, and the properties of the model for the administrative data and measurement-error-free survey data are identified. The case is “labeled” in their notation. The analyst makes two further assumptions: (a) there is no measurement error in the administrative earnings variable, only mismatch (data matched from the wrong individual), (b) survey measurement error, which comes in two types, must increase the variance of the survey measure relative to the administrative measure when there is no mismatch. These two assumptions for the “unlabeled” cases effectively declare the administrative measure to be the truth whenever it is correctly matched to the survey record and provide enough information to estimate posterior probabilities of truth. Meijer et al. (2012) generalize these assumptions and show that the posterior probabilities are sensitive to the identifying assumptions, as expected.

Our model shares the property that if there are only two observed measures and they are equal, then there is no observed measurement error. But we make no use of the structure of the error components in the two measures that would identify measurement error specifically. Moreover, during an audit, an analyst could discover that neither measure contained, for example, off-the-book payments like tips. Then neither measure would be true, and the auditor would have to estimate such payments or declare a latent third measure to be true. Our model explicitly allows this outcome. Furthermore, an analyst who wished to use prior information to characterize the effects of different types of errors on the distribution of \( y \) could calculate the posterior distribution of \( \omega \) given these model components. This is what Meijer et al. (2012) have done using the mixture factor analysis model.

Our method makes clear than in the absence of an audit, measurement error is entirely defined a priori and not from any observed data. Classical validation studies like the ones discussed in section I and their more sophisticated recent counterparts all identify the measurement error by placing very strong priors on the data generation process. Our contribution is to generalize this identification strategy to accommodate other priors and a broader set of potential measures.
Reliability statistics. To compute reliability statistics, we require estimates of \( \text{Var}[\text{Sig}(y_{ist})|X,Z] \) and \( \text{Var}[\text{ME}(y_{ist})|X,Z] \). These can be computed from the stochastic structure of equation (2):

\[
\text{Cov}[y_{ist}, y_{ist}'|X,Z] = G^{(0)} + G^{(\psi)} + \Sigma_{[t,t']},
\]

(10)

\[
\text{Cov}[\text{Sig}(y_{ist}), \text{Sig}(y_{ist}'|X,Z)] = \omega^T (G^{(0)} + G^{(\psi)} + \Sigma_{[t,t']}) \omega,
\]

(11)

\[
\text{Cov}[\text{ME}(y_{ist}), \text{ME}(y_{ist}'|X,Z)] = \left[ I - \omega^T \right]^T (G^{(0)} + G^{(\psi)} + \Sigma_{[t,t']}) \left[ I - \omega^T \right].
\]

(12)

The traditional measures for the case \( Q = 2 \) can be computed at all lags using \( \omega = (0, 1)^T \).

Some care must be taken in using the formulas in equations (11) and (12) because they do not represent an orthogonal decomposition of equation (10). The conventional reliability ratio for measure \( q \) is defined as the ratio of its signal variance to its total variance. With the SIPP and DER measures in positions 1 and 2, respectively, the traditional reliability ratios are

\[
\text{TRR}_{0,SIPP} = \frac{\omega^T (G^{(0)} + G^{(\psi)} + \Sigma_0) \omega}{\{G^{(0)} + G^{(\psi)} + \Sigma_0\}_{11}},
\]

(13)

and

\[
\text{TRR}_{0,DER} = \frac{\omega^T (G^{(0)} + G^{(\psi)} + \Sigma_0) \omega}{\{G^{(0)} + G^{(\psi)} + \Sigma_0\}_{22}},
\]

(14)

where the notation \( \{ij \} \) means to extract the \( i,j \)th element of the matrix in \( \{ \} \). The difficulty with equations (13) and (14) is that they are not bounded above by unity because either of the two measures in isolation (SIPP or DER) can omit elements that should be measured or include elements that should not. Our measurement error model has only the traditional reliability ratio property of being bounded above by unity when either the SIPP or DER is true or when the two measures are exchangeable, as in conventional survey reliability estimation where the measures are obtained by repeated application of the same survey instrument (see Groves et al., 2004).

We choose instead to define the reliability statistic by generalizing the index of inconsistency, the ratio of the measurement error variance to the total variance, and subtracting it from unity so that it has the interpretation of a reliability statistic:

\[
\text{RR}_{0,SIPP} = 1 - \left[ \frac{\{ I - \omega^T \}^T (G^{(0)} + G^{(\psi)} + \Sigma_0) \left[ I - \omega^T \right]}{\{G^{(0)} + G^{(\psi)} + \Sigma_0\}_{11}} \right]_{11}
\]

(15)

The reliability statistics in equations (15) and (16) reproduce the conventional reliability ratios when either of the measures is true or when the two measures are exchangeable. Because of the serial correlation caused by the structure of the individual, employer, and time effects, we also define reliability statistics at different lags. These are given by

\[
\text{RR}_{[t,t'-1],SIPP} = 1 - \left[ \frac{\{ I - \omega^T \}^T (G^{(0)} + G^{(\psi)} + \Sigma_{[t-t']} \left[ I - \omega^T \right]}{\{G^{(0)} + G^{(\psi)} + \Sigma_{[t-t']}\}_{11}} \right]_{11},
\]

(17)

and

\[
\text{RR}_{[t,t'-1],DER} = 1 - \left[ \frac{\{ I - \omega^T \}^T (G^{(0)} + G^{(\psi)} + \Sigma_{[t-t']} \left[ I - \omega^T \right]}{\{G^{(0)} + G^{(\psi)} + \Sigma_{[t-t']}\}_{22}} \right]_{22},
\]

(18)

These definitions require consistent estimates of the variance parameters \( G^{(0)}, G^{(\psi)} \) and \( \Sigma_{[t-t']} \), but they do not impose multivariate normality on the underlying data.

C. Defining True Values, Associated Measurement Error, and Reliability Statistics for the Fixed Effects

Using the MLMM specification in equation (2) and following the method from equation (8), we define

\[\text{Sig}(B) \equiv B_0,\]

(19)

where \( B \) is the matrix of fixed effects associated with the design \( X \). The true fixed effect is the \( \omega \)-weighted average of the SIPP and DER fixed-effect coefficients. These end points define the range of each fixed effect. The SIPP and DER measurement errors are defined as

\[\text{ME}(B_{(q)}) \equiv B_{(q)} - B_0,\]

(20)

It is worth noting here that it makes no sense to assume that the two measures are exchangeable because that is equivalent to assuming that the labels “SIPP” and “DER” are meaningless. The whole point of the analysis is that we know that outcomes were collected based on different measurement concepts (survey versus administrative records), so they should not have the same joint distribution if we exchange the labels.
where \( q = 1, 2 \), and the generalization to arbitrary \( Q \) is straightforward. We note that defining the signal and the measurement error in terms of the theoretical fixed-effect coefficients is exactly comparable to the methods discussed in section II, when the weight vector takes the value \((0, 1)\)^7, that is, when we assume that DER is truth.

Applying the MLMM model estimator in ASREML produces the estimates \( \hat{\mathbf{B}} \) and \( \hat{\mathbf{V}} \), which can be used directly to estimate the signal, measurement error, and associated standard errors for the fixed-effect coefficients. These are reported in section V for the same values of \( \omega \) as we use for the random effects.\(^5\)

We do not compute reliability statistics for the fixed effects because the measurement error in equation (20) has traditionally been interpreted as a bias, which is the interpretation we adopt—making point estimates and their standard errors more appropriate than reliability statistics.

D. Person-level Models

For comparison to the literature on the assessment of survey measurement errors using administrative data, we also estimate the MLMM for person-level outcomes. A person-level outcome for individual \( i \) in year \( t \) is defined as the sum of all observations \( y_{ist} \) over all jobs \( s \), including those jobs that matched between the SIPP and the DER and those that did not. Thus, for each person, there are \( q \) outcomes per year, and these outcomes differ across sources because of differences in reporting at the job level and because of differences in the number of jobs reported.

The base specification becomes

\[
y_{ist} = x_{ist} \mathbf{B} + d_i \mathbf{\Theta} + f_{it} \mathbf{\Psi} + \eta_{ist},
\]

where \( x_{ist} \) can be defined unambiguously because our observed covariates do not vary by employer. In order to be comparable to the literature, we do not include the design \( f_{it} \) in most of the person-level specifications. But we do report one set of results where the employer is defined as the one for which \( y_{ist} \) is a maximum over \( s \), that is, the employer with the greatest DER earnings during the year.\(^6\) All MLMM and reliability statistics for person-level models are defined in a manner that is strictly comparable to the job-level models, so we do not repeat any of the formulas here.

IV. Data Description

The fundamental unit of observation in this paper is a job, defined as a match between an individual and an employer. Data on jobs come from two sources: five Survey of Income and Program Participation (SIPP) panels conducted during the 1990’s and the Detailed Earnings Records (DER) extracted from the Social Security Administration Master Earnings File for the respondents in each of the five panels.\(^7\)

In the SIPP, data on earnings were reported monthly, while in the DER, earnings were reported annually. In both sources, there were multiple records per job from repeated interviews and annually filed W-2s. Hence, in order to compare earnings, we first had to identify jobs and group earnings records over time in each data source. After job records were created, individuals in each data set were linked by Social Security number and, then, job records from the SIPP and the DER were matched to each other for each individual. We describe each step of this process.

A. Creating a SIPP Jobs Data Set

All the SIPP panels conducted in the 1990s collected detailed labor force information from respondents every four months, or approximately three times per year, over the course of two and a half to four years. Respondents were asked questions about at most two jobs held during the previous four months, where the term job was loosely defined as working for pay. We used the longitudinal SIPP person ID, the wave (interview) number, and an edited longitudinal job ID that we created to combine records and create one observation per person per job. Appendix A in the online supplement, which includes all appendixes, describes the problems we found with the original job ID and gives a summary of how we created our edited version. The first column of table 1 shows the number of respondents in each SIPP panel who report working and the total number of jobs reported, using the three identifiers listed above to count jobs.\(^8\) Once we defined a set of jobs for each SIPP panel, we created annual earnings measures for each year covered by the survey by summing the appropriate monthly earnings reports.

In order to understand the comparison to another data source, it is important to first understand the concept of earnings as used during the SIPP interview. The field representative sought to ask each adult about his or her earnings, but if an adult was not present at the time of the interview, another adult could answer as a proxy. During the 1990–1993 SIPP panels, respondents (or proxies) were asked to report gross earnings from a specific employer in the following way: “The next question is about the pay … received from this job during the 4-month period. We need the most accurate figures you can provide. Please remember that certain months contain 5 paydays for workers paid weekly and 3 paydays for workers paid every 2 weeks. Be sure to include any tips, bonuses, overtime pay, or commissions. What was the total amount of pay that … received BEFORE deductions…?”


The edited SIPP job ID for the 1990–1993 panels was released by the Census Bureau as an update to the public use files and is available on the SIPP FTP website. The edited job ID is described in Stinson (2003).
on this job in . . . ?” The field representative read the name of each month and separately recorded earnings for that month. In the 1996 survey instrument, which was conducted using a computer-assisted personal Interview (CAPI) system, individuals (or proxies) could report earnings payments over a variety of time periods and the instrument automatically calculated monthly earnings. Field representatives (FRs) asked, “Each time he/she was paid by [Name of Employer] in [Month X], how much did he/she receive BEFORE deductions?” The field representative then followed up with questions about whether there were any other payments such as tips, bonuses, overtime pay, or commissions. Built-in consistency checks flagged earnings amounts outside certain bounds and prompted the FR to make corrections. Respondents were also asked to refer to earnings records if possible so as to give accurate responses. Thus, in the most accurate cases, these earnings reports most likely reflected the gross pay from monthly pay stubs.

B. Creating a DER Job-Level Data Set

The second source of data, DER, was a specialized extract from the SSA’s Master Earnings File that contained earnings histories for each SIPP respondent in the 1990, 1991, 1992, 1993, and 1996 panels with a validated SSN. The creation of the DER was a joint project between the Census Bureau and SSA. The Census Bureau asked each SIPP respondent at the time of the survey to provide an SSN. SSA then compared self-reported name, sex, race, and date of birth to their counterparts for the matching SSN on the Numident, an administrative database containing demographic information collected when every SSN was issued and updated when the individual had subsequent contacts with SSA. If a respondent’s name and demographics were deemed close enough to the name and demographics associated with the SSN in the Numident, the SSN was declared valid. This list of validated SSNs was the basis for extracting detailed earnings records from the SSA Master Earnings File.

A W-2 history for a SIPP respondent consisted of annual earnings, broken down by employer, from 1978 to 2000. The primary earnings variable came from box 1 of the W-2 form: wages, tips, and other compensation. This earnings variable was uncapped and represented all earnings that were taxable under federal income tax. For the purposes of this earnings comparison study, jobs with an employer (i.e., non-self-employment) held during the time period covered by the survey questions were used. In the second column of table 1, we show the number of SIPP respondents with DER records and a count of unique person-employer matches. Employers were identified in the DER by an IRS-assigned employer identification number (EIN). The EIN linked employers to the Business Register, the master list of all businesses maintained by the Census Bureau as the sampling frame for establishment-level surveys. Using this link, we merged information from the Business Register about the industry and name of the employer to each relevant job report.

<table>
<thead>
<tr>
<th>SIPP Panel</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>People with jobs</td>
<td>SIPP</td>
<td>DER</td>
<td>SIPP</td>
<td>DER</td>
<td>Job Match</td>
<td>Drop Missing Sample Years</td>
</tr>
<tr>
<td>1989–1992</td>
<td>Total jobs held</td>
<td>37,291</td>
<td>35,032</td>
<td>30,993</td>
<td>28,313</td>
<td>26,615</td>
<td>21,776</td>
</tr>
<tr>
<td>1990–1993</td>
<td>Total jobs held</td>
<td>40,818</td>
<td>58,020</td>
<td>32,447</td>
<td>25,258</td>
<td>21,761</td>
<td>18,331</td>
</tr>
<tr>
<td>1991–1995</td>
<td>Total jobs held</td>
<td>60,278</td>
<td>59,925</td>
<td>51,650</td>
<td>39,729</td>
<td>37,021</td>
<td>29,750</td>
</tr>
<tr>
<td>1992–1995</td>
<td>Total jobs held</td>
<td>61,094</td>
<td>81,320</td>
<td>47,723</td>
<td>74,317</td>
<td>25,314</td>
<td>24,599</td>
</tr>
<tr>
<td>1993–1996</td>
<td>Total jobs held</td>
<td>65,278</td>
<td>99,524</td>
<td>51,650</td>
<td>90,360</td>
<td>39,729</td>
<td>37,021</td>
</tr>
<tr>
<td>1994–1996</td>
<td>Total jobs held</td>
<td>60,116</td>
<td>55,894</td>
<td>48,542</td>
<td>44,626</td>
<td>42,203</td>
<td>7,654</td>
</tr>
<tr>
<td>1995–2000</td>
<td>Total jobs held</td>
<td>121,450</td>
<td>192,720</td>
<td>173,623</td>
<td>75,110</td>
<td>72,805</td>
<td>13,553</td>
</tr>
</tbody>
</table>

Job counts in the SIPP are from internal Census files using the job tracking identifier created by authors. Job counts in the DER are done using the employer tax identifier (EIN). Person counts in the SIPP are calculated using the internal longitudinal person identifier. Person counts in the DER are calculated using the SSN.
in the DER data. Details about this merge can be found in appendix B.

C. Matching SIPP and DER Jobs

After the creation of the SIPP and DER job-level data sets, the next step was to create a common sample of people who had job reports in both files. In the third column of table 1, we show the number of people found in both sources and the total jobs they held according to the SIPP and the DER. Here the timing of the survey plays an important role. In every SIPP panel, the survey asked employment questions of at least some respondents in the last few months of the year preceding the official beginning year and in only a subset of months in the final year of the panel. For DER jobs, we did not have any subannual information about the dates the job was held. In order to attempt to match as many SIPP and DER jobs as possible, all DER jobs from the years either partially or fully covered by the survey were included in the potential match set, as appropriate for each respondent. We did this to allow the best possible chance for a given SIPP job to match, feeling that we did not wish to impose the requirement that timing between the SIPP and the DER be exact. However, this has the effect of making the SIPP and DER job counts noncomparable. We report more comparable job counts in appendix table C2, where we count only SIPP and DER jobs reported in the full survey years.

After we matched by SSN, a job-to-job match was performed using probabilistic record linking based primarily on name matching. The primary basis for matching was self-reported name of the employer from the SIPP and administrative name of the employer from the Business Register. Earnings were not used in the match in order to prevent bias in the subsequent comparison of earnings. Appendix C gives the details of this match, including which additional matching variables were used, how duplicate matches were handled, and how company ownership changes affected the matching. The fourth column of table 1 shows the number of SIPP jobs that were successfully matched to a counterpart job in the DER. While the percentage of SIPP jobs that match ranges from 76% to 78% across the panels, the percentage of total person earnings represented by these matching jobs is much higher, ranging from 91% to 94%.

Of the jobs that matched, we dropped those that did not have at least one DER and one SIPP earnings report in one of the full years covered by the panel, but we did not require these reports to be in the same year. For example, a SIPP job could have earnings reports for 1996 and 1997 but not 1998, while the DER job could have reports for all three years. As a result, the SIPP and DER sample sizes were slightly different for each year. In our mixed-effects modeling, 0s were treated as missing values and were modeled as ignorable in the sense of Rubin (1976) given all effects in the fixed- and random-effects design matrices, facilitating estimation with an unbalanced panel. The decision not to require exact matching in the earnings years was based on the fact that earnings essentially reported as 0 in one source and positive in another source was a type of measurement error that we did not wish to exclude. The fifth column of table 1 shows the decrease in our sample of jobs due to missing DER or SIPP reports.

Finally, there were jobs that matched and had both SIPP and DER earnings in full survey years but the SIPP earnings were incomplete due to respondents missing an interview in the middle of the panel. When an entire household missed an interview, the Census Bureau did not impute responses for this wave of the survey, and the data were left missing. We dropped individuals who ever had a missing wave of SIPP data. In the sixth column of table 1, we show the final total number of jobs per panel that were used in the analysis. Combined across panels, our sample has 116,781 jobs, 80,792 people, and 70,081 unique employers.

In months where a SIPP respondent was interviewed but failed to answer the earnings questions, responses were imputed by the Census Bureau. Our main sample includes both reported and imputed values. In addition, we split our sample into person-job observations that never have imputed monthly earnings and those that do and estimate our model on both subsamples. This allows us to show the effect of the Census Bureau’s imputation method on reliability statistic calculations. The last column of table 1 shows people and jobs that remain when these imputations are dropped.

Tables 2 and 3 describe the covariance structure of the SIPP and DER earnings over time. Variances are shown on the diagonals, covariances are listed below the diagonal, and correlations are listed above. A job contributes an observation for any year in which it has nonzero SIPP or DER earnings or both. In the SIPP data, the correlations between adjacent years range from 0.53 to 0.76. In the DER data, they are higher, ranging from 0.80 to 0.83. For 1992 to 1994 and 1999, the variance of earnings is higher in the DER than in the SIPP, while in 1996, the SIPP has higher variance. In the remaining years (1990–1991 and 1997–1998), the variance is quite close between the two sources.

16 In earlier versions of this paper, we did not make this assumption because we wished to show the effect of these missing data on the measurement of annual earnings. We changed our sample for two reasons. First, we wanted to be more comparable to the literature, which generally drops missing data when estimating measurement error. Second, we concluded that most data users would first do an imputation of some kind for the missing months when calculating annual earnings. Since at the moment there is no standard method for doing this imputation, we decided to drop these cases from our modeling.

17 When dropping imputed values, we use the flag that indicates a monthly earnings value was imputed and the interview flag that tells when all of a respondent’s answers were imputed using a hot-deck method that assigns a donor. This latter type of imputation is called Type Z by the Census Bureau and is used when a household is interviewed but some members are not able to be interviewed.
Table 2.—Covariance/Correlation Matrix for Natural Log of SIPP Job Annual Earnings

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>2.038</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1991</td>
<td>0.983</td>
<td>2.054</td>
<td>0.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>0.833</td>
<td>1.795</td>
<td>0.56</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>0.887</td>
<td>1.808</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>0.699</td>
<td>1.060</td>
<td>1.935</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td></td>
<td></td>
<td></td>
<td>2.094</td>
<td>0.76</td>
<td>0.71</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td></td>
<td></td>
<td></td>
<td>1.074</td>
<td>1.946</td>
<td>0.76</td>
<td>0.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td>0.808</td>
<td>1.030</td>
<td>1.923</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td></td>
<td></td>
<td></td>
<td>0.659</td>
<td>0.772</td>
<td>0.961</td>
<td>1.805</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variances on diagonal; covariances below the diagonal; correlations above the diagonal. Sample is the matched SIPP/DER jobs with at least one SIPP and DER earnings report in the panel years and no missing interview waves.

Table 3.—Covariance/Correlation Matrix for Natural Log of DER Job Annual Earnings

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>1.965</td>
<td>0.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1991</td>
<td>1.209</td>
<td>2.095</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>1.266</td>
<td>2.088</td>
<td>0.80</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>1.284</td>
<td>2.195</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>1.040</td>
<td>1.276</td>
<td>2.261</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td></td>
<td></td>
<td></td>
<td>1.898</td>
<td>0.82</td>
<td>0.77</td>
<td>0.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td></td>
<td></td>
<td></td>
<td>1.147</td>
<td>1.931</td>
<td>0.83</td>
<td>0.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td>0.968</td>
<td>1.154</td>
<td>1.919</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td></td>
<td></td>
<td></td>
<td>0.882</td>
<td>1.000</td>
<td>1.186</td>
<td>1.974</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variances on diagonal; covariances below the diagonal; correlations above the diagonal. Same sample as table 2.

Table 4.—Correlation Matrix for Natural Log of SIPP/DER Job Annual Earnings

<table>
<thead>
<tr>
<th></th>
<th>Ln(SIPP Job Annual Earnings)</th>
<th>Ln(DER Job Annual Earnings)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.75</td>
<td>0.60</td>
</tr>
<tr>
<td>1991</td>
<td>0.59</td>
<td>0.76</td>
</tr>
<tr>
<td>1992</td>
<td>0.59</td>
<td>0.78</td>
</tr>
<tr>
<td>1993</td>
<td>0.62</td>
<td>0.87</td>
</tr>
<tr>
<td>1994</td>
<td>0.55</td>
<td>0.71</td>
</tr>
<tr>
<td>1996</td>
<td>0.89</td>
<td>0.73</td>
</tr>
<tr>
<td>1997</td>
<td>0.72</td>
<td>0.87</td>
</tr>
<tr>
<td>1998</td>
<td>0.67</td>
<td>0.72</td>
</tr>
<tr>
<td>1999</td>
<td>0.64</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Same sample as table 2.

Table 4 gives the cross-source correlations between each year of DER and SIPP data. For earnings in the same year, correlations range from 0.75 to 0.89, and for adjacent years, from 0.55 to 0.73. In general, correlation between SIPP and DER earnings has increased over time, with the high point occurring in 1996. Adjacent years of DER and SIPP data have lower cross-source correlations than the autocorrelation of adjacent years of DER data but are mixed when compared to the autocorrelation of adjacent years of SIPP data. For the 1996 panel, the cross-source correlations are lower in adjacent years than the SIPP-to-SIPP correlations. However, for the years 1992 to 1994, the correlation of SIPP earnings with the prior year is stronger when those earnings come from the DER instead of the SIPP.

D. Why Administrative Data Might Not Be Truth

Before comparing earnings, we discuss three reasons that considering administrative data to be truth might be problematic. First, there are some definitional differences between the two data sources. Second, there is likely to be error in the administrative data themselves. Third, the matching process between the data sources may also introduce error. We briefly discuss each of these in turn and summarize how they might affect the comparison between SIPP and DER earnings.

**Conceptual differences between SIPP and DER: Jobs and earnings definitions.** Conceptual differences between SIPP and DER stem from different definitions of earnings and jobs. There are at least two parts of earnings that would be reported on an employee’s pay stub in gross earnings that are not included in box 1 of the W-2 form: pre-tax health insurance plan premiums paid by the employee and pre-tax elective contributions made to deferred compensation arrangements such as 401(k) retirement plans. In the latter case, these contributions are reported elsewhere on the W-2 form (e.g., box 13 in 1999) and the DER file contains reports of these deferred contributions.
earnings that can be added to box 1 earnings to approximate gross earnings. While pre-tax health insurance plan premiums are reported on the W-2 form, they are not contained in the DER extract created for research use. This omission represents one important way in which administrative records may differ from survey records that is not the result of error in the survey data collection process. DER will be less than SIPP earnings if, as instructed, the respondent reported gross earnings during the survey that included health insurance premiums.

There are other possible differences between box 1 on the W-2 form and gross earnings reported in the survey. These involve an employee benefit that the employee is unlikely to consider wages and is unlikely to be reported as such on a pay stub but that the employer is required to report as taxable income. In these cases, DER earnings are likely to be higher than SIPP earnings, because respondents, again as instructed, do not report these benefits as gross earnings.

A final potential problem with DER employer reports is that EINs do not necessarily remain constant over time. This poses problems for defining an employer-employee relationship. Unlike Social Security numbers, which serve as good longitudinal identifiers for individuals, EINs can change for reasons that do not involve a person moving to a new employer. Company reorganizations through mergers, acquisitions, or spinoffs may result in a worker having two W-2 forms for a tax year, each with a different EIN, without having changed employers. In such cases, the DER earnings will be less than the SIPP earnings because a portion of the earnings for the year is missing. As part of the linking process between DER and SIPP earnings, we attempted to identify these kinds of successor-predecessor problems and merge the two DER jobs determined to be related to a single SIPP job (see appendix C for details).

To summarize, the exclusion of health insurance premiums from the DER implies DER less than SIPP; the inclusion of employee benefits in the DER implies DER is greater than SIPP; and the EIN changes due to firm reorganization imply DER is less than SIPP.

Error in the administrative data. Government agencies that collect administrative data recognize that mistakes are made in the reporting process although researchers commonly do not consider that. In the case of W-2 records, the SSA has a process for employers to file amended returns, and these are incorporated into the Master Earnings File. Data managers at SSA generally suggest that most amended returns are filed and processed within two years of the original filing. Since our MEF research extract was created in 2002, we believe that our DER data from tax years 1990 to 1999 contain most of the relevant amended filings. However, we are also confident that not all filing mistakes are recognized and corrected and that some corrections first happen many years later. While this type of error may be less common than survey reporting error, it does exist.

Another type of error in administrative data is processing error. Sometimes employers make a processing mistake when filing returns for all their employees, such as adding extra zeros to the ends of numerical amounts. There is some evidence that automated read-in processes sometimes malfunction and dollar amounts are created that are nonsensical. This type of error process is very different from the one typically postulated for self-reported data as it is unrelated to the actual amount or the person reporting. More research is needed to determine the extent of this error and quantify its specific impact.

Error in matching. Record mismatches are the final source of data error that we consider here. While much clerical review has convinced us that our probabilistic linking process produced high-quality matches, some (small, we hope), error is always introduced by this type of record matching. In our case, if the SSN for a respondent is correct, then a mismatch in the job means the earnings belonged to the same person but did not come from the same employer. If the SSN is incorrect, then the difference in SIPP and DER earnings is likely to be larger because the source person was incorrect. We believe it is unlikely that an incorrect SSN would have many job-level matches due to our use of the employer name, but it is a possibility, especially for large employers.

V. Results

A. Random Effects and Reliability Statistics

We report parameter estimates for the elements of $\Sigma_0$, $\Sigma_{t[t-1]}$, $G^{(0)}$, and $G^{(1)}$ from the estimation of our MLMM model in equation (2) in table 5. We report the variance of the SIPP and DER, the variance of the signal, the variance of the measurement error, and the reliability statistics, according to equations (10) to (18) in table 6. We present results for job-level specifications using the full sample and then separately using jobs with and without SIPP earnings imputations. We also present results for person-level specifications where the earnings are summed across all jobs, matched and non-matched. Again, we use the full sample, a subsample with only individuals with no imputed earnings, the complementary subsample with individuals with at least one month of imputed earnings, and finally the full sample with a dominant employer effect included in the random effects design matrix. For each sample, we show in table 6 calculations done with four different definitions of $\omega$: SIPP as truth $(1,0)$, SIPP and DER equally likely to be truth $(0.5,0.5)$, DER more likely than SIPP to be truth $(0.1,0.9)$, and DER as truth $(0,1)$.

---

18 These include educational assistance above a certain monetary level, business expense reimbursement above the amount treated as substantiated by the IRS, payments made by the employer to cover the employee’s share of Social Security and Medicare taxes, certain types of fringe benefits such as the use of a company car, golden parachute payments, group-term life insurance over $50,000 paid for by the employer, potentially some portion of employer contributions to medical savings accounts, nonqualified moving expenses, and, in some circumstances, sick pay from an insurance company or other third-party payer.
As shown in table 5, the person, employer, and time-period specific variance components are uniformly higher in the DER than the SIPP, which, using equation (10), means a higher conditional variance for the DER overall than in the SIPP, as shown in the third column of table 6 (Variance of Y, DER). Person effects, employer effects, and measure-specific random time effects all contribute to the greater conditional variance of the DER measure as compared to SIPP. This observation holds over all samples and subsamples and for both job- and person-level specifications.

In the model estimated with no imputed values, the variance of the person and employer effects rises for both SIPP and DER, and the difference between them remains similar. However, the variance of the time-period specific components, $\sigma_2^2$ and $\sigma_3^2$, falls for both SIPP and DER, and the gap narrows. In the model estimated with only jobs that had at least one month of imputed earnings, the DER person and employer variance components remain similar to the other two models, but the SIPP person and employer components become smaller. The variance of the time-period specific component rises for both the SIPP and DER, and the gap between them increases. This result is reflected in table 6 where the no-imputations sample has the highest variance in the SIPP and the imputations-only sample has the lowest SIPP variance. For the DER, the opposite is true: the imputations-only sample has the highest variance and the no-imputations sample has the lowest variance. Our hypothesis is that this ranking of overall levels of conditional variance is due to the Census Bureau’s survey imputation methods, specifically, that earnings non-responders are not ignorably missing given the conditioning data used in the hot-deck imputation.

The overall conditional variance for each data source remains constant regardless of our definition of truth. The variance of the signal moves between the SIPP variance and the DER variance, depending on the value of $\omega$. When $\omega = (1, 0)$—SIPP is truth—the signal equals the variance of the SIPP. When $\omega = (0, 1)$—DER is truth—the signal equals the variance of the DER. The variance of the measurement error moves accordingly and rises for each source as we place less weight on that source being the truth. Our reliability statistic as defined in equations (15) and (16) is shown in column 6 of table 6. Without additional assumptions or data about which source is truth, there is no way to choose one reliability statistic over the others. However, the range is informative in that it shows how much the SIPP statistic might change if researchers were to move away from the concept of administrative data as truth. Depending on the assumption about truth, the statistic ranges from 1 to 0.6 for the SIPP and from 1 to 0.68 for the DER in the full job sample. Not surprisingly, the range of both the SIPP and DER reliability statistics is larger for the model estimated with only jobs with imputations (1 to 0.37 for the SIPP and 1 to 0.55 for the DER) and smaller for the model using only jobs without imputations (1 to 0.68 for the SIPP and 1 to 0.73 for the DER).
The reliability statistic is predictably higher for the models estimated with person-level data and ranges from 1 to 0.7 for the SIPP and 1 to 0.74 for the DER when estimated using the whole sample. These results are consistent with those of Gottschalk and Huynh (2000), who found a reliability statistic of 0.67 for the 1996 panel, which rose to 0.73 when earnings imputations were dropped. Although we include four additional SIPP panels, when we declare the DER to be the truth, our SIPP reliability statistics of 0.70 and 0.78 for the full and no-imputation samples, respectively, are close. For the person-level specification that includes the dominant employer effect, there is very little change in the reliability statistics for either the SIPP or the DER. The only major difference is the fall in the variance of the person effects, because employer effects are rarely included at the person level. The effect of multiple job holding over the year, either simultaneous or sequential, is to reduce the total variance of earnings and to reduce the contribution of individual differences in earnings to that variance, particularly when we control for the effect of the dominant employer. Second, persistent differences in the pay policies of employers are mitigated by the structure of the error, in particular its property of mean reversion. Even when we place some probability on the SIPP being true and the DER having error, our results support this hypothesis.

An interesting feature of tables 5 and 6 is that when our model is fit at the person level, the variance of the person and employer effects is substantially less than when the model is fit at the job level. This may be an important new finding because employer effects are rarely included at the person level, and we speculate on its causes. First, there is much more variance in the SIPP and DER earnings measures at the job level than at the person level. The effect of multiple job holding over the year, either simultaneous or sequential, is to reduce the total variance of earnings and to reduce the contribution of individual differences in earnings to that variance, particularly when we control for the effect of the dominant employer. Second, persistent differences in the pay policies of employers are mitigated by changing employers but not by enough to eliminate employer effects in annual earnings from all sources.
B. Fixed Effects

We turn next to the fixed effects from our estimation of the MLMM model and present results from the job-level model estimated with all the observations in table 7. Using the definition from equation (19), in the first four columns, we present different estimates of the vector $\Sigma(B)$, calculated using the same priors for $\omega$ as we used to calculate the reliability statistics in section VA. Since equation (19) defines the truth to be the weighted average of the SIPP and DER coefficients, the first and fourth columns of table 7 [$\omega = (1, 0)$ and $(0, 1)$], correspond to coefficients from a SIPP earnings equation and a DER earnings equation, respectively. These end points define the range of each fixed effect. The SIPP and DER measurement errors are defined according to equation (20) and are reported in columns 5 to 10. When one source is
declared to be the truth, by definition the measurement error for this source is 0; hence, each source has only three columns of measurement error reported. Standard errors are reported for both the true coefficients and the measurement error. Measurement error significantly different from 0 is equivalent to stating that there are significant differences between the SIPP and DER estimates of a particular coefficient. Negative measurement error for either source means that the source coefficient was smaller than the weighted average coefficient. By definition, the DER and SIPP measurement errors will have opposite signs.

We remind readers that in mixed-effects modeling, the fixed effects are the coefficients on the observed characteristics of the individuals. These effects may vary over time and differ from the random effects because the effect is estimated directly instead of being inferred from its distribution. The fixed effects included in our model are education in five levels (no high school diploma (excluded case); high school diploma; some college; college degree; graduate degree); a piecewise linear spline in labor force experience with bend points at 2, 5, 10, and 25 years of experience; an overall intercept; SIPP panel effects (excluded case is 1996); and a linear time trend. The education, experience, and overall intercept coefficients are all interacted with race and gender to produce separate estimates for white males, nonwhite males, white females, and nonwhite females.

The only fixed effects that are significantly different between the SIPP and the DER are the college and graduate degree indicators for white males and the panel effects. For these two education effects, SIPP measurement error estimates range from $-0.018$ to $-0.037$, meaning that the SIPP estimates of the returns to a college degree are approximately 2 to 4 log points lower than in the DER. Estimates for the return to a graduate degree vary by approximately 1.5 to 3 log points, with the SIPP returns again being lower. The panel effects represent average differences in annual earnings for each SIPP panel relative to the 1996 panel. These effects are negative in both the SIPP and DER, meaning that 1996 earnings are higher on average. SIPP measurement error ranges from $-0.04$ to $-0.1$, which means that the DER panel effects are less negative, that is, for DER earnings, the differences between panels are lower. Here, however, we caution against assigning too much importance to this result. The 1996 panel was three waves longer than the longest panel of the early 1990s and suffered much more from problems with individuals missing waves. Thus, as shown in table 1, our 1996 sample size becomes very small when we drop individuals with missing waves and likely leaves us with a group of respondents with different characteristics from those in the panels from the early 1990s. Thus, comparisons across panels are difficult to make. All the other coefficients in table 7 are very similar between the SIPP and the DER with measurement error usually less than 1 percentage point and not significantly different from 0.

In table 8, we show the effect on earnings of 2, 5, 10, 25, and 30 years of experience, calculated using the five coefficients from the piece-wise linear spline in the main job-level model. These effects are split by demographic group and, as with the coefficients in table 7, we report the true effect based on four different priors for $\omega$ and also the SIPP and DER measurement error. Each effect is followed by its standard error. For both white males and females, there are significant differences between the SIPP and DER experience effects at 2 and 5 years, with the SIPP effect being larger by 5 to 15 log points. At 10 and 25 years, there are no significant differences between the SIPP and DER effects. For white men at 30 years, the DER effect is 2 to 4 log points significantly larger, but for white women, the SIPP and DER effects are not significantly different at 30 years. For nonwhite males and females, the differences between the SIPP and the DER are significant only at 2 years. For nonwhite men, the measurement error effects are relatively large at 5 years and 30 years, ranging from 5 to 10 log points, but these results are imprecisely estimated and are not significant. Nonwhite females have similarly large standard errors and hence no significant effects after 2 years, but even the magnitude of the effects stays small after 5 years. In this sense, the profiles of white and nonwhite women are similar to each other. The SIPP effect is initially higher, and then the SIPP and DER converge and are quite similar at 30 years. For both white and nonwhite men, the SIPP effect is initially larger and then converges to the DER effect, and then the DER effect becomes larger, although only for white males is this pattern significant.

These results change somewhat when imputations are dropped or when the analysis is done at the person level. In particular, returns for women stay higher in the SIPP relative to the DER across the whole experience profile. We refer interested readers to appendix D, where we discuss the full set of comparable results for the job-level model estimated using jobs without imputed earnings and jobs with at least one month of imputed earnings and for the person-level model using all three samples and the dominant employer specification. This appendix also contains graphical summaries of the experience effects.

VI. Conclusion

We used linked survey data from the Census Bureau’s SIPP and administrative data from the SSA’s DER, matched at both the job and person levels, to estimate and analyze measurement error models based on a multivariate linear mixed-effects model for the pair of SIPP and DER annual earnings outcomes. We showed that linking survey and administrative data at the job level is a substantially more complicated and nuanced process than linking the same data at the individual level. The potential for measurement error due to mismatching is substantial, and we documented the steps taken to control that error. In the statistical specification, we find that the conditional variance of the DER measures, given the factors in both the fixed- and random-effect design matrices, is greater than that of the SIPP component by component for the person, employer, and time effects. There
<table>
<thead>
<tr>
<th>Experience</th>
<th>Weight(SIPP,DER)</th>
<th>Male White</th>
<th>Female White</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 years</td>
<td>1.154 1.079 1.019 1.004</td>
<td>2.009 2.172 2.174 2.174</td>
<td>2.406 2.388 2.373 2.370</td>
</tr>
<tr>
<td>5 years</td>
<td>0.036 0.035 0.037 0.038</td>
<td>0.033 0.033 0.035 0.036</td>
<td>0.031 0.031 0.033 0.034</td>
</tr>
<tr>
<td>10 years</td>
<td>0.000 0.075 0.135 0.150</td>
<td>0.000 0.047 0.085 0.095</td>
<td>0.000 0.001 0.032 0.036</td>
</tr>
<tr>
<td>25 years</td>
<td>0.000 0.012 0.021 0.023</td>
<td>0.000 0.010 0.018 0.020</td>
<td>0.000 0.010 0.017 0.019</td>
</tr>
<tr>
<td>30 years</td>
<td>0.000 0.001 0.002 0.000</td>
<td>0.020 0.010 0.002 0.000</td>
<td>0.019 0.010 0.002 0.000</td>
</tr>
<tr>
<td>Effect</td>
<td>-0.150 -0.075 -0.015 0.000</td>
<td>-0.095 -0.047 -0.009 0.000</td>
<td>-0.036 -0.018 -0.004 0.000</td>
</tr>
<tr>
<td>SE</td>
<td>0.023 0.012 0.002 0.000</td>
<td>0.020 0.010 0.002 0.000</td>
<td>0.019 0.010 0.002 0.000</td>
</tr>
</tbody>
</table>

**Effect on earnings of 2, 5, 10, 25, and 30 years of experience. Calculated using five coefficients from piece-wise linear spline (0-2, 2-5, 5-10, 10-25, 25+) in job-level mixed effects model, full sample. The asterisks refer to the SIPP ME effects: Significant at *5%, **1%, and ***0.1%.**
is more variability in the DER job- and person-level data, even controlling for demography, education, and labor force experience.

In our model, neither the SIPP nor the DER measure was treated as “true.” Instead, we specified a prior weight vector that was used to define “truth” as a weighted average of SIPP and DER. Such a specification allowed us systematically to consider the implications of errors in either measure on the resulting conclusions about conditional means (fixed effects) and variance components (random effects). Considering the random components of the error process, we found that the reliability statistics for SIPP and DER earnings measures were quite comparable except for the subsample of SIPP person-jobs, where at least one year of SIPP earnings contained a Census Bureau imputation. These measures were less reliable than the DER. For the fixed effects, we found very little statistically meaningful measurement error, with most of the error being found in the highly educated white male groups and the early-career experience profiles of male and female whites.

Overall, our results point to the need to allow for measurement error in both the survey and administrative data when doing validation studies. However, there are certain situations, particularly when the SIPP measure is based on partially imputed data, where we find strong evidence that the administrative measure contains less error. An important next step is to combine our modeling procedure with an audit study that determines the correct value of the earnings measure as a function of variables that are measured for all cases. Results from such work could be used by statistical agencies to produce a measure of “true earnings” that is a hybrid of survey and administrative data, a valuable measure for researchers that would allow agencies to release information from administrative data while limiting confidentiality concerns.

REFERENCES


