Bias From Classical and Other Forms of Measurement Error

Dean R. HYSLOP
Department of Economics, University of California at Los Angeles, Los Angeles, CA 90095 (*dhyslop@pana.sscuet.ucla.edu*)

Guido W. IMBENS
Department of Economics, University of California at Los Angeles, Los Angeles, CA 90095;
Department of Economics, University of California, Berkeley, CA 94720;
National Bureau of Economic Research, Cambridge, MA

We consider the implications of an alternative to the classical measurement-error model, in which the observed, mismeasured data are optimal predictions of the true values, given some information set. In this model, any measurement error is uncorrelated with the reported value and, by necessity, correlated with the true value of interest. In a regression model, such measurement error in the regressor does not lead to bias, whereas measurement error in the dependent variable leads to bias toward 0. In general, the measurement-error model, together with the information set, is critical for determining the bias in econometric estimates.

KEY WORDS: Classical measurement error; Optimal prediction error; Regression analysis.

Many variables used in econometric analyses are recorded with error. These errors may have occurred at various stages of the data collection. They may be the result of misreporting by subjects, miscoding by the collectors of the data, or incorrect transformation from initial reports into a form ready for analysis. Often such errors are ignored. In cases in which explicit attention is paid to measurement error, it is typically assumed to be "classical measurement error," in which the error is independent, or at least uncorrelated with, the true value of the underlying variable (e.g., Klepper and Leamer 1984; Li and Vuong 1998; Chesher and Schluter 1999; Schennach 1999; see Grilliches 1987; Angrist and Krueger 2000; and Bound, Brown, and Mathiowetz in press for surveys). However, when responses have been validated (Bound and Krueger 1991; Pischke 1995; Card and Hyslop 1997) empirical support for classical measurement error (CME) has typically been limited.

The implications of deviations from CME are only rarely considered. Card (1996) and Bollinger (1996) studied models with measurement error in binary variables in which the CME assumptions cannot hold. Kane, Rouse, and Staiger (1999) investigated categorical response models exploiting the presence of two measures with uncorrelated errors. They did allow the errors to be correlated with both the true and reported values. Horowitz and Manski (1995) studied bounds when a fraction of the observations are mismeasured in an unrestricted manner. Bound et al. (in press) surveyed some of these approaches.

In this article, we explore the consequences of a class of alternative models of measurement error. We argue that, if errors occur in reports by agents based on limited information, there are specific alternatives to the CME model based on the view that respondents are actively choosing a best response in the presence of limited information. Such models have been considered before in settings where explicit account was taken of the agent's awareness of the limits on his/her knowledge and incentives for accurate reporting. Examples include the modeling of preliminary reports of macroeconomic aggregates (Mankiw and Shapiro 1986), in the analysis of the effect of financial incentives on accuracy in surveys (Philipson in press), and the analysis of responses to questions about future events (Manski 1990; Das, Dominitz, and Van Soest 1999). The alternatives we consider, like some of the models analyzed by Berkson (1950) and Durbin (1954) assume that, in response to the question "What is the value of X?" respondents report their best estimate of this quantity given their information set. In contrast, under the CME model, respondents can be viewed as reporting an unbiased estimate with higher expected mean squared error.

We explore the implications of these alternative models in the context of linear regression models. We find that the standard argument that measurement error in regressors leads to underestimation of the magnitude of the relationship between the true variables depends critically on the measurement-error model. In particular, under alternative assumptions, measurement error can lead to over– as well as underestimation of the underlying coefficients. We derive signs of the bias for a number of leading cases. Finally, we present some calculations showing how sensitive regression estimates can be to measurement error under different models in the context of wage regressions, using the Panel Survey of Income Dynamics (PSID) validation study (e.g., Bound and Krueger 1991; Pischke 1995) to obtain estimates of the reliability of reported earnings and the Ashenfelter and Krueger (1994) twins study for estimates of the reliability of reported years of education. Although in many cases one may not be able to credibly choose between the different types of measurement error, one may be able to assess the amount of measurement error in each variable using previously collected data from validation studies. In such cases one can explore the range of parameter

Journal of Business & Economic Statistics
October 2001, Vol. 19, No. 4

values consistent with the amount of measurement error under the various models, as we illustrate in Section 4. These analyses are in the spirit of the sensitivity and bounds analyses of Rosenbaum and Rubin (1983), Leamer (1987), Horowitz and Manski (1995), and Bollinger (1996).

1. A DECOMPOSITION OF MEASUREMENT ERROR

Let X^* denote the true value of a variable of interest and X the recorded value. The measurement error is the difference between the recorded and true value:

$$\varepsilon \equiv X - X^*. \tag{1}$$

We decompose ε into three components: $\varepsilon = \varepsilon_{\text{CME}} + \varepsilon_{\text{OPE}} + \varepsilon_r$. The first component is not predictable by the true value and is therefore CME:

$$\begin{aligned}\varepsilon_{\text{CME}} \equiv \varepsilon - E[\varepsilon|X^*] &= X - X^* - E[X - X^*|X^*] \\ &= X - E[X|X^*]. \end{aligned} \tag{2}$$

The second component is not predictable by the reported value, which we refer to as optimal prediction error (OPE):

$$\begin{aligned}\varepsilon_{\text{OPE}} \equiv \varepsilon - E[\varepsilon|X] &= X - X^* - E[X - X^*|X] \\ &= -X^* + E[X^*|X]. \end{aligned} \tag{3}$$

The third and final component, ε_r, is defined as the remainder of the error:

$$\begin{aligned}\varepsilon_r &\equiv \varepsilon - \varepsilon_{\text{CME}} - \varepsilon_{\text{OPE}} \\ &= X - X^* - (X - E[X|X^*]) - (-X^* + E[X^*|X]) \\ &= E[X|X^*] - E[X^*|X]. \end{aligned} \tag{4}$$

This decomposition is definitional in that it does not require any assumptions (beyond finiteness of the appropriate expectations). It is unique, and any assumptions on the measurement error can therefore be formulated as assumptions on the three components.

2. TWO MODELS FOR MEASUREMENT ERROR

2.1 Classical Measurement Error

The standard classical measurement error (CME) model assumes that the measurement error is independent of the true value. Assuming that the measurement error has mean 0, this implies $E[\varepsilon|X^*] = 0$. Since by definition $\varepsilon_{\text{CME}} = \varepsilon - E[\varepsilon|X^*]$, it follows that, for this model to be correct, it must be that $\varepsilon = \varepsilon_{\text{CME}}$ and the last two components ε_{OPE} and ε_r sum to 0. This model is typically defended by reference to physical measurement models in which often passive recording of measurements based on imprecise measuring instruments takes place.

2.2 Optimal Prediction Error

The alternative model, which we refer to as the optimal prediction error (OPE) model, is based on the assumption that the measurement error is independent of the reported value. This implies that $E[\varepsilon|X] = 0$ and, since $\varepsilon_{\text{OPE}} \equiv \varepsilon - E[\varepsilon|X]$, $\varepsilon = \varepsilon_{\text{OPE}}$ so that $\varepsilon_{\text{CME}} + \varepsilon_r = 0$. Berkson (1950) and Durbin (1954) discussed such measurement error in regressors in the context of a regression model.

An argument in support of this model views the agent reporting the data as fully aware of the lack of precision of the measuring instrument. Suppose the agent is asked to provide the value of some variable. The agent has no way of ascertaining the true value X^* of this variable but has available a flawed or noisy measure, $\widetilde{X} = X^* + \eta_X$, with the measurement error η_X independent of the true value of the variable, exactly as in the CME model. However, suppose that the agent is aware of the lack of precision of the measurement and corrects for this by reporting the best estimate of the underlying true value X^* based on this measurement $\widetilde{X}$. To operationalize this, we interpret "best" in terms of a quadratic loss function, which implies that the agent would report the expected value of the true value given the agent's information set. [An alternative would be to assume absolute value loss, in which case the agent would report the median of X^* given the information set. For most of the following illustrative calculations, the mean and median will give the same answers because we assume normality. A third possibility arises when there is a constant loss if the reported value differs from the true value, in which case respondents should report the mode of the distribution. This arises in Philipson's (in press) survey of physicians who, with some probability, get a reward if their response matches administrative records.] Under this interpretation the error, $\varepsilon = X - X^*$, should have mean 0 given the information set of the agent. Since the reported value is clearly in the information set, this implies that the error has mean 0 given the reported value.

Critical in this model is the active role of the respondent. Thus, to assess the impact of measurement error, the researcher needs to understand how the respondent views the survey question. If the respondent is aware of not having exact information regarding the value of the variable requested, presumably the question "What is the value of X?" is interpreted as "What is your best estimate of the value of X?" In that case the answer should *not* be the unbiased measurement even if that is the basic piece of information available to the respondent. Although the respondent need not have the exact probability model underlying the unbiased measurement and true value, it is plausible that outliers are adjusted in a way that leads to some correlation between the true value and the measurement error. On the other hand, this model is less likely to be appropriate if the measurement error is the result of miscoding of survey answers.

A crucial ingredient in the OPE model is the information set. It may be that the respondent has only a single unbiased measurement of the underlying true variable. Alternatively, other variables, which themselves may enter the econometric model of interest, may be used to produce this estimate. [For example, Ashenfelter and Krueger (1994) surveyed twins

and asked each sibling to report both his/her own education and his/her sibling's education. To the extent that a respondent is not fully aware of his/her sibling's education level but has knowledge of related items, such as occupation, it is plausible that such information would be used to infer the education level.] In the next section we consider, in the context of a linear regression model, two variations of the model that differ in the information set exploited in the calculation of the optimal prediction of the quantity of interest.

Models similar to this OPE model have been used in other contexts in which agents are asked to provide information about variables whose values they do not know exactly. The behavior of government agencies reporting macroeconomic quantities can be viewed as predicting the underlying variable of interest given the agency's information set, which includes signals of the true value. In this context, the measurement error is expected to be independent of the information set used, which again necessarily includes the reported value. For example, Mankiw and Shapiro (1986) modeled the revision in gross national product between the preliminary and final reports and found that the revisions are uncorrelated with the early reports. In addition, the revisions are correlated with the final reports which, if the final reports are assumed to be the truth, is consistent with an OPE model but not the CME model.

Philipson (in press) carried out experiments to see how the reliability of survey responses varies with incentives. Philipson asked physicians the value of a categorical variable (medical specialization) that can be verified from administrative data. He offered some of the physicians an incentive scheme in which with some probability he would check the value of the variable and pay a sum of money if the reported and true values agreed. He found that such incentive schemes increased the probability that the respondents answered the question correctly.

Manski (1990) and Das et al. (1999) analyzed data in which individuals were asked about future events including future income. In that case, individuals clearly cannot know the exact value of these variables, and Manski and Das et al. modeled the qualitative responses as the best predictions (modes) given current information.

Each of these examples suggests that economic agents responding to questions about uncertain quantities can sometimes usefully be modeled as solving a prediction problem rather than as passively reporting noisy measurements. We therefore investigate the implications of this model for the estimation of regression coefficients in a linear model.

3. IMPLICATIONS OF MEASUREMENT ERROR IN THE LINEAR REGRESSION MODEL

Let us consider a homoscedastic linear regression model for two scalar variables Y^* and X^*:

$$Y^* = \alpha + \beta \cdot X^* + \nu, \tag{5}$$

where $\nu \perp X^*$, $E[\nu|X^*] = 0$, $\text{var}(\nu|X^*) = \sigma_\nu^2$, and the parameter of interest, β, is the ratio of the covariance of Y^* and X^* over the variance of X^*. Possibly mismeasured values Y and X are recorded. We consider the implications for least squares estimates of β based on a random sample from (Y, X) of various properties of the measurement error.

The basic piece of information available to the respondent is assumed to be a pair of noisy measures of the underlying variables: $\widetilde{X} = X^* + \eta_X$, $\widetilde{Y} = Y^* + \eta_Y$. We assume that the basic measurement errors (η_X, η_Y) are independent of the true value of the regressor X^* and of ν, but potentially correlated with each other. For expositional reasons, we also assume joint normality:

$$\begin{pmatrix} X^* \\ \nu \\ \eta_X \\ \eta_Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & 0 & 0 & 0 \\ 0 & \sigma_\nu^2 & 0 & 0 \\ 0 & 0 & \sigma_{\eta_X}^2 & \rho_{\eta_X\eta_Y}\sigma_{\eta_X}\sigma_{\eta_Y} \\ 0 & 0 & \rho_{\eta_X\eta_Y}\sigma_{\eta_X}\sigma_{\eta_Y} & \sigma_{\eta_Y}^2 \end{pmatrix} \right). \tag{6}$$

We consider three cases relating the basic measurements $\widetilde{X}$ and $\widetilde{Y}$ to the reported values X and Y. The first case is the CME model in which the reported value is identical to the unbiased measurement. In addition we consider two versions of the OPE model. The first [OPE(1)] is one in which the respondent reports his/her best estimate based only on the noisy measure of the mismeasured variable itself. The second [OPE(2)] is one in which the respondent reports his/her best estimate based on the noisy measures of both variables. Table 1 summarizes the three models. In each of the three cases we consider the sign of the difference between the probability limit of the least squares estimator $(\hat{\beta})$ using the noisy measures (Y, X) and the limit of the least squares estimator (β^*) using the true values (Y^*, X^*): $\text{sign}(\hat{\beta} - \beta^*)$.

3.1 Measurement Error in the Regressor

First we consider the case with $\sigma_{\eta_Y}^2 = 0$, where the measurement error is confined to the regressor. Thus $Y = \widetilde{Y} = Y^*$ under all three models. First we briefly review the CME case. The reported value is $X_{\text{CME}} = \widetilde{X} = X^* + \eta_X$. The least squares estimator therefore underestimates the coefficient in the regression with the true values:

$$\beta_{\text{CME}} = \frac{\text{cov}(Y^*, X_{\text{CME}})}{\text{var}(X_{\text{CME}})} = \beta \cdot \frac{\sigma_X^2}{\sigma_X^2 + \sigma_{\eta_X}^2},$$

which is less than β in absolute value. This is the standard case of CME leading to a bias toward 0.

Next consider the OPE(1) case. This is the case studied by Berkson (1950) and Durbin (1954). The reported value X

Table 1. Three Models for Measurement Error

Reporting model	*X*	*Y*
Classical measurement error	$X_{\text{CME}} = \widetilde{X}$	$Y_{\text{CME}} = \widetilde{Y}$
Optimal prediction error (1)	$X_{\text{OPE(1)}} = E[X^*\|\widetilde{X}]$	$Y_{\text{OPE(1)}} = E[Y^*\|\widetilde{Y}]$
Optimal prediction error (2)	$X_{\text{OPE(2)}} = E[X^*\|\widetilde{X}, \widetilde{Y}]$	$Y_{\text{OPE(2)}} = E[Y^*\|\widetilde{X}, \widetilde{Y}]$

is linear in the unbiased measurement $\widetilde{X}$ with coefficient $(1/\sigma^2_{\eta_X})/(1/\sigma^2_X + 1/\sigma^2_{\eta_X})$:

$$X_{\mathrm{OPE}(1)} = E[X^*|\widetilde{X}] = E[X^*|X^* + \eta_X] = \mu_X \cdot \frac{1/\sigma^2_X}{1/\sigma^2_X + 1/\sigma^2_{\eta_X}} + \widetilde{X} \cdot \frac{1/\sigma^2_{\eta_X}}{1/\sigma^2_X + 1/\sigma^2_{\eta_X}}.$$

To see the bias from this model, consider the regression function $Y^* = \alpha + \beta \cdot X^* + \nu = \alpha + \beta \cdot X_{\mathrm{OPE}(1)} + \tilde{\nu}$, with the composite error terms $\tilde{\nu}$ equal to $\tilde{\nu} = \nu + \beta \cdot (X^* - X_{\mathrm{OPE}(1)})$. Since by assumption in the OPE(1) model the reporting error $X_{\mathrm{OPE}(1)} - X^*$ is independent of the reported value $X_{\mathrm{OPE}(1)}$, the composite error term $\tilde{\nu}$ is independent of $X_{\mathrm{OPE}(1)}$ and there is no bias resulting from the measurement error, or $\beta_{\mathrm{OPE}(1)} = \beta$.

Finally, consider the case in which the respondent adjusts the report to take into account not just the unbiased measurement $\widetilde{X}$ but also the (accurately measured) outcome Y^*: $X_{\mathrm{OPE}(2)s} = E[X^*|\widetilde{X}, \widetilde{Y}] = E[X^*|\widetilde{X}, Y^*]$. This can be interpreted as estimating X^* based on two noisy measurements, $\widetilde{X} = X^* + \eta_X$ and $(Y^* - \alpha)/\beta = X^* + \nu/\beta$, with uncorrelated errors η_X and ν/β. The resulting reported value is therefore a weighted average of the population mean μ and the two unbiased measurements:

$$\begin{aligned} X_{\mathrm{OPE}(2)} &= \frac{1/\sigma^2_X}{1/\sigma^2_X + 1/\sigma^2_{\eta_X} + \beta^2/\sigma^2_\nu} \cdot \mu_X \\ &\quad + \frac{1/\sigma^2_{\eta_X}}{1/\sigma^2_X + 1/\sigma^2_{\eta_X} + \beta^2/\sigma^2_\nu} \cdot \widetilde{X} \\ &\quad + \frac{\beta^2/\sigma^2_\nu}{1/\sigma^2_X + 1/\sigma^2_{\eta_X} + \beta^2/\sigma^2_\nu} \cdot \frac{Y^* - \alpha}{\beta} \\ &= \lambda_1 \cdot \mu_X + \lambda_2 \cdot \widetilde{X} + \lambda_3 \cdot \frac{Y^* - \alpha}{\beta}, \end{aligned}$$

with all $\lambda_j \geq 0$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. We can rewrite this as a linear function of the true value and independent disturbances:

$$X_{\mathrm{OPE}(2)} = \lambda_1 \cdot \mu_X + (\lambda_2 + \lambda_3) \cdot X^* + \lambda_2 \cdot \eta_X + \lambda_3 \cdot \frac{\nu}{\beta}.$$

Simple but tedious calculations show that the probability limit of the least squares estimator is equal to

$$\beta_{\mathrm{OPE}(2)} = \beta \cdot \left(1 + \frac{1/\sigma^2_X + 1/\sigma^2_{\eta_X} + \beta^2/\sigma^2_\nu}{\beta^2/\sigma^2_\nu + 1/\sigma^2_{\eta_X} + \sigma^2_X \cdot (1/\sigma^2_{\eta_X} + \beta^2/\sigma^2_\nu)^2}\right),$$

which is greater than β in absolute value. In this case the least squares estimator overestimates the magnitude of the regression coefficient, due to the correlation between the reported value and the disturbance ν in the regression, which is induced by the use of Y^* in producing the best estimate of the regressor X^*.

3.2 Measurement Error in the Outcome Variable

In this subsection we consider measurement error in the outcome variable and assume that the regressor is accurately measured: $\sigma^2_{\eta_X} = 0$, and thus $X = \widetilde{X} = X^*$. Under the CME assumption, we can write the regression model as $Y = \widetilde{Y} = Y^* + \eta_Y = \alpha + \beta \cdot X + \nu + \eta_Y$. By assumption, both components of the composite error term $\nu + \eta_Y$ are independent of X, so there is no bias, and $\beta_{\mathrm{CME}} = \beta$.

Next, consider the case in which the agent reports $Y = E[Y^*|\widetilde{Y}]$. The unconditional mean of Y^* is $\alpha + \beta \cdot \mu_X$, with variance $\beta^2 \cdot \sigma^2_X + \sigma^2_\nu$, so the best estimate of Y^*, based on $\widetilde{Y}$, is

$$Y_{\mathrm{OPE}(1)} = (\alpha + \beta \cdot \mu_X) \cdot \frac{1/(\beta^2 \cdot \sigma^2_X + \sigma^2_\nu)}{1/(\beta^2 \cdot \sigma^2_X + \sigma^2_\nu) + 1/\sigma^2_{\eta_Y}} + \widetilde{Y} \cdot \frac{1/\sigma^2_{\eta_Y}}{1/(\beta^2 \cdot \sigma^2_X + \sigma^2_\nu) + 1/\sigma^2_{\eta_Y}}.$$

The slope coefficient in a regression of $\widetilde{Y}$ on X^* is β, so the slope coefficient in a regression of $Y_{\mathrm{OPE}(1)}$ on X^* is

$$\beta_{\mathrm{OPE}(1)} = \beta \cdot \frac{1/\sigma^2_{\eta_Y}}{1/(\beta^2 \cdot \sigma^2_X + \sigma^2_\nu) + 1/\sigma^2_{\eta_Y}},$$

which means $\beta_{\mathrm{OPE}(1)}$ is biased toward 0.

Finally, consider the case in which the respondent reports the best estimate of Y given $\widetilde{Y}$ and X^*. Based on X^* alone, the best estimate of Y^* would be $\alpha + \beta \cdot X^*$. Knowledge of both X^* and $\widetilde{Y}$ can be interpreted as knowledge of both $\alpha + \beta \cdot X^*$ and $\widetilde{Y} - \alpha - \beta \cdot X^* = \eta_Y + \nu$. Hence we can write

$$\begin{aligned} Y_{\mathrm{OPE}(2)} &= E[Y|\widetilde{Y}, X^*] = \alpha + \beta \cdot X^* + E[\nu|X^*, \widetilde{Y}] \\ &= \alpha + \beta \cdot X^* + E[\nu|X^*, \eta_Y + \nu] \\ &= \alpha + \beta \cdot X^* + E[\nu|\eta_Y + \nu], \\ &= \alpha + \beta \cdot X^* + (\eta_Y + \nu) \cdot \frac{1/\sigma^2_\nu}{1/\sigma^2_\nu + 1/\sigma^2_{\eta_Y}}. \end{aligned}$$

Because ν and η_Y are independent of X^*, again there is no bias from regressing $Y_{\mathrm{OPE}(2)}$ on X^*.

3.3 Measurement Error in Both the Regressor and Outcome Variable

In this subsection we consider the case in which both regressor and outcome are measured with error. In each case, the individual reporting the variables has available an unbiased measurement, $\widetilde{X} = X^* + \eta_X$, $\widetilde{Y} = Y^* + \eta_Y$, with possibly correlated errors,

$$\begin{pmatrix} \eta_X \\ \eta_Y \end{pmatrix} \Big| X^*, \nu \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\eta_X} & \rho_{\eta_X \eta_Y} \sigma_{\eta_X} \sigma_{\eta_X} \\ \rho_{\eta_X \eta_Y} \sigma_{\eta_X} \sigma_{\eta_X} & \sigma^2_{\eta_Y} \end{pmatrix}\right).$$

We look at the bias resulting from the three models considered before, CME, OPE(1), and OPE(2). In general, with the errors in $\widetilde{X}$ and $\widetilde{Y}$, η_X and η_Y, respectively, potentially correlated, the biases from measurement error cannot be signed. If the correlation between the measurement errors is 0, the direction of the bias follows intuitively from the previous calculations. These results, combined with those of Sections 3.1 and 3.2, are reported in Table 2. If the correlation between η_X and η_Y is close enough to 1, the bias will always be upward, and if it is close enough to negative 1, the bias will always be downward. To see how big these effects can be, we report in Section 4 some numerical calculations, based on numbers relevant for wage regressions.

Table 2. Measurement-Error Bias in Slope Coefficient

Source of measurement error	σ_{η_X}	σ_{η_Y}	Reporting model: CME $X=\widetilde{X}$ $Y=\widetilde{Y}$	OPE(1) $X=E[X^*\|\widetilde{X}]$ $Y=E[Y^*\|\widetilde{Y}]$	OPE(2) $X=E[X^*\|\widetilde{X},\widetilde{Y}]$ $Y=E[Y^*\|\widetilde{X},\widetilde{Y}]$
No error	0	0	No bias	No bias	No bias
Error in regressor only	>0	0	Toward 0	No bias	Away from 0
Error in outcome only	0	>0	No bias	Toward 0	No bias
Error in both (zero correlation)	>0	>0	Toward 0	Toward 0	Away from 0

3.4 Instrumental-Variables Estimation

One standard approach to dealing with CME is to use instrumental-variables methods [see the Bound et al. (in press) survey for a general discussion]. Here we explore what instrumental-variables methods do when the measurement error is of the OPE variety. We maintain the preceding linear model structure and assume that Y^* is observed without error, but X^* is measured with error and two noisy measures are available. We also assume that the two reports are optimal predictions based on unbiased and independent measurements: $X_1 = E[X^*|\widetilde{X}_1]$, $X_2 = E[X^*|\widetilde{X}_2]$, where $\widetilde{X}_1 = X^* + \eta_1$, $\widetilde{X}_2 = X^* + \eta_2$, $(\eta_1, \eta_2) \perp X^*$, and $\eta_1 \perp \eta_2$. From Section 3.1, we know that regressing Y^* on X_1 (or X_2) leads to unbiased estimates of β because the measurement error is uncorrelated with the reported value. If instead we use the second measure as an instrument for the first one, we estimate β as

$$\hat{\beta}_{iv} = \frac{\text{cov}(Y^*, X_2)}{\text{cov}(X_1, X_2)} = \beta \cdot \frac{\sigma_{\eta_2}^2 + \sigma_X^2}{\sigma_X^2}.$$

Instrumenting for the mismeasured regressor now leads to a bias away from 0, proportional to the inverse of the reliability ratio of the noisy measure. Note that in this case, as before, the data are not informative about the nature of the measurement error. The finding that, as in the Ashenfelter and Krueger (1994) study, instrumenting leads to considerably higher estimates than ordinary least squares estimates is consistent with the CME story as well as with the OPE model. The interpretation of the results is very different, however, under the two models.

4. MEASUREMENT ERROR IN WAGE REGRESSIONS

In this section we consider the regression of the logarithm of wages on education when both may be measured with error and the interest is in the coefficients from the regression based on the true values. (In some cases one can argue that interest should be in the regression on perceived values. For example, if individuals do not know their own income with certainty, one may argue that their estimated income is more relevant for consumption decisions than true income. Here we would argue that in answering a survey an individual may have insufficient incentive to carefully check his or her records and that, if the value of the variable is needed for making economically meaningful decisions, one might acquire the relevant information.) We calculate some of the moments of hourly wages and education levels from National Longitudinal Survey of Youth (NLSY) data. (See Hellerstein and Imbens 1999 for details of the sample used.) The earnings measure used is the logarithm of the usual weekly wage, and the education measure is years of completed schooling. The estimated regression function based on these data is

$$\widehat{Y}_i = \underset{(0.09)}{5.16} + \underset{(0.006)}{0.061}\ X_i.$$

The standard deviations of the log wage is $\sigma_Y = 0.43$, and the standard deviation of the education level is $\sigma_X = 2.2$.

To find appropriate numbers for the measurement-error variances, we turn to some of the validation studies. For the measurement error in the education level, we take our numbers from the Ashenfelter and Krueger (1994) study. Ashenfelter and Krueger asked twins about their own education as well as their twin sibling's level of education. Using those data, they estimated a reliability ratio of approximately 90%, implying that the variance of the measurement error is approximately 10% of the variance of education. We therefore use $\sigma_{\eta_X} = \sqrt{0.1} \times \sigma_X = 0.63$. For log wages we take our numbers from Bound and Krueger (1991) and Pischke (1995), who analyzed the validation study of the PSID. Their numbers suggest a reliability ratio of 75%, and hence $\sigma_{\eta_Y} = \sqrt{0.25} \times \sigma_Y = 0.3$. Although these validation studies are obviously different from the NLSY in the way individuals were selected and in the formulation of the questions and the estimates are all based on the CME assumption, they may be informative about the relative amount of measurement error for the earnings and education measures.

Based on these error variances and the distribution of the observed variables, we calculate the true parameter values, β^*, and percentage bias, $(\hat{\beta} - \beta^*)/\hat{\beta} \times 100\%$, under different measurement-error scenarios. Table 3 summarizes the results. The results in the first three rows, with measurement error in, at most, one variable, reflect the qualitative results in Sections 3.1 and 3.2. For example, in the second row, with only measurement error in the regressor, comparing the estimated parameter of 0.061 with the true parameter value of 0.069 implies that the estimated value is biased downward by 12%. The largest bias in these three rows is on the order of 27%. When both variables are measured with error and with the errors correlated, the bias can get much larger. With zero correlation, the bias for the CME model is 12%. Allowing the correlation between measurement errors to go to -0.90, the bias goes to 76%, and with the correlation up to 0.90, the bias goes to 50%. Similarly for the other reporting models the bias

Table 3. True Returns to Education in the Presence of Measurement Error

				Reporting model		
				CME	OPE(1)	OPE(2)
Source of measurement error	σ_{η_X}	σ_{η_Y}	$\rho_{\eta_X \eta_Y}$	$X = \widetilde{X}$ $Y = \widetilde{Y}$	$X = E[X^*\|\widetilde{X}]$ $Y = E[Y^*\|\widetilde{Y}]$	$X = E[X^*\|\widetilde{X}, \widetilde{Y}]$ $Y = E[Y^*\|\widetilde{X}, \widetilde{Y}]$
No error	0.00	0.00	—	0.061 (0%)	0.061 (0%)	0.061 (0%)
Error in regressor	0.63	0.00	—	0.069 (−12%)	0.061 (0%)	0.055 (9%)
Error in outcome	0.00	0.30	—	0.061 (0%)	0.077 (−27%)	0.061 (0%)
Error in both	0.63	0.30	−0.90	0.108 (−76%)	0.109 (−77%)	0.076 (−24%)
	0.63	0.30	−0.50	0.090 (−48%)	0.095 (−55%)	0.068 (−11%)
	0.63	0.30	0.00	0.069 (−12%)	0.077 (−26%)	0.057 (6%)
	0.63	0.30	0.50	0.047 (22%)	0.060 (1%)	0.046 (25%)
	0.63	0.30	0.90	0.030 (50%)	0.046 (24%)	0.036 (40%)

NOTE: Estimated Returns to Education are equal to 0.061.

goes up considerably, although not quite as much as under the CME model.

One conclusion is that the CME model may overstate the biases associated with measurement error, as well as understate them. A second point is that although CME alone in the dependent variables does not lead to bias, if correlated with measurement error in the regressors it can affect the results considerably.

5. CONCLUSION

The behavioral model of reporting that underlies the CME model is that the individual passively reports a flawed but unbiased measurement. In contrast, the OPE model implies that the individual is fully aware of his/her ignorance and actively seeks to provide an optimal response given his/her information set. This leads to measurement error that is uncorrelated with variables in his/her information set and therefore, by necessity, correlated with the true value of interest.

We show that, in the linear regression framework, under plausible alternatives to CME, measurement error in the regressor can lead to over- as well as underestimation of the coefficients of interest. In addition, when both regressor and outcome variables are measured with correlated errors, biases can be away from 0 even in the CME model. Critical for determining the bias is the model for the individual reporting the mismeasured variables, the content of the individuals' information set, and the correlation structure of the errors. In particular, the range of estimates consistent with amounts of measurement error found in wages and years of schooling can be quite wide, especially if the measurement error in variables is correlated.

Although the classical model for measurement error may be suitable in certain cases, in general the appropriate model is likely to be context specific. The implication of our analysis is that the choice of model for measurement-error issues deserves greater consideration than typically given.

ACKNOWLEDGMENTS

We are grateful to Manuel Arellano, David Card, the editor, an associate editor, and a referee for comments and discussions.

[Received November 1998. Revised March 2001.]

REFERENCES

Angrist, J., and Krueger, A. (2000), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics* (Vol. 3), eds. O. Ashenfelter and D. Card, Amsterdam: Elsevier, pp. 1277–1366.

Ashenfelter, O., and Krueger, A. (1994), "Estimates of the Economic Return to Schooling From a New Sample of Twins," *American Economic Review*, 84, 1157–1173.

Berkson, J. (1950), "Are There Two Regressions," *Journal of the American Statistical Association*, 45, 164–180.

Bollinger, C. (1996), "Bounding Mean Regressions When a Binary Regressor Is Mismeasured," *Journal of Econometrics*, 73, 387–400.

Bound, J., Brown, C., and Mathiowetz, N. (in press), "Measurement Error in Survey Data," in *Handbook of Econometrics* (Vol. 5), eds. J. Heckman and E. Leamer, Amsterdam: Elsevier.

Bound, J., and Krueger, A. (1991), "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics*, 9, 1–24.

Card, D. (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64, 957–979.

Card, D., and Hyslop, D. (1997), "Does Inflation 'Grease the Wheels of the Labor Market'?" in *Reducing Inflation: Motivation and Strategy*, eds. Christina Romer and David Romer, Chicago: University of Chicago Press, pp. 71–114.

Chesher, A., and Schluter, C. (1999), "Welfare Measurement and Measurement Error," unpublished manuscript, University College London, Dept. of Economics.

Das, M., Dominitz, J., and Van Soest, A. (1999), "Comparing Predictions and Outcomes: Theory and Application to Income Changes," 94, 75–85.

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

Griliches, Z. (1987), "Economic Data Issues," in *Handbook of Econometrics* (Vol. 3), eds. Z. Grilliches and M. Intrilligator, Amsterdam: Elsevier, pp. 1465–1514.

Hellerstein, J., and Imbens, G. (1999), "Imposing Moment Restrictions by Weighting," *Review of Economics and Statistics*, 81, 1–14.

Horowitz, J., and Manski, C. (1995), "Identification and Robustness With Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

Kane, T., Rouse, C., and Staiger, D. (1999), "Estimating Returns to Schooling When Schooling Is Mismeasured," Working Paper 7235, National Bureau of Economic Research, Cambridge, MA.

Klepper, S., and Leamer, E. (1984), "Consistent Sets of Estimates for Regressions With Errors in All Variables," *Econometrica*, 52, 163–183.

Leamer, E. (1987), "Errors in Variables in Linear Systems," *Econometrica*, 55, 893–909.

Li, T., and Vuong, Q. (1998), "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

Mankiw, N. G., and Shapiro, M. (1986), "News of Noise. An Analysis of GNP Revisions," *Survey of Current Business*, 66, 20–25.

Manski, C. (1990), "The Use of Intentions Data to Predict Behavior: A Best-Case Analysis," *Journal of the American Statistical Association*, 85, 934–940.

Philipson, T. (in press), "Missing Data and Incentives," *Econometrica*, 69.

Pischke, S. (1995), "Measurement Error and Earnings Dynamics: Some Estimates From the PSID Validation Study," *Journal of Business & Economic Statistics*, 13, 305–314.

Rosenbaum, P., and Rubin, D. (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212–218.

Schennach, S. (1999), "Estimation of Nonlinear Models with Measurement Error," unpublished Ph.D. dissertation, Massachusetts Institute of Technology, Dept. of Economics.