MEASUREMENT ERROR IN HUMAN CAPITAL AND THE BLACK-WHITE WAGE GAP

Christopher R. Bollinger*

Abstract—Proxy variables are frequently used in economics to control for unavailable variables in a linear regression setting. For example, AFQT scores have been used to control for human capital accumulation in measuring black-white wage differentials. This practice may bias the coefficient estimates for the correctly measured variables as well. This paper models proxy variables as a measurement error process and derives bounds for the coefficients on the correctly measured variables under a variety of assumptions. The results show that the coefficient on race in a linear regression is an overstatement of the actual black-white wage gap. Sensitivity analysis suggests that if human capital could be correctly measured it would be unlikely that the coefficient on black would be negative.

I. Introduction

THERE are many circumstances in microeconomic empirical work where a researcher cannot easily or directly measure the variable theory requires as an explanatory variable. In many of these cases, researchers rely upon a *proxy* variable. Researchers casually expect the proxy variable to "absorb" the effects of the unobtainable variable called for by theory. Often the proxy is merely a control variable and the researcher has primary interest in another variable. An important example of this is presented in Neal and Johnson (1996). The authors are primarily interested in the black-white wage differential, but desire to control for "human capital youths have attained by their late teens as a predetermined initial condition that constrains the future path of human capital and, hence, future wages" (Neal and Johnson, 1996, pp. 871–872). However, human capital (HC) is, at best, difficult to measure. Neal and Johnson (1996) utilize the Armed Forces Qualification Test (AFQT) and the Armed Services Vocational Aptitude Battery (ASVAB) as measures of acquired HC. The approach is creative and potentially useful. However, a maintained, but incorrect, assumption is that the coefficients on other variables are identified correctly when this proxy variable is used. In fact, proxy variables can be modeled as an errors-in-variables problem. Hence, the results of Neal and Johnson (1996), though interesting, may be severely biased due to measurement error. This paper presents a general model of proxy variables and applies it to measuring the black-white wage differential using the data and model posited by Neal and Johnson (1996).

The model presented here assumes that the vector of explanatory variables can be broken down into (1) proxy variables which are linearly related to the true variable of interest and (2) correctly measured variables. Krasker and Pratt (1986) demonstrate that in this type of model, the coefficients on the slopes are not identified, and that coefficients on the proxy variables are not bounded. The linear relationship posited below is relatively general: it requires only uncorrelated error structure rather than strict mean independence. Since any pair of variables can be related through a linear projection, the model is not overly restrictive. Moreover, it allows the proxy variables to be linearly related to all of the unmeasured variables, not simply the variable for which it is a proxy. Extending results based upon Klepper and Leamer (1984), bounds for coefficients on all correctly measured variables in the model of Neal and Johnson (1996) are estimated. The possibility of identifying and estimating bounds for the coefficients where the variable is measured using a proxy variable is explored. Similar bounds have been proposed by Klepper and Leamer (1984), Klepper (1988), Leamer (1987), Erikson (1993), and Bollinger (1996), among others. Except for Klepper (1988), Erikson (1993), and Bollinger (1996), all these results require the measurement error to be an additive white noise process.

The finding is that the black-white wage differential, after controlling for early accumulation of HC, is likely to be smaller than estimated by Neal and Johnson (1996). Hence, the role played by early accumulation of HC is even more important in explaining black-white wage differentials than Neal and Johnson (1996) find. Further, a sensitivity analysis reveals that it is unlikely that the true coefficient on black is even negative. For females, the bounds on the black coefficient are entirely contained in the positive real line. For males, it is found that the correlation between wage and HC would have to be remarkably low to allow the coefficient on black to be negative. Furthermore, it is found that differences between males and females in the response to HC accumulation and the black-white wage differential are potentially smaller than previously thought, due to differences in the structure of the measurement error.

Section II describes the model and utilizes the results of Klepper and Leamer (1984) to establish bounds for coefficients on the correctly measured variables. Section III presents estimated bounds for the model of Neal and Johnson (1996) and examines the robustness of their results to the use of a proxy variable. Section IV presents conclusions and suggestions for future research.

II. Bounds for a General Errors-in-Variables Model

A general version of the errors-in-variables model, which is well suited to the proxy-variable problem, is given by the following equations:

Received for publication May 21, 2001. Revision accepted for publication June 28, 2002.

* University of Kentucky.

I would like to thank (without implication) Chuck Manski, Art Goldberger, Mark Berger, Dan Black, Amitabh Chandra, and an anonymous referee for comments and suggestions. I also thank Derek Neal for providing the data.

$$Y = \underline{Z}_1'\underline{\beta} + \underline{Z}_2'\underline{\alpha} + u, \qquad E[u|\underline{Z}_1, \underline{Z}_2] = 0, \tag{1}$$

$$\underline{X} = \underline{\delta} + \Gamma\underline{Z}_1 + \underline{\epsilon}, \tag{2}$$

where $\underline{Z}_1$ is a $k_1 \times 1$ vector of unobserved variables, $\underline{X}$ is a $k_1 \times 1$ vector of observed proxy variables, $\underline{Z}_2$ is a $k_2 \times 1$ vector of observed variables, and the following assumptions hold:

Assumption 1. $V(\underline{\epsilon}) = D$ is a $k_1 \times k_1$ diagonal matrix with diagonal elements $d_i \geq 0$.

Assumption 2. $\mathrm{Cov}(\underline{Z}_1, \underline{\epsilon}) = 0$ and $\mathrm{Cov}(\underline{Z}_2, \underline{\epsilon}) = 0$.

Assumption 3. $\mathrm{Cov}(\underline{\epsilon}, u) = \underline{0}$.

Assumption 4. Γ is a positive definite $k_1 \times k_1$ matrix with diagonal elements γ_{ii}.

The researcher can only observe the scalar Y, the $k_1 \times 1$ vector of proxy variables $\underline{X}$, and $\underline{Z}_2$ (it is presumed that the intercept is included in the $k_2 \times 1$ vector $\underline{Z}_2$). The researcher is primarily interested in the parameters $\underline{\alpha}$, and $\underline{\beta}$ is of secondary importance. The noise variable $\underline{\epsilon}$ is only assumed to be uncorrelated with $\underline{Z}$. This subsumes many measurement error processes. The assumptions on Γ ensure that the proxy variables are positively correlated with their respective true variables. This assumption is simply a normalization; the key part of it is that the sign of the correlation is known. An implication of the assumptions above is that $\underline{X}$ and $\underline{Z}_1$ have the same number of elements. If there exist extra proxy variables, then it is possible to identify the parameter vector $\underline{\alpha}$ via an instrumental variables approach (this is shown in the appendix below). However, the parameter vector $\underline{\beta}$ is not generally identified (again see appendix). For the case where there are fewer proxy variables, it becomes problematic, for there is now additionally an omitted variable bias. One may then consider reformulating the model in terms of fewer latent variables. The case where there is a one-to-one correspondence between proxies and latent variables is actually quite common. Here, for example, a single score represents ability.

The failure of identification of $\underline{\beta}$ is easily seen. If it were known that Γ was the identity matrix, then this would simply be a classical errors-in-variables model, which is well known to be unidentified. Since Γ is unknown, the model cannot be identified. However, it is commonly misunderstood that $\underline{\alpha}$ is also unidentified unless $\underline{Z}_1$ and $\underline{Z}_2$ are uncorrelated, in which case $\underline{\alpha}$ is identified simply by the regression of Y on $\underline{Z}_2$. Since a major concern in the literature on the black-white wage gap is that the average difference between black earnings and white earnings is explained, in part, by differential attainment of HC during the individual's youth (see the discussion in Neal & Johnson, 1996), this assumption is untenable.

For simplicity, this section considers the case where $\underline{Z}_1$ and $\underline{Z}_2$ are scalars ($k_1 = k_2 = 1$). Without loss of generality, the variables are all assumed to have mean zero. The more general case is considered in the appendix.

To understand the identification failure and to apply the theorems of Klepper and Leamer (1984), the above model can be transformed. Let a be the slope coefficient, and let Y^* be the residuals from the linear projection of Y on Z_2. Similarly, let F be the coefficient, and let $\underline{X}^*$ be the residuals from the linear projection of X on Z_2. Let Φ be the coefficient, and Z_1^* be the vector of residuals from the linear projection of Z_1 on Z_2. Using standard omitted variable bias results, a number of relationships can be described:

$$a = \alpha + \Phi\beta, \tag{3}$$

$$F = \Phi\Gamma, \tag{4}$$

$$Y^* = Z_1^*\beta + u, \tag{5}$$

$$X^* = Z_1^*\Gamma + \varepsilon. \tag{6}$$

Additionally, define $Z_1^{**} = \underline{Z}_1^*\Gamma$ and $\theta \equiv (\Gamma^{-1})\beta$. Using these definitions, equations (5) and (6) can be rewritten as

$$Y^* = \underline{Z}_1^{**\prime}\underline{\theta} + u, \tag{7}$$

$$X^* = Z_1^{**} + \varepsilon, \tag{8}$$

which is now a classical errors-in-variables model with slope parameter $\underline{\theta}$. The identification failure for α is now readily seen: both Φ and β must be identified from the observed data to identify α. Even if Γ were known to be I (thus allowing identification of Φ), β would still be unidentified due to measurement error. Using the reparameterization above, the results of Klepper and Leamer (1984) can be applied directly to bound θ. Bounds on θ can then be used, through equation (3), to bound α.

It is helpful to adopt notation similar to that used by Klepper and Leamer. Let b^d be the coefficient from the direct regression of Y^* on X^*. Let b^{rev} be the reciprocal of the regression coefficient from the reverse regression of X^* on Y^* (see the appendix for the more general case involving k_1 such reverse regressions). The result of Klepper and Leamer (1984) can be directly applied to bound θ.

Theorem 1 (Application of result of Klepper and Leamer [1984]) One has

$$|b^d| \leq |\theta| \leq |b^{\mathrm{rev}}|,$$

and $\mathrm{sign}(b^d) = \mathrm{sign}(b^{\mathrm{rev}}) = \mathrm{sign}(\theta)$.

This univariate result can be attributed to Frisch (1934); a more general multivariate result (used in the appendix) was developed by Klepper and Leamer (1984). In the multivariate case, the bounds can fail if the direct and reverse regressions change sign. When θ is a scalar, the direct and reverse regressions must have slopes of the same sign, and so the bounds exist. The upper bound is achieved when $V(u) = 0$: when the only noise in the system is from

a single mismeasured variable. Hence, Klepper and Leamer (1984) and Klepper (1988) show that tighter bounds can be achieved when restrictions are placed on the correlation of Y^* with Z_1^{**}. These tighter bounds can be used as a specification analysis of the type called for by Leamer (1985).

The corollary below gives bounds on the parameter α. Using equation (3) and the bounds on θ, we have

Corollary 1. The set of feasible values for α can be represented as

$$\alpha \in \{c \mid c = a - F(\lambda b^{d} + (1 - \lambda)b^{rev})\},$$

where $0 \leq \lambda \leq 1$.

If b^{d} and hence b^{rev} are both positive and F is negative (as is the case in the example below), then $a - Fb^{d} \leq \alpha \leq a - Fb^{rev}$. In general, the two bounds are given by $a - Fb^{rev}$ and $a - Fb^{d}$, with the upper bound being the larger of the two expressions, and the lower bound being the smaller. It is interesting to note that even though the parameter β is neither identified nor even bounded (see Krasker & Pratt, 1986), the remaining slope coefficients can be bounded. However, it is also important to note that the slope coefficients on correctly measured variables are not identified if any variable is measured with error. The exception to this is if $\Phi = 0$, and hence $F = 0$: the case where the unobserved variable Z_1 is uncorrelated with the observed variable Z_2. In this case, though, it would be unnecessary to control for the variable Z_1 in the first place.

The above results derive a set of bounds which are population parameters. That is, given the underlying population structure described in equations (1) and (2), and assumptions 1–4, the true parameter α will lie within the bounds in the population. In estimation there are two interpretations. The first, and most obvious, is that the estimated bounds are estimates of the population bounds. Hence, inference on the estimated bounds (such as confidence intervals) gives information on where the true bounds lie. The researcher can construct a 95% confidence interval for the lower and upper bounds. One could then use these to construct a 95% confidence interval for the true parameter. Unlike confidence intervals for point estimates, this interval would not collapse to a point as the sample size grew, but would rather collapse to the population bounds. It can also be shown (see Bollinger, 1993, or Klepper & Leamer, 1984) that these bounds hold in sample as well as in population. The estimated bounds would be exact bounds for the estimated slope if the variable Z_1 were observed in this data set. That is, if the true Z_1 for these data were suddenly available, the estimated slopes would fall within the estimated bounds.

In many cases, the resulting bounds for θ and hence α are too wide to be of practical use. Klepper and Leamer (1984) suggest two approaches to tightening the bounds. In each case, they require knowledge about the strength of the underlying relationships in the model, either through the correlation between Z_1 and X, or through the R^2 of the regression of Y on Z_1 and Z_2. The use of these approaches is explored in the next section. However, it is important to note that bounds for β cannot be obtained by tightening the bounds on θ.

In the classical errors-in-variables literature, information on the variance of Z is sufficient to obtain identification of all parameters. Here, that is not the case, but this information can be used to then derive the bounds for β. Without loss of generality, assume $\beta > 0$.

Corollary 2. If there exist known U and L such that $U \geq V(Z_1) \geq L$, then

$$b^{d}\sqrt{\frac{V(X)}{U}} \leq \beta$$

and

$$\beta_i \leq b^{rev}\sqrt{\frac{V(X)}{L}}.$$

The proof is presented in the appendix. In the case where β_i is negative, the bounds are reflected into the negative orthant. This information is not useful for tightening the bounds on α, because it does not tighten the bounds on θ. The information which can tighten the bounds for θ does not provide bounds for β, and conversely, information which provides bounds for β does not tighten the bounds for θ. The issue here is that bounds for β can only be achieved when information about the relative scale of X and Z is available either through the variance of Z, or through information about Γ (as in the classical errors-in-variables case where Γ is known). Tightening the bounds for θ, and hence α, requires information about the signal-to-noise ratio. In the classical errors-in-variables case, the scale and the signal-to-noise ratio contain the same information.

III. Estimation of Wage Equations

Neal and Johnson (1996) present a pathbreaking set of results, demonstrating that the black-white wage differential may, in large part, be due to differences in the HC stock obtained by the late teen years. They argue that this stock of HC is largely exogenous, and is important in predetermining future education and wage paths. If this stock could be measured, the black-white wage differential could be isolated. They argue that later schooling is, in part, endogenous when the HC stock is not measured. The opportunities for additional schooling or other training may be largely correlated with race. Hence, using total education to control for HC may bias the measurement of the black-white wage differential due to endogeneity. This implies a model

$$\log wage = \alpha_0 + \beta\, HC + \alpha_1\, Black + \alpha_2\, Hispanic + \alpha_3\, Age + u. \tag{9}$$

TABLE 1.—SAMPLE DESCRIPTIVE STATISTICS

	Men (N = 1593)		Women (N = 1449)	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
log wage	6.793	0.463	6.602	0.486
AFQT	0.013	1.000	0.076	0.895
Black	0.293	0.455	0.296	0.457
Hispanic	0.190	0.392	0.202	0.402
Age	27.06	0.806	27.07	0.805

Neal and Johnson (1996) utilize data derived from the National Longitudinal Survey of Youth. These same data are used here.[1] As noted in Neal and Johnson (1996), the sample consists of respondents born between 1962 and 1964 who took the AFQT (over 90% of the original sample) and who had a valid wage for either 1990 or 1991. A complete discussion of the construction of the data set can be found in Neal and Johnson (1996). Table 1 presents the means of the variables used in the analysis. As with Neal and Johnson (1996), men and women are analyzed separately.

In the model of Neal and Johnson (1996), Y is the natural log of wages, and Z_1 is the immeasurable HC attainment discussed by the authors. The variable X is the score on the AFQT. The variable Z_2 comprises age and indicators for black and Hispanic. Clearly, the AFQT score is not a perfect measure of HC attainment. Moreover, it is unlikely that the relationship between the HC variable from an earnings equation and the test score on the AFQT is simply an additive white-noise measurement error process. In fact, it is likely to be far more complicated. However, the measurement error model in the previous section does not rule out more complicated relationships. It simply expresses the relationship between the observed score and the true HC variable as a linear projection, which of course exists trivially for any pair of variables with finite variances.

Assumption 4 does restrict the linear projection to have a positive slope. It is difficult to dismiss this assumption as trivial, but the fact that the military and other organization use the test as a crude measure of practical skill suggests, at least, that it is likely. Certainly this assumption underlies the analysis of Neal and Johnson (1996).

A. *Basic Bounds*

Table 2 presents the direct regression results as found in Neal and Johnson (1996). The coefficient on AFQT is b^d, the *direct* regression. The coefficients on the other variables in the direct regression of table 2 also provide the left bounds for those coefficients. The first row of table 4 also presents the bounds for the black coefficient in this case. Table 3 presents the reverse regression slope and the other half of the bounds for α.

[1] I gratefully acknowledge Neal and Johnson's willingness to provide these data. Some minor differences in the results are explained by editing performed on the NLSY since Neal and Johnson (1996) was completed.

TABLE 2.—DIRECT LOG *WAGE* REGRESSION RESULTS

	Men		Women	
Variable	(1)	(2)	(1)	(2)
AFQT		0.171		0.231
		(0.012)		(0.015)
Black	−0.244	−0.072	−0.185	0.035
	(0.026)	(0.027)	(0.029)	(0.031)
Hispanic	−0.114	0.006	−0.028	0.145
	(0.030)	0.030	(0.033)	(0.033)
Age	0.048	0.040	0.010	0.023
	(0.014)	(0.013)	(0.015)	(0.015)
Constant	5.591	5.738	6.386	5.926
	(0.379)	(0.357)	(0.425)	(0.395)

It should be noted that the bounds on α correspond to a particular value for θ. That is, one cannot combine tables 2 and 3 and pick coefficients for each term from both. If you argue that $\theta = 0.171$ (for men), then it must be that the coefficient on black (for men) is -0.072. As this corresponds to the direct regression, it must also be that the variance of ε (the measurement error term) is zero. Similarly, if you argue that $\theta = 1.489$, then the coefficient on black must be 1.255, and the variance of u must be zero (the true model is deterministic).

The results in tables 2 and 3 shed interesting light on the black-white wage differential. Due to measurement error, the coefficient that Neal and Johnson (1996) present for men is actually the largest negative magnitude possible. Thus, their results overstate the actual black-white wage differential. They conclude (see p. 874) that HC differentials (at age approximately 18) explain nearly three-fourths of the marginal black-white wage differential. This is, in fact, an understatement. Error in measuring the HC has caused this coefficient to be inflated. Indeed, if the measurement error is large enough, it may actually be that the wage differential, after controlling for HC stock, is positive.

For women the small positive coefficient estimated by Neal and Johnson (1996) is also an understatement. Indeed, if HC were perfectly measured, black women might actually make more than their white counterparts by 3.5% or more. Although the upper bound is highly unlikely, it clearly

TABLE 3.—STANDARDIZED LOG *WAGE* REVERSE REGRESSIONS

Variable	Men	Women
AFQT	1.489	1.648
	(0.070)	(0.065)
Black	1.255	1.393
	(0.127)	(0.126)
Hispanic	0.929	1.198
	(0.111)	(0.115)
Age	−0.023	0.100
	(0.039)	(0.039)
Constant	6.864	3.101
	(1.056)	(1.074)

places the true coefficient on black women in the positive range.

To understand why Neal and Johnson (1996) get an understated estimate of the coefficient on black, note that the bounds for the coefficient on black are

$$[a_{\text{black}} - F_{\text{black}}b^{\text{d}}, a_{\text{black}} - F_{\text{black}}b^{\text{rev}}]. \tag{10}$$

The term a_{black} is found in the first and third columns of table 2 (for men and women respectively). The term F_{black} is the coefficient on black from a regression of AFQT on the two race variables and age. For men it is -1.006; for women it is -0.952. Since $b^{\text{d}} \leq b^{\text{rev}}$, the slope from the direct regression (which assumes no measurement error and is thus one end of the spectrum) is the smaller—or more negative—of the two bounds. Intuitively, this is similar to an omitted variables regression. The AFQT proxy behaves like true HC, but is an imperfect measure. Hence, some of the variation in HC is not taken into account in the direct regression estimated by Neal and Johnson (1996). If black and HC are correlated, then the coefficient on black will pick up some of the unaccounted variation in HC. Since, by assumptions 2 and 4, the covariance of black and AFQT has the same sign as covariance of black and human capital, it follows that on average blacks had lower HC accumulation than whites (at the time of the AFQT test). Since black and HC are negatively related whereas HC and log *wage* are positively related, the coefficient on black will be biased downward by the unaccounted variation in HC.

The use of direct and reverse regressions in examining discrimination and racial wage differentials has a substantial history. For example, see Kamalich and Polachek (1982) or Conway and Roberts (1983). In these cases, the coefficients on race (or gender) in the reverse regression (ability on salary and race) were interpreted as measures of how productivity or ability were distributed across race, holding salary constant. Often the reverse regression shows coefficients on the race (or gender) variable which are small in magnitude and insignificant. This is taken to be evidence of fairness. The framework above allows this claim to be evaluated. Rearranging equation (9) yields

$$HC = \frac{-\alpha_0}{\beta} + \frac{1}{\beta}\log \text{wage} - \frac{\alpha_1}{\beta}\text{Black} - \frac{\alpha_2}{\beta}\text{Hispanic} - \frac{\alpha_3}{\beta}\text{Age} - \frac{1}{\beta}u. \tag{11}$$

Hence, the coefficient on black, for example, is a rescaled version of the original coefficient. In addition, the variable u is correlated with log wage, and hence introduces a downward bias on the coefficient on log wage (see Klepper & Leamer, 1984—this gives the upper bound on β) and hence on all other coefficients as well (also resulting in these being bounds). The reverse regression coefficient on black may be quite small in magnitude, depending on the magnitude of β and the variance of u. Thus tests based upon simply the reverse regression coefficient estimate (as opposed to the rescaled coefficient used here) may not have much power. Failure to reject should not be taken as evidence of fairness. The approach here formalizes the information contained in the reverse regressions in a way that allows the question of interest to be addressed.

TABLE 4.—SUMMARY OF BOUNDS FOR θ, β (THE COEFFICIENT ON HUMAN CAPITAL), AND α_{BLACK} (THE COEFFICIENT ON BLACK)

Scenario	θ	β	α_{black}
Men			
Basic result (tables 2, 3)	[0.17, 1.49]	—	[−0.07, 1.26]
Assume $V(HC) = 1$	[0.17, 1.49]	[0.17, 1.49]	[−0.07, 1.26]
$R^2_{\max} = 0.1627$ ($\alpha_{\text{black}} < 0$)	[0.17, 0.24]	—	[−0.07, 0]
$\rho^2_{\min} = 0.707$ ($\alpha_{\text{black}} < 0$)	[0.17, 0.24]	—	[−0.07, 0]
Women			
Basic result (tables 2, 3 above)	[0.23, 1.65]	—	[0.04, 1.39]
Assume $V(HC) = 1$	[0.23, 1.65]	[0.22, 1.56]	[0.04, 1.39]
$R^2_{\max} = 0.1627$ ($\alpha_{\text{black}} < 0$)	—	—	—
$\rho^2_{\min} = 0.707$ ($\alpha_{\text{black}} < 0$)	—	—	—

IV. Bounds for Human Capital Coefficient

To obtain bounds on the HC coefficient β, additional information must be used. Since HC has no intrinsic scale, an approach that has appeal is to assume that its variance is 1. This represents an arbitrary scaling of the variable. Then, comparisons of HC can be done in terms of standard deviations from the mean (of zero). Using the results in corollary 2 gives the bounds presented in the second row of table 4. The fact that the variance of the AFQT score for men is 1 (see table 1) implies that the bounds reported in tables 2 and 3 represent bounds on the response of log wages to a 1 standard deviation change in HC accumulation. That is, log wages will rise between 0.171 and 1.489. Note that this additional information gives no improvement for the bounds θ. Since the bounds on the coefficient for black derive only from the bounds on θ, there is no improvement there either.

For women the variance for AFQT scores is 0.801 (see table 1), so the bounds from theorem 2 imply that a 1 standard deviation change in HC attainment results in a change in log wages between 0.219 and 1.559. This is also presented in the second row of table 4. It is interesting in that it more closely aligns the female response to HC with the male response. This suggests that previous results which suggest female wages respond more to HC attainment may indeed be due to differences in the measurement error structure between men and women, rather than differences in market forces. In fact, empirical research into response error in other variables suggests that women and men do have different processes (see, for example, Bollinger, 1998; Bound & Krueger, 1991; Bollinger & David, 1997). Again,

note that the added information gives no improvement for the bounds on θ and hence the coefficient on black.

A. *Sensitivity Analysis*

The upper bound on θ can only be achieved if $V(u) = 0$, that is, only if the structural equation (9) is deterministic. This seems unlikely. However, it is interesting to note that in this case, bounds appear relatively tight (see Bollinger, 1996, or Bollinger, 2001, for examples where the bounds are very wide). However, it is difficult to judge, for θ has little economic content as it stands. Klepper and Leamer (1984) demonstrate how information about the R^2 of the relationship between Y^* and Z_1^{**} (which is equivalent to the R^2 of the regression of Y on Z_1 and Z_2 here) can be used to tighten the bounds on θ, and consequently the bounds on α. The upper bound on θ and the resulting positive upper bound on the coefficient on black are associated with an R^2 of 1 for the structural regression of Y on Z_1, Z_2. Hence, bounding that R^2 will move the positive bound on black toward zero. Klepper and Leamer (1984) show that if $R^2 \leq R^2_{\max}$, then $0 \leq R^2_{\max} b^{\text{rev}}$ (for the scalar case studied here). Rather than posit a particular $R^2_{\max}$, one can find the value for $R^2_{\max}$ which gives a zero upper bound on the coefficient for black. This is asking what conditions support a negative coefficient on black. Since a negative coefficient on black is the usual finding, the answer to this question provides a measure of how sensitive that finding is to measurement error. Using the results in theorem 1, the condition can be found by solving for $R^2_{\max}$ in the equation $a_{\text{black}} - F_{\text{AFQT,black}} b^{\text{rev}} R^2_{\max} = 0$. For the males, the resulting value is 0.1627. The associated bounds are presented in the third row of table 4. The very low R^2 seems an unlikely condition and suggests that the coefficient on black may indeed be zero or even slightly positive, as with the results for women. Failing to properly measure the actual HC attainment at age 18 may account for all of the black-white wage differential.

Similarly, information about the correlation between the true regressor (in this case ΓZ_1^*) and the observed X^* can be used to tighten the bounds (Klepper & Leamer, 1984). Again, the correlation between Z_1^* and X^* is equal to the correlation between Z_1 and X, due to the assumption that the error ε is uncorrelated with Z_2. For the simple case where only one variable is measured with error, the results from Klepper and Leamer (1984) are greatly simplified. Given an upper and a lower bound (ρ^{*2} and ρ^2_* respectively) on the correlation between Z_1 and X, the results of Klepper and Leamer and Theorem 1 above can be used to show that

$$\rho^{*2} \geq \frac{b^{\text{d}}}{\theta} \geq \rho^2_* \tag{12}$$

Note that if $\rho^{*2} = 1$, this establishes that b^{d} is the lower bound on θ, while if $\rho^2_* = 0$, it establishes that b^{d} and θ have the same sign. Hence, tightening ρ^{*2} will tighten the lower bound, whereas tightening ρ^2_* will tighten the upper bound. As above, the value for ρ^2_* that bounds θ low enough to ensure that the coefficient on black is negative can be solved by finding $a_{\text{black}} - F_{\text{AFQT,black}}(b^{\text{d}}/\rho^2_*) < 0$. Solving this equation implies that the correlation between HC attainment at age 18 and the AFQT must be *at least* 0.707 to ensure that the coefficient on black is negative. This result is shown in the fourth row of table 4. If ρ^2 were less than 0.707, then the coefficient on black would be nonnegative. This statement is much more difficult to analyze than the restriction on the R^2 of the wage equation. However, it seems unlikely that the AFQT is that precise a measure of human capital.

The two restrictions are related. In fact, it is not difficult to show that a correlation between X and Z_1 of at least 0.707 requires that R^2 of the regression of Y^* on Z_1^* must be no larger than 0.1627. The intuition is straightforward: There are two sources of error in the model, u and ε. The residual from the regression of Y^* on X^* (and the regression of X^* on Y^*) measures the total of these errors. The upper bound on θ assumes that this total is allocated entirely to ε, and the lower bound assumes it is allocated to u. The conditions on R^2 or ρ^2 above force an allocation of variance to each source of error. Hence, given the additivity ensured by the uncorrelated errors, the conditions must be identical. Hence, in order to have a coefficient on black that is negative in sign, the R^2 between Y^* and X^* must be no larger than 0.1627, which ensures that at least 70% of the variation in AFQT is due only to variation in HC.

V. Conclusions

The above analysis suggests that HC attainment at age 18 may explain all of the gross differences in wages between blacks and whites. The fact that HC is, at best, only grossly controlled for leads to potential bias in the measure of the black-white wage differential. In this case, that bias is large enough to switch signs for the men. For the women, as showed by Neal and Johnson (1996), the coefficient is measured as positive, but that may also be an understatement. This finding suggests that policy-makers should focus on early education rather than ex post labor market outcomes.

Moreover, the paper finds that some of the observed difference between male and female coefficients on proxy variables for HC may be due to differences in the measurement error structure rather than actual differences in returns to human capital. Similarly, the usual result that black men are the most affected by black-white wage differentials is in question also. It may be simply that there are different measurement error processes at work rather than actual differences in labor market outcomes.

This paper also brings a general overall warning to researchers: when proxy variables are used, coefficients on other variables may also be biased—in some cases severely. Using the approach here, an understanding of the effect of measurement error may be gained. Since the use of proxy variables is widespread in microeconomic work, a general

call for this kind of specification analysis is warranted. Further research into analytic solutions for more complicated proxy-variable specifications is also warranted.

REFERENCES

Black, Daniel A., Mark C. Berger, and Frank A. Scott, "Bounding Parameter Estimates with Nonclassical Measurement Error," *Journal of the American Statistical Association* 95:451 (2000), 739–748.

Bollinger, Christopher R., "Measurement Error in Binary Regressors, with an Application Bounding the Union Wage Differential," University of Wisconsin PhD dissertation (1993).

Bollinger, Christopher R., "Bounding Mean Regressions When a Binary Regressor is Mismeasured," *Journal of Econometrics* 73 (1996), 387–399.

Bollinger, Christopher R., "Measurement Error in the CPS: A Nonparametric Look," *Journal of Labor Economics* 16:3 (1998), 576–594.

Bollinger, Christopher R., "Response Error and the Union Wage Differential," *Southern Economic Journal* 68:1 (2001), 60–76.

Bollinger, Christopher R., and Martin H. David, "Modeling Discrete Choice with Response Error: Food Stamp Participation," *Journal of the American Statistical Association* 92:439 (1997), 827–835.

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

Conway, Delores A., and Harry V. Roberts, "Reverse Regression, Fairness, and Employment Discrimination," *Journal of Business and Economic Statistics* 1 (1983), 75–85.

Erikson, Timothy, "Restricting Regression Slopes in the Errors-in-Variables Model by Bounding the Error Correlation," *Econometrica* 61:4 (1993), 959–970.

Frisch, R., *Statistical Confluence Analysis by Means of Complete Regression Systems* (Oslo, Norway: University Institute for Economics, 1934).

Kamalich, Richard, and Solomon Polachek, "Discrimination: Fact or Fiction? An Examination Using an Alternative Approach," *Southern Economic Journal* 49 (1982), 1227–1229.

Klepper, Steven, "Bounding the Effects of Measurement Error in Regressions Involving Dichotomous Variables," *Journal of Econometrics* 37 (1988), 343–359.

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

Krasker, William S., and John W. Pratt, "Bounding the Effects of Proxy Variables on Regression Coefficients," *Econometrica* 54:3 (1986), 641–655.

Leamer, Edward E., "Sensitivity Analysis Would Help," *American Economic Review* 75 (1985), 308–313.

Leamer, Edward E., "Errors in Variables in Linear Systems," *Econometrica* 55:4 (1987), 893–910.

Neal, Derek A., and William R. Johnson, "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy* 104:5 (1996), 869–895.

APPENDIX

1. Multivariate Version of Theorem 1

Equations (3) through (8) easily generalize to a multivariate model. Hence equations (7) and (8) along with assumptions 1–4 define a classical errors-in-variables system with k_1 mismeasured right-hand-side variables. It is straightforward to then apply the results of Klepper and Leamer (1984). They define the direct regression and $k_1 + 1$ reverse regressions. The reverse regression coefficients are arrived at by regressing each observed X^*-variable upon Y^* and the remaining X^*-variables and rearranging the resulting fitted equation so that Y is back on the left-hand side. Let $\underline{b}^d$ be the coefficient vector from the direct regression, and let $\underline{b}^j$ be the vector of coefficients from the j^{th} reverse regression. Then if the $k + 1$ direct and reverse regression coefficient vectors are all in the same orthant, the set of feasible values for $\underline{\theta}$ is the convex hull of the $k + 1$ regressions (Klepper & Leamer, 1984). That is,

$$\underline{\theta} \in \{\underline{b} \mid \underline{b} = \lambda_d \underline{b}^d + \lambda_1 \underline{b}^1 + \cdots + \lambda_k \underline{b}^k\} \tag{A1}$$

where $\lambda_j \geq 0$ and $\lambda_d + \lambda_1 + \cdots + \lambda_k = 1$.

The lower bound for each variable will be the smallest coefficient value from the $k_1 + 1$ direct and reverse regression vectors, and the upper bound will be the largest coefficient value from the $k_1 + 1$ regression vectors. The requirement that all $k_1 + 1$ vectors be in the same orthant ensures that the sign of the coefficient on each variable is the same across all $k_1 + 1$ coefficient vectors. If the coefficients change sign, the bounds do not exist.

The result for the case where there is only one X-variable (as discussed in the text) is straightforward. As Frisch (1934) showed, the sign of the direct and reverse regressions coefficients will always be the same. Hence, for the scalar case, the general expression above reduces to

$$\underline{\theta} \in \{b \mid b = \lambda b^d + (1 - \lambda) b^{rev}\}. \tag{A2}$$

Thus, for $\theta > 0$, we have $b^d \leq \theta \leq b^{rev}$, with the result reflected into the negative orthant when $\theta < 0$.

2. Proof of Corollary 1

As noted above, equation (3) is easily expressed for the multivariate case. The coefficient vector $\underline{a}$ from the regression of Y on $\underline{Z}_2$ can be written as

$$\underline{a} = \underline{\alpha} + \Phi\underline{\beta}. \tag{A3}$$

The $k_2 \times k_1$ matrix Φ is composed of the coefficient vectors from the projection of each variable in $\underline{Z}_1$ on $\underline{Z}_2$ (the $k_2 \times 1$ vectors from each regression are stacked side by side). Rearranging yields

$$\underline{\alpha} = \underline{a} - \Phi(\Gamma\Gamma^{-1})\underline{\beta}. \tag{A4}$$

Using the definitions of $\underline{\theta} = \Gamma^{-1}\underline{\beta}$ and $F = \Phi\Gamma$, results in

$$\underline{\alpha} = \underline{a} - F\underline{\theta}. \tag{A5}$$

The factors $\underline{a}$ and F are estimable from the sample. The bounds on $\underline{\theta}$ above give the final bounds on $\underline{\alpha}$. An important point to note here is that each feasible value for the vector $\underline{\theta}$ described by equation (A-2) yields a feasible value for α. This implies that some combinations of values for individual coefficients in α are not feasible. That is, one cannot simply pick any combination of values within the upper and lower bounds for each coefficient.

3. Proof of Corollary 2

A more detailed proof can be found in Bollinger (1993).

Utilizing the framework of Klepper and Leamer (1984), and defining the i^{th} row of V as V_i, it can be shown that

$$\Sigma = \Gamma^{-1}\left[V - \text{diag}\left\{\frac{V_i(\underline{\theta} - \underline{b}^d)}{\theta_i}\right\}\right]\Gamma^{-1}. \tag{A6}$$

Thus the bounds on σ_{ii} imply that

$$U \geq \left(\frac{1}{\gamma_i}\right)^2\left[V_{ii} - \frac{V_i(\underline{\theta} - \underline{b}^d)}{\theta_i}\right] \geq L. \tag{A7}$$

Defining the ratio $C_{ij} = v_{ij}/v_{ii}$, using the definition of θ_i, and rearranging yields

$$\frac{V_{ii}}{L\gamma_i}\left(b_i^d + \sum_{j \neq i} C_{ij}(b_j^d - \theta_j)\right) \geq \beta_i, \tag{A8}$$

$$\frac{V_{ii}}{U\gamma_i}\left(b_i^d + \sum_{j \neq i} C_{ij}(b_j^d - \theta_j)\right) \leq \beta_i. \tag{A9}$$

Define $\underline{b}^{\wedge}$ as an arbitrary convex linear combination of all regression coefficient vectors except $\underline{b}^{r}$. By the results of Klepper and Leamer (1984), the feasible values of $\underline{\theta}$ are of the form $a\underline{b}^{\wedge} + (1 - a)\underline{b}^{r}$ with $a \in [0, 1]$. It can be shown that

$$b_i^{d} + \sum_{j \neq i} C_{ij}[b_j^{d} - ab_j^{\wedge} - (1 - a)b_j^{r}] = ab_i^{\wedge} + (1 - a)R_{A_i}^2 b_i^{r} \tag{A10}$$

and

$$\frac{b_i^{d}}{b_i^{r}} \leq R_{A_i}^2. \tag{A11}$$

Using the expressions (A-20) and (A-11), it can be shown that the minimum of the left-hand side of equation (A-9) occurs at $a = 1$. The maximum of the left-hand side of equation (A-8) occurs at $a = 1$ if $R_{A_i} b_i^{r} \leq b_i^{\max\wedge}$, and occurs at $a = 0$ otherwise. Hence

$$\frac{V_{ii} b_i^{\min}}{U\gamma_i} \leq \beta_i \leq \frac{V_{ii} b_i^{\max s}}{L\gamma_i} \text{ or } \frac{V_{ii} b_i^{\max\wedge}}{L\gamma_i}. \tag{A12}$$

Since $b_i^{\max} \geq \beta_i/\gamma_i \geq b_i^{\min}$ from proposition 1, then $b_i^{\max}/\beta_i \geq 1/\gamma_i \geq b_i^{\min}/\beta_i$. This last fact is used to establish the theorem. In the univariate case, there are only b^{d} and b^{rev}. Substituting these for $b^{\wedge}$ and b^{rev} respectively, and recognizing that $R_{A_i} b_i^{r} \leq b_i^{\max\wedge}$ must always hold for this case, the result follows.

4. Identification Issues for Multiple Proxies

We consider the identification of parameters when the number of proxy variables exceeds the number of latent variables. Consider the model with two proxy variables X_1 and X_2 for a single latent variable Z_1 and an observed variable Z_2. Using the framework above, consider the residuals Y^*, X_1^*, and X_2^* from the regressions of Y, X_1, and X_2 on Z_2:

$$Y^* = \beta Z_1^* + u, \tag{A13}$$

$$X_1^* = \gamma_1 Z_1^* + \varepsilon_1, \tag{A14}$$

$$X_2^* = \gamma_2 Z_1^* + \varepsilon_2. \tag{A15}$$

The covariance of Y^* and X_1^* is $\gamma_1 V(Z_1^*)$, and the covariance of Y^* and X_2^* is $\gamma_2 V(Z_1^*)$. The covariance of X_1^* and X_2^* is $\gamma_1\gamma_2 V(Z_1^*)$. Hence, both $\theta_1 = \beta/\gamma_1$ and $\theta_2 = \beta/\gamma_2$ are identified by the ratios of covariances. The slope coefficient from the regression of Y on Z_2 is again

$$a = \alpha + \Phi\beta = \alpha + \Phi\begin{pmatrix}\gamma_1\\ \gamma_1\end{pmatrix}\beta = \alpha + F_1\theta_1. \tag{A16}$$

Since the regression of X_1 on Z_2 yields F_1, and θ_1 is estimable from the IV regression, the term α is identified. The results easily extends to the multivariate case. Black, Berger, and Scott (2000) provide an approach for a related case.