Estimating the Returns to College Quality with Multiple Proxies for Quality

Dan A. Black, *Syracuse University and NORC*

Jeffrey A. Smith, *University of Michigan*

Existing studies of the effects of college quality on wages typically rely on a single proxy variable for college quality. This study questions the wisdom of using a single proxy given that it likely contains substantial measurement error. We consider four econometric approaches to the problem that involve the use of multiple proxies for college quality: factor analysis, instruments variables, a method recently proposed by Lubotsky and Wittenberg, and a GMM estimator. Our estimates suggest that the existing literature understates the wage effects of college quality and illustrate the value of using multiple proxies in this and other similar contexts.

I. Introduction

A growing literature in economics estimates the labor market effects of the quality of the college an individual attends. The literature proceeds by estimating the parameters of linear “education production functions” with an outcome of interest (such as wages) on the left-hand side and...
some measure of college quality, such as the average Scholastic Aptitude Test (SAT) score of the entering class, on the right-hand side, usually along with a wealth of covariates designed to take account of nonrandom selection of students into schools of differing qualities. A related literature performs similar analyses to investigate the effects of primary and secondary school quality. This article reconsiders the standard education production function literature in a context where multiple measures of college quality are available. Our analysis applies to many other similar contexts both within and outside labor economics and the economics of education.

We motivate our analysis in Section II by carefully considering the parameters of interest in studies of college quality and the link between these parameters and the estimates in much of the existing literature. Most papers in the existing literature include only a single measure of college quality, which they interpret as a proxy for a latent one-dimensional “college quality” variable. To the extent that the proxy variable measures college quality with error, we expect bias toward zero. Additional bias of unknown direction may arise if we allow the scale of the proxy variable to differ from that of latent college quality. As a result of these issues, existing estimates of the effect of college quality may exhibit substantial biases.

In light of these concerns, our article considers different ways of using multiple proxies to better estimate the impact of college quality than the current literature. After introducing the data and, in particular, our multiple proxies for college quality, in Section III, in Section IV we explicitly model the problem of multiple proxies and derive a measurement error model that allows the variance of each proxy to differ from the variance of unobserved college quality. In Section V, we present several different sets of estimates. First, for comparison purposes, we present ordinary least squares (OLS) estimates similar to those in the rest of the literature. Second, we combine the information in the proxies using factor analytic methods to produce a college quality index that we then include in the outcome equation. Third, we adopt the standard solution to classical measurement error and use instrumental variables (IV) methods, with the additional proxies serving as instruments. Fourth, we derive a generalized method of moments (GMM) estimator that identifies, subject to a required normalization, the structural parameters of our more general measurement error model, and we argue for the superiority of this approach on econ-
omestic grounds. Fifth, we present some sensitivity analyses. Section VI discusses the bounds developed in Lubotsky and Wittenberg (forthcoming) and reports the results of applying their method to our data. Section VII summarizes our contributions and highlights our main finding that the existing literature appears to understate the wage effect of college quality.

II. The Parameter of Interest and the Literature

Consider, in somewhat more formal terms, the education production function, defined as

\[ Y = f(q_1, \ldots, q_k, X), \]

where \( Y \) denotes an outcome of interest, such as wages; \( q_1, \ldots, q_k \) denote various college-level inputs (which we also refer to as measures or dimensions of college quality), such as the average SAT score of the entering class, expenditures per student, and so on; and \( X \) denotes other factors affecting earnings and college quality choice.

Based on this version of the production function, we can define various parameters of interest; in particular, we can define derivatives with respect to various college-level inputs. Consider input \( k \) and the usual linear approximation to the production function, so that the parameters of interest become derivatives of the linear conditional expectation function. A natural parameter of interest is the partial derivative with respect to one dimension of quality, holding the other dimensions (and the \( X \)) constant.1 In notation,

\[ P_i = \frac{\partial E(Y|q_1, \ldots, q_k, X)}{\partial q_k}, \]

where \( P_i \) denotes parameter 1. This parameter is of particular interest to policy makers and college administrators making choices regarding which dimensions to focus on when cutting (to minimize the damage) or augmenting (to maximize the improvement) a college budget. Monks (2000) estimates \( P_i \) using data from the National Longitudinal Survey of Youth (NLSY—the same data we employ in this study). Zhang (2005) estimates \( P_i \) for state university systems using the Baccalaureate and Beyond data. Long (2004) estimates this parameter with a very large number of school characteristics in his study of secondary school quality.

The literature on college quality (but not that on primary and secondary

\[ 1 \text{ We follow the existing college quality literature, which treats the slope coefficient on quality as the same across individuals. In a world of heterogeneous slope coefficients, the standard production function regression estimates, under some additional assumptions, what Wooldridge (2002b) calls the average partial effect.} \]
school quality) often implicitly adopts the simplifying assumption of a “one-factor” model, in which quality has a single dimension. In this case, the production function simplifies to

\[ Y = f(Q^*, X), \]

where the variable \( Q^* \) is a single factor that we refer to as “college quality.” The asterisk indicates that the variable is latent. The assumption that \( Q^* \) is a scalar is a strong one, as schools may have multiple dimensions, with, for instance, University of Chicago excelling at liberal arts training and MIT excelling at technical training.

The partial derivative with respect to college quality represents the obvious parameter of interest in the one-factor model. In terms of our notation,

\[ P_2 = \frac{\partial E(Y|Q^*, X)}{\partial Q^*}. \]

This parameter indicates the effect of an increase in (latent) college quality on the outcome of interest, holding \( X \) constant. The one-factor model has the virtue of both conceptual simplicity and ease of interpretation in cases where budgetary allocations within a college are not the primary policy issue of interest.

Empirically, aside from Fitzgerald (2000), Monks (2000), Zhang (2005), and our own papers (Daniel, Black, and Smith 1995, 1997; Black and Smith 2004; Black, Daniel, and Smith 2005), most of the literature adopts the following strategy. A single college quality measure \( q_j \) is chosen—often some measure of selectivity in admissions—and included in outcome equations along with covariates. Most studies assume what Heckman and Robb (1985) term “selection on observables,” in the hope that the inclusion of a sufficiently rich \( X \), including at least some measure of individual “ability” (usually a test score), controls for the nonrandom matching of students and colleges. Some more recent studies adopt alternative identification schemes that attempt to take account of selection on unobservables. These include the Behrman, Rosenzweig, and Taubman (1996) study, which uses data on twins, and the Brewer, Eide, and Ehrenberg (1999) study, which uses a parametric polychotomous selection model with variables related to net college costs as exclusion restrictions. Dale and Krueger (2002) represents an intermediate case because of their access to data on which colleges students applied to and which colleges accepted them, variables not normally observed in studies of this type.

In this article, we assume selection on observables and focus instead on the issue of how to interpret the parameter estimated in most of the literature; this issue arises regardless of the chosen econometric identification strategy. We can interpret this parameter in three ways, none of
Table 1
College Quality Variables, NLSY Men, 1989

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum Value</th>
<th>Maximum Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty-student ratio</td>
<td>.0663</td>
<td>.0264</td>
<td>.02</td>
<td>.25</td>
</tr>
<tr>
<td>Rejection rate</td>
<td>.255</td>
<td>.165</td>
<td>.00</td>
<td>.82</td>
</tr>
<tr>
<td>Freshman retention rate</td>
<td>.750</td>
<td>.123</td>
<td>.24</td>
<td>.98</td>
</tr>
<tr>
<td>Mean SAT score (+ 100)</td>
<td>9.36</td>
<td>1.44</td>
<td>5.50</td>
<td>13.75</td>
</tr>
<tr>
<td>Mean faculty salaries</td>
<td>.0550</td>
<td>.0107</td>
<td>.0236</td>
<td>.0958</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty-Student Ratio</td>
</tr>
<tr>
<td>Faculty-student ratio</td>
</tr>
<tr>
<td>Rejection rate</td>
</tr>
<tr>
<td>Freshman retention rate</td>
</tr>
<tr>
<td>Mean SAT score</td>
</tr>
<tr>
<td>Mean faculty salaries</td>
</tr>
</tbody>
</table>


which is very satisfactory. First, we can interpret the existing literature as estimating \( P_3 \), defined as

\[
P_3 = \frac{\partial E(Y|q_k, X)}{\partial q_k}.
\]

In words, \( P_3 \) denotes the partial derivative of the conditional expectation function with respect to one dimension of quality, holding \( X \) but not the other dimensions of quality constant, a parameter that lacks both a clear economic interpretation and any obvious policy relevance. We cannot interpret the literature as estimating \( P_1 \) because, as we show in table 1, the various dimensions of quality have nontrivial positive correlations with one another. As a result, we expect that \( P_1 > P_3 \). Put differently, when the different dimensions of college quality have positive correlations, including only one dimension means that its coefficient incorporates some of the effects of the other dimensions. (This point holds more generally, and indicates that studies that seek to estimate \( P_1 \) require data on all of the relevant inputs to the production function in order to avoid confusing the effects of omitted inputs with those of included inputs.) Because \( P_3 \) has a clear interpretation, while
does not, the necessity of interpreting the existing studies as estimating the latter renders their estimates problematic.

Second, we can interpret the various quality measures \( q_1, \ldots, q_k \) as proxies for the latent quality variable rather than as inputs. In this view, the existing studies estimate \( P_2 \) using the single quality measure \( q_k \) as a proxy for the latent quality \( Q^* \). Most authors in the existing literature implicitly or explicitly interpret their estimates in this way. For example, Dale and Krueger (2002) treat quality as synonymous with selectivity and interpret the coefficients on their average SAT score variable (or on the *Barron’s* magazine quality category dummies that they employ in a separate specification) as estimates of the effect of both quality and selectivity. Other studies talk primarily about selectivity, prestige, or competitiveness but clearly intend these as synonyms for quality. Recent papers in this group include Fox (1993), Loury and Garman (1995), Hoxby (1998), and Chevalier and Conlon (2003).

The second interpretation of the existing literature also raises important conceptual issues. Using only a single proxy variable means that the obtained estimates likely understate the parameter of interest because the proxy variable measures the latent variable with error. The extent of the bias toward zero depends on the extent of measurement error in the proxy variable. Additional biases of unknown direction arise when the proxy variable and the latent variable do not share the same scale. We discuss these scaling issues in more detail below.

The third interpretation returns to treating the observed quality measures as inputs into the production of latent quality, rather than as proxies for quality, and then makes some strong assumptions. To begin, assume that all universities and colleges face the same prices and that the underlying college quality production function is homogeneous of degree \( r \) in the inputs. Universities pick different levels of quality because of difference in endowments, and, in the case of public universities, because of political constraints. Because the production function is homogeneous of degree \( r \) in the inputs and all universities face the same prices, we know that for \( \lambda > 0 \) inputs differ only by the scale of production or

\[
\lambda Q^*(q_1, \ldots, q_k) = Q^*(\lambda q_1, \ldots, \lambda q_k).
\]

The *Barron’s* college quality categories, which are also used in Brewer et al. (1999), and related categorizations, such as the prestige categories in Chevalier and Conlon (2003), raise interesting issues when considered as proxies. Such categorical measures implicitly combine information from multiple measures of college quality (albeit in a less formal way than methods we consider here), which should reduce the amount of measurement error they embody. At the same time, these measures throw out all the variation within categories, which works in the opposite direction.
In this framework, should a researcher enter multiple inputs into the wage equation, the inputs should be perfectly collinear. Indeed, in this framework we need only one input to capture perfectly the production of quality. Under these assumptions, the usual approach in the literature estimates the returns to the latent quality variable.

The very strong, and implausible, assumptions of homogeneity of the production function and common prices derail this approach in our view. We know of no evidence in favor of the homogeneity of this particular production function. It also seems clear that input prices differ among colleges. For example, in the case of student ability, colleges in locations with a large population of highly educated parents and/or with more amenities will face a lower cost of student ability. Finally, this approach requires input measures without error. If, for example, the average SAT score measures student ability with substantial error, as the large sums spent on admissions offices suggest, then the need for the approaches examined in this article returns, although the coefficient estimates we obtain have a somewhat different interpretation under these assumptions.

In sum, the existing estimates in the literature present important interpretational difficulties. They likely represent biased estimates, perhaps substantially biased estimates, of \( P_2 \), the usual implicit or explicit parameter of interest in these studies.\(^3\)

III. Data Description

Our data come from three sources. Our primary data source is the 1979 Cohort of the National Longitudinal Survey of Youth (NLSY), a panel data set based on surveys of a sample of men and women who were 14–21 years old on January 1, 1979. Respondents were first interviewed in 1979, then reinterviewed annually from 1979 to 1994 and biennially since 1994. Because we are interested only in the postcollege earnings of these respondents, we use earnings data from 1989. We chose 1989 because, given the subsequent attrition in the NLSY, it maximizes our sample size. We limit our sample to men who have attended postsecondary schools for which we have measures of quality, which is roughly the set of 4-year comprehensive colleges and universities.\(^4\) We focus on men to

\(^3\) See McClellan and Staiger (1999) for a related discussion in the context of measuring hospital quality.

\(^4\) In the course of our earlier work (Daniel et al. 1995, 1997), we compared estimates constructed using all NLSY men and estimates constructed using a subsample of those who attended college, where the latter is broader than the sample we employ here because it includes individuals who attended colleges for which we do not have quality measures. The substantive results did not differ very much. Despite this, we prefer to err on the side of caution and exclude individuals who either did not attend college or attended a college for which we
avoid having to deal with labor force participation issues, which are not our primary concern in this article. See Black and Smith (2004) for more details about the construction of the sample.

The NLSY suits our purpose well for several reasons. First, the timing means that we have information on wages for a relatively recent cohort of college graduates that is old enough for the vast majority of those who will attend college to have already done so. Furthermore, those who will attend graduate school have largely completed doing so as well. Second, the NLSY confidential files provide information on individual colleges attended, which allows us to match up information on specific colleges from external sources. Third, the NLSY allows us to construct a compelling “ability” measure using the Armed Services Vocational Aptitude Battery (ASVAB), which was administered to over 90% of the sample. Fourth, the NLSY is rich enough in other covariates to make the assumption that conditioning on observable characteristics alone solves the problem of nonrandom sorting into colleges of varying qualities plausible. These covariates include detailed information on family background, home environment, and high school characteristics.

Our sources for college characteristics are the Department of Education’s Integrated Post-Secondary Education Data System (IPEDS) for 1992 and the US News and World Report’s Directory of Colleges and Universities (US News and World Report 1991). We only include information for 4-year colleges; roughly one half of the men in the NLSY data with some postsecondary education attended a 4-year college. The analysis sample includes 398 distinct colleges.

We focus on five measures of quality: faculty-student ratio, the rejection rate among those who applied for admission, the freshman retention rate, the mean SAT score of the entering class, and mean faculty salaries. We focus on these measures for two reasons. First, many of them have been used in previous studies as measures of quality. Second, the response rates do not have quality measures in order to avoid having the estimated relationship driven by observations from outside our population of primary interest.

5 Neal and Johnson (1996) describe the test in detail and discuss the issues of interpretation surrounding it.

6 Although the timing of these college quality measures differs somewhat from the timing of college attendance for most of our sample, these measures have a very high serial correlation, so that only a small amount of measurement error likely results from the timing difference.

7 The first four measures are from the US News and World Report’s Directory of Colleges and Universities, and the last is from the IPEDS data. For schools that report an average ACT score rather than an average SAT score, we impute an average SAT score. We redefine the raw college characteristics so that larger values correspond to obvious notions of quality, e.g., we recode the “acceptance rate” as a “rejection rate” and use the latter.
for these measures are relatively high, which is important because we limit
our sample to individuals whose colleges reported all five measures. The
top panel of table 1 displays the summary statistics for these measures,
and the bottom panel displays their correlations. The correlations range
from a maximum of 0.70 to a minimum of just 0.31.

If each of the quality variables perfectly measured college quality, the
correlations would always be one, which they clearly are not. Thus, we
must interpret these variables as proxies for college quality. Proxy variables
are a staple of econometric models; see Wooldridge (2002a, 2003) for
textbook treatments and Bollinger (2003) for further discussion. We refer
to the difference between one of our proxy variables and latent college
quality as "measurement error." If our proxy variables embody classical
measurement error, then every pair among them should have the same
covariance (equal to the unknown variance of the latent college quality
variable). The data strongly reject this restriction, which indicates that we
require a more general measurement error model. The next section out-
lines such a model.

IV. Econometric Model

The relatively low correlations among the various measures of quality
suggest that they contain much measurement error. To focus our discus-
sion, consider the following model of wage determination:

$$\ln (w_{ij}) = X_i \beta + \delta S_i + \gamma Q^*_i + e_{ij},$$

where $\ln (w_{ij})$ is the natural logarithm of the wage rate of the $i$th person
attending the $j$th college, $X_i$ is a vector of covariates, $S_i$ is the number of
years of schooling, $Q^*_i$ is the latent quality variable defined in Section III
for college $j$, $e_{ij}$ is an error term assumed to be uncorrelated with the
regressors, and $(\beta, \delta, \gamma)$ are parameters to be estimated.$^9$ The parameter
$\gamma$ corresponds to our parameter $P_2$ above.

The inclusion of years of completed schooling is controversial. As Black

$^8$ These variables may measure not only latent college quality with error but
also the quantity to which they directly correspond with some error due to data
collection or definition problems and to gaming of these measures by the colleges
involved. We do not treat this type of measurement error separately here.

$^9$ We examined alternative specifications in which college quality entered non-
linearly, but we found little evidence of departures from linearity. Because it is
not the main point of our article and because we appear to lack the sample size
to precisely estimate a model with nonlinear quality effects, we focus on the linear
specification in our empirical work. We also experimented with interacting the
quality measures with the years of schooling variable, but we obtained only
substantively and statistically insignificant coefficients on the interaction term.
and Smith (2004) discuss in some detail, there is a strong correlation between years of college completed and the quality of the institution attended. To the extent that attending a high-quality university increases the number of years of schooling, the model given in equation (1) will underestimate the returns to attending a high-quality school. To keep our results comparable with the previous literature, however, we will condition on completed years of schooling in this study.

To make the exogeneity of schooling and college quality plausible, we need to condition on a rich set of covariates. Our specification of $X$, includes quadratics in the first two principal components of the age-adjusted ASVAB scores (as suggested by Cawley, Heckman, and Vytlacil [2001]), a black indicator, a Hispanic indicator, a quartic in age, and region of birth dummies. We also include variables measuring home characteristics (whether at age 14 the respondent’s household subscribed to a magazine, whether it subscribed to a newspaper, and whether the respondent had a library card), parental characteristics (the years of schooling of each parent, whether the parents were living together in 1979, whether the mother was alive in 1979, whether the father was alive in 1979, and parental occupations in 1978), and high school characteristics (size of the high school, number of books in the library, fraction of the student body that was economically disadvantaged, and mean teacher salary). Rather than dropping observations with missing values on one or more of the home, parental, and high school characteristics due to item nonresponse, we recode the missing values to zero and add indicators for missing values.

Unfortunately, we do not measure college quality directly but rather must rely on noisy proxies, defined as

$$q_{kj} = \alpha_k Q^p + u_{kj},$$

where $\alpha_k > 0$ is a scale coefficient and $u_{kj}$ is measurement error that we
assume is uncorrelated with both \( Q_{ij}^* \) and \( X \). Our model (modestly) generalizes the classical measurement error model, which requires that \( q_{ij} = Q_{ij}^* + u_{ij} \). In particular, the inclusion of the scale coefficients allows the covariances of the various proxies, \( q_{ij} \), to differ.

Rather than work directly with equation (1), it is convenient to consider a simple transformation that allows us to ignore the \( X \) and \( S \) in (1). Let \( \ln(\tilde{w}_{ij}) \) denote the residual from the regression of \( \ln(w_{ij}) \) on \( X \) and \( S \) and let \( q_{ij} \) denote the residuals from similar regressions of \( q_{ij} \) on \( X \) and \( S \). We refer to \( \ln(\tilde{w}_{ij}) \) and \( q_{ij} \) as “Yulized residuals” in honor of Yule’s (1907) discovery of this decomposition. Using the Yulized residuals, we have

\[
\ln(\tilde{w}_{ij}) = \gamma \tilde{q}_{ij} + \epsilon_{ij}.
\]

Ordinary least squares or instrumental variables estimation of equation (3) will provide the same estimates of \( \gamma \) as OLS or IV estimation of equation (1). Equation (3), however, provides some insights into the measurement problems. When the covariates explain a substantial portion of the total variation in \( Q_{ij}^* \) (by assumption they explain none of variation in \( u_{ij} \)), then noise necessarily makes up a larger proportion of the Yulized residuals and the resulting estimates must be attenuated more than when \( X \) and \( S \) account for less of the variation in \( Q_{ij}^* \). As the OLS estimate of \( \gamma \) equals \( \text{Cov}(\log(\tilde{w}_{ij}), \tilde{q}_{ij})/\text{Var}(\tilde{q}_{ij}) \), the more of the variation in \( Q_{ij}^* \) that the covariates remove, the larger the noise-to-signal ratio and the greater the attenuation bias in the estimate of \( \gamma \). While this point has been recognized when dealing with panel data and fixed-effect or first-differenced estimation—see, for example, Griliches and Hausman (1986) and Bound and Krueger (1991)—it may not be fully appreciated in a cross-sectional context.

11 Our problem is similar to the problems addressed in the MIMIC (Multiple Indicators Multiple Causes) and LISREL frameworks (see, e.g., Jöreskog and Goldberger 1975; and Bollen 1989; see also the discussion of LISREL and partial least squares in Dijkstra [1983]). Heckman, Stixrud, and Urzua (2006, in this issue) consider a model in which both cognitive and noncognitive skills are imperfectly measured and extend the LISREL and MIMIC models in a matching framework where the matching variables are not observed, but their distributions are estimated nonparametrically.

12 In the regression context, we may also allow the proxies to be of the form \( q_{ij} = \beta_i + \alpha_i Q_{ij}^* + u_{ij} \) without any added complexity; the parameter \( \beta_i \), however, is not identified.

13 Using \( \ln(w_{ij}) \) would increase notational consistency, but we prefer \( \ln(\tilde{w}_{ij}) \) for simplicity.

14 We thank Terra McKinnish for the reference to Yule’s work. See also the discussion of “double residual regression” in Goldberger (1991) and the discussion of the Frisch-Waugh-Lovell Theorem in Davidson and MacKinnon (1993, 62–69).
As is well known, classical measurement error generally attenuates the coefficient estimates. As our discussion of the Yulized residuals demonstrates, estimation of a model that includes a rich set of covariates exacerbates the attenuation bias because the covariates explain a portion of the error term $u_{ij}$ and so increase the noise-to-signal ratio. This effect has empirical relevance in our context; when we regress each quality measure on $X$ and $S$, we account for 18% of the variation in the faculty-student ratio, 24% of the variation in the rejection rate, 29% of the variation in the freshman retention rate, and 25% of the variation in average SAT scores and average faculty salaries. Of course, we cannot solve this problem by simply dropping variables from the model because only a rich covariate set makes our “selection on observables” identification strategy plausible.

Given the modest correlations among the quality measures in table 1, removing a substantial fraction of the systematic variation may lead to a lot of attenuation. To see why, we note that, under our form of nonclassical measurement error,

\[ \text{plim } \hat{\gamma}_{\text{OLS}} = \frac{\gamma}{\alpha_e} \left( 1 + \frac{\text{Var}(\tilde{u}_{ik})}{\alpha_e \text{Var}(\tilde{Q}_{jk})} \right)^{-1}, \]  

Equation (4) is also useful for seeing the fundamental identification problem faced when using proxy variables. Even in the absence of measurement error, so that $\text{Var}(\tilde{u}_{ik}) = 0$, the OLS estimate will be biased unless $\alpha_e = 1$. In the presence of measurement error, we cannot determine whether the estimates are biased upward or downward. For instance, when $\alpha_e < 1$, the estimates may be biased upward despite the attenuation bias that results from the measurement error.
Table 2
Impact Estimates from Regressions with Each Quality Variable Individually and with All Quality Variables, NLSY Men, 1989

<table>
<thead>
<tr>
<th>Quality Measure</th>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
<th>Column 5</th>
<th>Column 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty-student ratio</td>
<td>.013</td>
<td>.025</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.0164)</td>
<td>(.0157)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rejection rate</td>
<td>.003</td>
<td>.026</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.0199)</td>
<td>(.0185)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freshman retention rate</td>
<td>.038</td>
<td>.048*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.0269)</td>
<td>(.0199)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean SAT score</td>
<td>.002</td>
<td>.037*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.0232)</td>
<td>(.0172)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean faculty salaries</td>
<td>.008</td>
<td>.035*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.0237)</td>
<td>(.0198)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significant at the 5% level in a one-tailed test.


Note.—College quality measure is for last college attended as of 1989. The regressions also include years of schooling, quadratics in the first two principal components of the age-adjusted ASVAB scores, a black indicator, a Hispanic indicator, a quartic in age, and region of birth dummies. We also include controls for home characteristics (whether at age 14 the household subscribed to a magazine, whether it subscribed to a newspaper, and whether the respondent had a library card), parental characteristics (education of the parents, whether their parents were living together in 1979, whether the mother was alive in 1979, whether the father was alive in 1979, and parental occupations in 1978), and high school characteristics (size of high school, number of books in the school library, fraction of student body that was economically disadvantaged, and mean teachers’ salaries). To avoid losing sample due to missing values resulting from item nonresponse, we recoded the home, parental, and high school characteristics missing values to zero and then added indicator variables that equal one if the corresponding data element is missing. The dependent variable is the natural log of the respondent’s wage, defined as earnings in 1988 (the year prior to the 1989 survey) divided by hours in 1988. Each college quality measure is normed to have unit variance. Huber-White standard errors appear in parentheses.

We believe that the failure to model this scale factor may be of first-order importance. For instance, few would dispute that the average SAT score of the entering class would be correlated with the quality of a college or university. Indeed, many authors use this as their sole measure of college quality. It is another matter, however, to assert without corroborating evidence that this measure varies on the same scale as college quality or, perhaps more important, given that latent quality lacks a natural scale, to assert that the average SAT score variable varies on the same scale as other measures of college quality.15

Equation (4) also suggests that, because of the differences in scale, our OLS estimates are not directly comparable. To make the OLS estimates roughly comparable to each other and to the factor analysis estimates in the next section, we normalize each of the college quality measures to have unit variance. In column 1 of table 2, we report estimates from a model that includes all five proxies for college quality. In such a specification, the coefficients on the individual quality measures correspond to our parameter $p_i$. None of the five coefficient estimates differ signifi-

15 Bollinger (2003) notes the fundamental nonidentification problem when there is a single proxy and derives bounds on the coefficient $\gamma/\alpha$. 
cantly from zero at conventional levels, despite the fact that we use one-tailed tests (both here and throughout the tables) due to the nature of our null hypothesis.

Given equations (2) and (3), however, this is hardly surprising. Identification of the parameters on the college quality measures rests on components that are orthogonal to the other quality measures. If the single factor model is correct, the Yulized residuals of the quality measures (which now condition on the other quality measures in each case, as well as on $X$ and $S$) will (asymptotically) contain only limited information. To see why, consider the case of two proxies (and to keep the analysis simple, assume no other covariates). The regression of $q_1$ on $q_2$ produces residuals that (asymptotically) equal

$$
\frac{\alpha_1 \text{Var} (u_2)}{\alpha_1^2 + \text{Var} (u_2)} Q^* + u_1 - \frac{\alpha_1 \alpha_2^2}{\alpha_2^2 + \text{Var} (u_2)} u_2.
$$

When either $u_2$ has a low variance or the first measure provides a weak signal (as indicated by a low value of $\alpha_1$), this residual embodies mainly noise rather than signal. Indeed, if the second measure has no measurement error so that $\text{Var} (u_2) = 0$, the Yulized residual contains nothing but noise. Thus, in the context of the single factor model, including multiple proxies and trying to interpret the individual coefficients makes little sense.

In columns 2–6 of table 2, we report estimates from regressions that include each one of the proxy variables in turn. When entering the equation alone, the estimated coefficient for each of the measures exceeds—usually substantially—the corresponding coefficient when the variables enter jointly. As discussed in Section II, we interpret these as estimates of the difficult-to-interpret parameter $P_3$ rather than as estimates of $P_1$, the parameter of primary interest within the single-factor model of college quality.

**B. Factor Analysis Estimates**

Intuitively, we should be able to combine the various measures of college quality to obtain a more reliable measure of $Q^*$. More formally, suppose that, across all colleges, $E(Q^*) = 0$, a harmless normalization that keeps the notation simple. Let $q = (q_1, \ldots, q_K)^T$ be a $K$-vector of noisy signals of the quality of each college, such that for a college with quality $Q^*$, the value of each signal is $q_{kj} = \alpha_j Q_j^* + u_{kj}$ with $E(q_{kj}) = 0$; $E(u_{kj}) = a_{kj}$; $E(u_{kj} u_{lj}) = 0$, $\forall j \neq l$; $E(u_{kj} u_{kl}) = 0$, $\forall k \neq l$; and $E(Q_j^* u_{kj}) = 0$. We construct a measure of college quality by taking a linear combination of the signals. Define $Q = \sum_{k=1}^K \tau_k q_{kj}$, where there is no need for an intercept
term because we normalized the expected value of $Q^*$ to zero. We select $Q^*_g$ to minimize the expected squared distance between $Q$ and $Q^*$, or

$$\min_{t, \ldots, t_K} E(Q^* - \hat{Q})^2. \tag{5}$$

The necessary conditions for minimization are

$$\alpha_k \text{Var}(Q^*) - \alpha_k \sum_{i=1}^K \alpha_i \text{Var}(Q^*) - \tau_k \sigma_k^2 = 0$$

$$\forall \ k \in \{1, 2, \ldots, K\}, \tag{6}$$

or

$$1 - \sum_{i=1}^K \alpha_i \tau_i - \alpha_i \tau_i \tau_k = 0$$

$$\forall \ l \in \{1, 2, \ldots, K\}, \tag{7}$$

where $\tau_k$ is the noise-to-signal ratio $\sigma_k^2 / [\alpha_k \text{Var}(Q^*)]$. Evaluating equation (7) at $k = 1$ and $k = l$ implies that

$$\alpha_l \tau_l = \alpha_l \frac{\tau_1}{\tau_l}. \tag{8}$$

Thus, we may rewrite equation (7) as

$$\alpha_l^{-1} \tau_l^{-1} - \tau_l \left( \sum_{i=1}^K \frac{\tau_i^{-1}}{l} + 1 \right) = 0. \tag{9}$$

Solving for $\tau_l$ we obtain

$$\tau_l = \frac{\alpha_l^{-1} \tau_l^{-1}}{1 + \sum_{i=1}^K \frac{\tau_i^{-1}}{l}}. \tag{10}$$

The remaining $\tau_k$ have similar formulas. Thus, $\tau_k$ decreases in the variance of the idiosyncratic error $u_k$, so that signals that more accurately reflect the latent college quality receive more weight in the forecast. When we use only two factors ($K = 2$) to construct the index of college quality, we refer to the model as a “two-variable model”; if we use all five variables ($K = 5$), we refer to the index as the “five-variable model.”

Readers familiar with the psychometrics literature may recognize this model as a transformation of Spearman’s (1904) factor model; see Harman (1976) for a good discussion of the historical development of this model.

To implement the model, we simply specify the variables (the signals) to
be used in the factor analysis.\textsuperscript{16} In the spirit of Carniero, Hansen, and Heckman (2003), we also looked for a second factor. We found a second factor only in the case of the five-variable model; in that case, we tried including the second factor in the regression model and could not reject the null of a zero coefficient. Thus, we report only results based on first factors in all cases. The implied college quality rankings based on the first factors accord with a priori notions of quality; for example, Stanford, Harvard, MIT, Yale, and Penn make up the top five schools attended by respondents in our sample based on the five-variable model. The first factors obtained using different combinations of variables correlate strongly with one another, as expected. For the two-variable factors, the correlations range from 0.53 to 0.90.\textsuperscript{17}

After obtaining the factor loadings, we estimate equation (1) with OLS, with the quality index included as the quality measure. In table 3, we provide factor analysis estimates from the two-variable and five-variable models. The use of multiple proxies makes it reasonable to interpret these estimates as estimates of parameter $\hat{P}$. In general, the estimates decrease as the number of variables used to construct the index increases; the estimates with two proxies range from 0.049 to 0.080, and the estimate using all five proxies equals 0.042. The estimates nearly always exceed the simple OLS estimates presented in table 2, as they should given that the use of multiple proxies reduces the extent of measurement error in college quality.\textsuperscript{18}

The factor analysis approach is simple to implement and makes it easy to construct a quality index for use in ranking colleges. At the same time, the factor analysis estimates remain attenuated relative to the true value because the use of multiple signals lowers but does not eliminate the resulting measurement error.\textsuperscript{19} Thus, we now turn to an alternative in the form of instrumental variables.

\textsuperscript{16} We estimated the factor loadings using both the sample of schools attended by individuals in our analysis sample and the sample of all schools attended by anyone in the NLSY data. The factor loadings differed little between the two samples; the estimates in table 3 are based on the first set of factor loadings.

\textsuperscript{17} The factor loadings for the five-variable model are as follows: faculty-student ratio (0.096), rejection rate (0.137), freshman retention rate (0.257), mean SAT score (0.385), and mean faculty salaries (0.245).

\textsuperscript{18} The estimated standard errors in table 3 do not reflect a correction for the estimation of the factor loadings.

\textsuperscript{19} Factor analysis provides an unbiased estimate of the underlying latent variable even when only two variables are used to construct the factor. Adding additional variables to the factor analysis reduces the variance of the resulting estimate of the latent variable, meaning that, on average, it contains less measurement error. This, in turn, reduces attenuation bias in the estimated regression coefficient in eq. (1).
Table 3
Estimates from Regressions including College Quality Indices Constructed Using Factor Analysis, NLSY Men, 1989

<table>
<thead>
<tr>
<th>Model Type/Factor</th>
<th>Huber-White Standard Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Two-variable models:</strong></td>
<td></td>
</tr>
<tr>
<td>Factor combines:</td>
<td>Estimate</td>
</tr>
<tr>
<td>Faculty-student ratio and rejection rate</td>
<td>0.060</td>
</tr>
<tr>
<td>Faculty-student ratio and freshman retention rate</td>
<td>0.080</td>
</tr>
<tr>
<td>Faculty-student ratio and mean SAT scores</td>
<td>0.061</td>
</tr>
<tr>
<td>Faculty-student ratio and mean faculty salaries</td>
<td>0.060</td>
</tr>
<tr>
<td>Rejection rate and freshman retention rates</td>
<td>0.064</td>
</tr>
<tr>
<td>Rejection rate and mean SAT scores</td>
<td>0.050</td>
</tr>
<tr>
<td>Rejection rate and mean faculty salaries</td>
<td>0.054</td>
</tr>
<tr>
<td>Freshman retention rates and mean SAT scores</td>
<td>0.056</td>
</tr>
<tr>
<td>Freshman retention rates and mean faculty salaries</td>
<td>0.062</td>
</tr>
<tr>
<td>Mean SAT scores and mean faculty salaries</td>
<td>0.049</td>
</tr>
<tr>
<td><strong>Five-variable model:</strong></td>
<td></td>
</tr>
<tr>
<td>Factor combines faculty-student ratio, rejection rate, freshman retention rate, mean SAT scores, and mean faculty salaries</td>
<td>0.042</td>
</tr>
</tbody>
</table>


**Note.**—\( N = 887 \). See note of table 2 for description of the contents of the regressions. The dependent variable is the natural log of the respondent’s wage, defined as earnings in 1988 (the year prior to the 1989 survey) divided by hours in 1988. Huber-White standard errors appear in parentheses. All of the estimates in this table are significant at the 5% level in a one-tailed test. We construct each college quality index using factor analysis.

C. Instrumental Variables Estimates

Economists have long recognized that instrumental variables estimation may eliminate the bias associated with estimates obtained using variables with classical measurement error. See Griliches (1986) for a review of the early literature.

With our slightly more general form of measurement error, standard IV estimation will not provide point identification. To see why, we note that the standard two-stage least squares (2SLS) estimator with a single instrument is

\[
\hat{\gamma}^{IV} = \frac{\sum_{i=1}^{N} \hat{q}_{bi} \ln (\hat{w}_{i})}{\sum_{i=1}^{N} \hat{q}_{bi} \hat{q}_{bi}},
\]

(11)

where \( i \) indexes observations and \( (q_{b1}, q_{b6}) \) are two of the quality measures. Taking the probability limit of the IV estimator, we obtain

\[
\text{plim} \quad \hat{\gamma}^{IV} = \gamma \cdot \frac{\alpha_{6} \text{Var}(\hat{Q}^{6})}{\alpha_{6} \alpha_{6} \text{Var}(\hat{Q}^{6})} = \frac{\gamma}{\alpha_{6}}.
\]

(12)
Table 4  
IV Estimates of the Effect of College Quality, NLSY Men, 1989

<table>
<thead>
<tr>
<th></th>
<th>OLS Estimate</th>
<th>IV Estimate</th>
<th>GMM Estimate</th>
<th>Partial F-Statistic from First-Stage Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty-student ratio</td>
<td>.025</td>
<td>.107*</td>
<td>.104*</td>
<td>22.5</td>
</tr>
<tr>
<td>Rejection rate</td>
<td>.026</td>
<td>.084*</td>
<td>.078*</td>
<td>55.4</td>
</tr>
<tr>
<td>Freshman retention rate</td>
<td>.048*</td>
<td>.059*</td>
<td>.057*</td>
<td>231.8</td>
</tr>
<tr>
<td>Mean SAT score</td>
<td>.025*</td>
<td>.064*</td>
<td>.065*</td>
<td>246.3</td>
</tr>
<tr>
<td>Mean faculty salaries</td>
<td>.035*</td>
<td>.069*</td>
<td>.066*</td>
<td>134.7</td>
</tr>
</tbody>
</table>


Note.—N = 887. College quality measure is for last college attended as of 1989. See the note of table 2 for description of the contents of the regressions. The dependent variable is the natural log of the respondent’s wage, defined as earnings in 1988 (the year prior to the 1989 survey) divided by hours in 1988. Each college quality measure is normed to have unit variance. Huber-White standard errors are in parentheses. p-values are in squared brackets.

* Significant at the 5% level in a one-tailed test.

The inclusion of more instruments does not remedy the inconsistency. Hence, the parameter of interest is only identified up to a positive constant. Of course, even if we assume that there is no measurement error, so that $\sigma_i^2 = \sigma_j^2 = \ldots = \sigma_k^2 = 0$, standard OLS only identifies the parameter of interest up to a positive constant as well.

With that caveat, in table 4, we present 2SLS estimates where we use one quality measure in the “structural equation” and the remaining four measures as instruments. Each of the estimates is statistically significant, and each is much larger than the corresponding OLS estimate. The 2SLS estimates range widely, from a low of 0.059 to a high of 0.107.

Asymptotically, the OLS and 2SLS estimates differ by the term

$$\left(1 + \frac{\text{Var}(\tilde{u}_n)}{\alpha_i^2 \text{Var}(Q^n)}\right)^{-1},$$

which strictly increases in the noise-to-signal ratio. Using this observation, we estimate that the faculty-student ratio is the noisiest measure of school quality and the freshman retention rate is the least noisy measure.

By way of comparison, we also reproduce the standard optimally weighted GMM estimates; see Wooldridge (2002a). In the presence of heteroskedasticity, these estimates should be relatively more efficient than simple IV estimates, and the estimated standard errors are always smaller than the corresponding 2SLS estimates. In our application, the two sets
of estimates differ very little (from 0.001 to 0.006), with the simple IV estimates slightly higher in four of the five cases. Given that the 2SLS and GMM estimators identify the parameter of interest only up to scale, we now turn our attention to an estimator that does identify the parameter of interest, subject only to a modest normalization.20

D. Method of Moments Estimates with a Convenient Normalization

1. Estimates

When we have two quality measures, we are unable to identify the general measurement error model presented in Section IV from the covariance matrix of the data. To see why, consider the covariance matrix of the data given by

\[
\text{Var}(\ln(\bar{w})) = \gamma^2 \sigma_{\bar{w}}^2 + \sigma_{\tilde{w}}^2;
\]

\[
\text{Var}(\tilde{q}_k) = \alpha_k \sigma_{\tilde{q}_k}^2 + \sigma_{\tilde{q}_k}^2, \quad k = 1, 2, \ldots, K;
\]

\[
\text{Cov}(\ln(\bar{w}), \tilde{q}_k) = \gamma \alpha_k \sigma_{\tilde{q}_k}^2, \quad k = 1, 2, \ldots, K; \quad (13)
\]

\[
\text{Cov}(\tilde{q}_k, \tilde{q}_l) = \alpha_k \alpha_l \sigma_{\tilde{q}_k}^2, \quad k, l = 1, 2, \ldots, K, \quad k \neq l.
\]

The number of equations in this system is \(2K + 1 + \sum_{l=1}^{K-1} l\), where \(K\) is again the number of quality measures. The number of unique parameters the system contains is \(3 + 2K\). With \(K = 2\) we have six equations and seven parameters, leaving the system underidentified. We do not, of course, ever observe \(Q^*\), which suggests normalizing \(\sigma_{\tilde{q}_k}^2\) to one. Doing so reduces the number of parameters to six, so that the three “off-diagonal” elements of the covariance matrix now suffice to identify \((\alpha_1, \alpha_2, \gamma)\). It is easy to show that

\[
\alpha_1 = \frac{(\text{Cov}(\ln(\bar{w}), \tilde{q}_1) \text{Cov}(\tilde{q}_1, \tilde{q}_2))^{1/2}}{(\text{Cov}(\ln(\bar{w}), \tilde{q}_1))^{1/2}},
\]

\[
\alpha_2 = \frac{(\text{Cov}(\ln(\bar{w}), \tilde{q}_2) \text{Cov}(\tilde{q}_1, \tilde{q}_2))^{1/2}}{(\text{Cov}(\ln(\bar{w}), \tilde{q}_1))^{1/2}}, \quad (14)
\]

\[
\gamma = \frac{(\text{Cov}(\ln(\bar{w}), \tilde{q}_1) \text{Cov}(\ln(\bar{w}), \tilde{q}_2))^{1/2}}{(\text{Cov}(\tilde{q}_1, \tilde{q}_2))^{1/2}}.
\]

In the factor analysis estimator, the covariances between the individual

---

20 We could combine the factor analysis and IV approaches by constructing the index with some of the proxies and then instrumenting it using the remaining ones. This approach does not, however, solve the problems associated with using either of the methods separately.
proxy variables and the wage play a role only indirectly via the correlation between the quality index and the wage, whereas this estimator makes use of these covariances directly.

Now consider \( K > 2 \). We might hope that additional proxies would allow us to identify the entire system without a normalization, but this turns out not to be possible. To see why, consider the off-diagonal equations

\[
\text{Cov} \left( \ln (\tilde{w}), \tilde{q}_k \right) = \gamma \alpha_k \sigma_{\bar{c}}^2, \quad k = 1, 2, \ldots, K;
\]

\[
\text{Cov} \left( \tilde{q}_k, \tilde{q}_l \right) = \alpha_k \alpha_l \sigma_{\bar{c}}^2, \quad k, l = 1, 2, \ldots, K, \quad k \neq l. \tag{15}
\]

By way of contradiction, suppose that \((\hat{\alpha}, \hat{\gamma}, \hat{\sigma}^2)\) represents a unique solution to the system. The vector \((\hat{\alpha}/\hat{\gamma}, \hat{\gamma}/\hat{\gamma}, \hat{\sigma}^2)\), for an arbitrary \( c > 0 \), also solves the system. Hence, the solution is not unique, which contradicts the hypothesis. Of course, this result is hardly surprising; we have no data on \( Q^* \), so we are unable to identify its moments.

When we normalize the variance of \( Q^* \) to one, the system becomes overidentified for \( K > 2 \) and we can use optimally weighted GMM to estimate the system (see Wooldridge [2002a] for a discussion). The GMM estimator in this section avoids both the inconsistency associated with the factor analysis estimator and the strong assumptions about the scales of the proxy variables required to justify the IV estimator; for this reason, we strongly prefer it on econometric grounds. At the same time, we note that, unlike the factor analysis approach, it does not provide a handy quality ranking of colleges as a by-product.

Using the five covariances with the wage measure and the 10 covariances of the college quality proxies, we estimate \( \gamma = 0.043 \), with a standard error of 0.0164. Table 5 displays this estimate along with the estimated \( \alpha \)'s for each of the college quality measures. Because we have normalized the variance of \( Q^* \) to one, the larger the estimated \( \alpha_k \) for a measure, the smaller the noise (the variance of the corresponding \( \tilde{u}_k \)). Using this ranking, we see that the least noisy proxy for college quality is average SAT, which supports the frequent use of this variable in the literature. The next least noisy is the freshman retention rate, followed by average faculty salaries, the rejection rate, and the faculty-student ratio, where the last two are noisy indeed.

Finally, we can use the estimated \( \alpha_k \) from the GMM estimator in this section to rescale the simple IV estimates obtained in Section V.C using the relation in equation (12). That is, we can use our estimates of \( \alpha_k \) to retrieve the implicit estimates of \( \gamma \) from the simple IV estimates. These estimates appear in the final column of table 5. The rescaled estimates possess two interesting properties: they vary much less among themselves (from 0.037 to 0.048 rather than 0.059 to 0.107), and they look a lot more like the estimates from both the GMM estimator in this section and the quality index estimates of Section V.B.
Table 5
Scale Independent GMM Estimates of the Effect of College Quality, NLSY Men, 1989

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM Estimate</td>
<td>0.043*</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>(0.0164)*</td>
<td>(0.0164)*</td>
</tr>
<tr>
<td>Estimated αi from</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the GMM Estimator of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sec. V.D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty-student ratio</td>
<td>0.348</td>
<td>0.337</td>
</tr>
<tr>
<td>Rejection rate</td>
<td>0.480</td>
<td>0.437</td>
</tr>
<tr>
<td>Freshman retention rate</td>
<td>0.629</td>
<td>0.637</td>
</tr>
<tr>
<td>Mean SAT score</td>
<td>0.738</td>
<td>0.748</td>
</tr>
<tr>
<td>Mean faculty salaries</td>
<td>0.619</td>
<td>0.641</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Note.—N = 887. College quality measure is for last college attended as of 1989. See the note of table 2 for description of the content of the regressions. The dependent variable is the natural log of the respondent’s wage, defined as earnings in 1988 (the year prior to the 1989 survey) divided by hours in 1988.

* Significant at the 5% level in a one-tailed test.

2. Sensitivity Analyses

We performed three sensitivity analyses on our estimates. First, because of the sensitivity of standard GMM to the estimation of the covariance matrix documented by Altonji and Segal (1996), we calculated the equally weighted minimum distance estimator. This estimator yields an estimate of $\gamma = 0.042$, which differs from the optimally weighted GMM estimate by only 0.01.

Second, we calculated the GMM estimate using all possible observations to calculate each moment condition rather than using the subset of observations with valid values for all of the variables used in constructing the estimate. The benefit from this procedure comes from not throwing out information; the downside is that the variables are likely not missing at random, which is what is required for this procedure to produce consistent estimates. Compared to the sample of 887 observations with valid values for all of the variables, the number of observations used ranges from 911 for SAT scores and wages to 1,593 for faculty salaries and wages, where the 911 and 1,593 do not fully overlap. This wide variation in the observations utilized in each case provides plenty of scope for selection issues to arise. As a result, we do not put too much weight on the resulting estimate of $\gamma = 0.036$, but it does suggest the value of filling in the data to create a large sample with valid values for all of the variables.

Finally, we calculated the GMM estimate using 1998 wages rather than
1989 wages.\textsuperscript{21} This reduces the sample size to 707, and yields an estimate of $\gamma = 0.038$.\textsuperscript{22} The ratio of the GMM coefficient to the OLS coefficient equals $(0.043/0.037) = 1.162$ in 1998 and $(0.038/0.027) = 1.407$ in 1989.

VI. Lubotsky and Wittenberg Estimator

Lubotsky and Wittenberg (forthcoming) propose a simple estimator that provides a lower bound on $\gamma$. Their estimator relies on a model of measurement error that closely resembles our own but allows nonzero covariances between the measurement errors associated with different proxy variables, so that, in our notation, $\text{Cov}(\hat{u}_k, \hat{u}_l) \neq 0$. By relaxing the zero covariance assumption, they add $K - 1$ parameters but no new information to the system. This, in turn, implies that they can only identify the $\alpha_k$ up to a normalization and, even with the normalization, that they can only offer a lower bound on $\gamma$ rather than achieving point identification. We follow them in setting $\alpha_1 = 1$.

The Lubotsky-Wittenberg estimator includes simplicity among its virtues; in terms of our notation, it equals

$$\hat{\gamma}_{\text{LW}} = \sum_{j=1}^{K} \frac{\text{Cov}(\ln \hat{w}, \hat{q}_j)}{\text{Cov}(\ln \hat{w}, \hat{q}_j)^{\text{OLS}}_{ij}},$$

where $\hat{\gamma}_{\text{OLS}}$ denotes the estimated coefficient on the $j$th quality measure in a regression of $\ln w$ on $X$ and all of the quality measures. When $\gamma > 0$, Lubotsky and Wittenberg show that this estimator produces the greatest lower bound on the parameter $\gamma/\alpha_1$ of any linear combination of the quality measures.\textsuperscript{23} Like the college quality index estimator based on factor analysis presented in Section V.B, the Lubotsky-Wittenberg estimator reduces but does not eliminate the attenuation bias.

In table 6, we present estimates obtained by applying the Lubotsky-Wittenberg estimator to our data, normalizing with the mean SAT score (the measure that we estimate has the largest $\alpha$). We obtain an estimated lower bound of 0.040, a value that lies very close to the OLS estimate obtained using just the SAT score variable in table 1.

Table 6 also presents the results of an examination, via the bootstrap, of bias in the Lubotsky-Wittenberg estimator, as well as bootstrap confidence intervals. The bootstrap sampling distribution, though quite skewed to the right, provides little evidence of bias. The unadjusted IV estimate of $\gamma/\alpha_1$ in table 4 equals 0.064. If $\text{Cov}(\hat{u}_k, \hat{u}_l) = 0$, this suggests a moderate amount

\textsuperscript{21} In Black et al. (2005), we estimated $\gamma$ for each year of the data between 1989 and 1998 and could not reject the null hypothesis of equality over these years.

\textsuperscript{22} The OLS, factor analysis, and IV estimates using the 1998 data also resemble their counterparts from the 1989 data.

\textsuperscript{23} If $\gamma < 0$, then their estimator represents the least upper bound on $\gamma/\alpha_1$ of any linear combination of the quality measures.
Table 6
Lubotsky-Wittenberg Estimates of the Effect of College Quality, NLSY Men, 1989

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lubotsky-Wittenberg, normed to the SAT covariance</td>
<td>.040 (.0182)</td>
</tr>
<tr>
<td>Estimated bias of estimator</td>
<td>.006</td>
</tr>
<tr>
<td>One-sided 95% confidence interval</td>
<td>.019</td>
</tr>
<tr>
<td>Two-sided 95% confidence interval</td>
<td>[.015, .088]</td>
</tr>
<tr>
<td>Corresponding OLS estimate</td>
<td>.037 (.0172)</td>
</tr>
<tr>
<td>Corresponding IV estimate</td>
<td>.064 (.0267)</td>
</tr>
<tr>
<td>N</td>
<td>887</td>
</tr>
</tbody>
</table>


Note.—College quality measure is for last college attended as of 1989. See the note of table 2 for description of the contents of the regressions. The dependent variable is the natural log of the respondent’s wage, defined as earnings in 1988 (the year prior to the 1989 survey) divided by hours in 1988. Standard errors (in parentheses) and confidence intervals (in brackets) are based on 999 bootstrap replications.

of remaining attenuation bias in the Lubotsky-Wittenberg estimator. In contrast, if Cov(\(u_k, u_i\)) ≠ 0, it suggests a preponderance of negative covariances, as required for the IV estimator to have an upward bias.

In the absence of direct evidence as to whether Cov(\(u_k, u_i\)) = 0 or not, how can we assess the usefulness of the Lubotsky-Wittenberg estimator either in our application or in general? In our application, all of the IV estimates exceed the Lubotsky-Wittenberg estimate, many of them substantially. This indicates one or more of the following: a lot of attenuation in the Lubotsky-Wittenberg estimate (something consistent with the modest correlations among the quality measures in table 1), Cov(\(u_k, u_i\)) < 0, or the magnitude of Cov(\(u_k, u_i\)) is small relative to Var(\(u_k\)) and Var(\(u_i\)). As our prior puts heavy weight on Cov(\(u_k, u_i\)) ≥ 0, we doubt the second explanation. Either of the other two explanations suggests a preference for the GMM estimator from Section V.D, which avoids any attenuation bias but assumes Cov(\(u_k, u_i\)) = 0.

More generally, in our view, the Lubotsky-Wittenberg estimator represents a useful, though limited, robustness check. In particular, a finding that the magnitude of the Lubotsky-Wittenberg estimate exceeds the magnitude of the IV estimate(s) would provide reasonable evidence of Cov(\(u_k, u_i\)) > 0, would suggest caution in the application and interpretation of the other estimators discussed in this article, and would suggest the wisdom of relying (at least in part) on the lower bound represented by the Lubotsky-Wittenberg estimate.
VII. Conclusions

Our analysis provides a number of important substantive findings. First, we show the importance of thinking about the parameter of interest and of linking the choice of parameter of interest to the choice of estimator. Second, our results indicate that papers in the existing literature that seek to estimate parameter $P_2$ using a single proxy for latent college quality likely underestimate the labor market effects of college quality. Specifically, our GMM estimator, which builds on a generalization of the classical measurement error model and makes use of information on four additional proxies for college quality, suggests a downward bias of around 20% relative to using the SAT variable as a single proxy for quality. This is not a huge effect, but it is not a trivial one either; given the easy availability of additional proxies, there is little excuse not to use them.\footnote{Our analysis suggests that the quality variables commonly used in the primary and secondary school literature, such as class size (often measured at the school or district level and so with substantial error), teacher experience, and whether or not the teacher has an advanced degree constitute weak proxies for (unobserved) school quality. Moreover, to the extent that these variables covary with dimensions of school quality not included in the model, they overstate the effects of these variables, holding the others constant, which is the policy parameter of interest in these papers. This interpretation comports with the findings of Rivkin, Hanushek, and Kain (2005), who estimate a very large variance of teacher quality not accounted for by observable teacher characteristics.}

Third, our preferred GMM estimator not only allows us to identify the parameter of interest up to a normalization (the same normalization invoked by factor analysis) but also allows us to estimate the reliability of our measures of college quality. Interestingly, we found that the average SAT score was the single most reliable signal about college quality, which supports its wide use in this literature. Fourth, we find that, once we account for differences in the reliability of the various college quality measures, the IV estimates look both less different from one another and much more like the estimates from our preferred GMM estimator. Fifth, our application of the estimator in Lubotsky and Wittenberg (forthcoming) leads us to conclude that correlation among the measurement error components does not represent a major problem in this application.

In terms of methods, we prefer the GMM estimator laid out in Section V.D. Unlike the quality index in Section V.B, it does not suffer from attenuation bias. Unlike the simple IV estimator in Section V.C, it does not suffer from scaling problems. It does rely on the assumption of no correlation among the error terms, on the validity of which the Lubotsky-Wittenberg (forthcoming) estimator can shed some light. In cases where the researcher has an independent interest in obtaining a quality ranking, the quality index provides this but should incorporate as many quality measures as possible. The GMM estimator then provides a check on the
quality index estimates. We note, finally, that the methods outlined in this article have applicability to a rich variety of contexts beyond the one considered here.

Future work should examine a couple of potentially serious problems not addressed by the estimators considered here. First, we have maintained the (quite strong) assumption that \( \text{Cov}(Q^*, u_k) = 0 \). As discussed in Bound, Brown, and Mathiowetz (2001), both OLS and IV yield inconsistent estimates when \( \text{Cov}(Q^*, u_k) \neq 0 \). While the literature has made progress on this issue in the case of binary \( Q^* \) (see, e.g., Kane, Rouse, and Staiger 1999; Black, Berger, and Scott 2000; and Frazis and Loewenstein 2003), it has not made much headway for continuous \( Q^* \).

Second, as noted in Dale and Krueger (2002) and by many others, it seems unlikely that colleges have a single quality dimension. The quality of particular departments and programs often varies widely within a given university with the result that someone seeking to study labor economics may correctly have a very different quality ranking of colleges than someone seeking to study Etruscan poetry. In addition, match-specific issues may mean that different individuals seeking to study the same thing rate different colleges differently due to the nature of the learning environments (e.g., class size, teaching style, and presence or absence of potentially distracting amenities) they provide.

References


Neal, Derek, and William Johnson. 1996. The role of premarket factors


