Estimating the Causal Effects of Education on Wage Inequality Using IV Methods and Sample Selection Models

Stacey H. Chen, SUNY at Albany*
Shakeeb Khan, University of Rochester

August 25, 2005

Abstract

We propose instrumental variables and semiparametric estimators to solve the problem of selection bias in comparisons of wage inequality across schooling groups. These estimators provide flexible schemes for the identification of (conditional) average treatment effects (ATE) on various inequality measures, including conditional variance and interquantile spreads. Our latent index framework captures the effect of a binary schooling choice on inequality. We also show how symmetry assumptions for the joint distributions of error terms in the outcome and selection equations, along with kernel weighting schemes, can be used to identify the ATE on inequality. Asymptotic theory is derived for the proposed estimators; a simulation study indicates that both estimators perform well in finite samples. Using college proximity as an instrument for schooling, we find little evidence that college education increased the degree of wage inequality in 1976. This runs contrary to the conventional OLS results.

JEL Classification: C24, C14, C13.

Key Words: schooling choice, average treatment effect, quantile regression, instrumental variable method, average scale ratio.

*Corresponding author: Stacey H. Chen. Mailing address: BA111B 1400 Washington Ave, Albany, NY 12222. E-mail: schen@albany.edu.
1 INTRODUCTION

One of the most important topics in labor economics is the relationship between education and inequality. For example, a substantial part of increasing inequality in the U.S. in the 1980s can be explained by increases in the economic returns to schooling. At least as important are differences in wage inequality conditional on schooling. While wage differences among college graduates have increased since the 1980s, the within-group inequality among high school graduates has exhibited an inverted-U shape, rising sharply during the 1980s but falling since the 1990s. This phenomenon has been explored most recently by Autor, Katz and Kearney (2004) and Lemieux (2004), using the Current Population Survey, and by Angrist, Chernozhukov and Fernandez-Val (2004) using the U.S. Census. Since the educational gap in wage inequality is widening and the workforce is more and more educated, the current research attention has turned to determining whether the recent growth in wage inequality can be attributed to the increasing importance of education.

Most of the research on differences in the degree of wage inequality among high school- versus college-educated individuals presumes that education choices are taken to be exogenous. However, if the choice to attend college is endogenous – that is, related to an individual’s potential earnings – then differences in wage dispersion across education levels no longer provide a causal explanation. For example, consider an economy in which college education has no effect on the distribution of earnings—the mean and variance of earnings in the total population are unchanged. Individuals with higher earnings potential still may self-select into the college-educated group, generating a non-zero college wage premium. This would result in reduced wage inequality among the college-educated relative to an economy in which education levels are assigned randomly. As Heckman and Honore (1990) and Acemoglu (2002) have suggested, the problem with the conventional measure of within-group inequality is that self-selection in the data on education choices may truncate the wage distribution, causing the degree of inequality at each education level to be understated. In light of this problem, observed differences in the degree of earnings inequality across education levels do not necessarily capture the causal effect of education on the dispersion of potential earnings.

A large theoretical and empirical literature attempts to deal with the problem of selection bias in research on the economic returns to schooling. Most of this literature focuses on

More recently, progress has been made on the problem of estimating causal effects on distributions. For example, there is the Quantile Treatment Effects estimator developed by Abadie, Angrist and Imbens (2002) and Abadie’s (2002, 2003) approach to the estimation of causal effects on cumulative distribution functions. Both of these approaches use the local average treatment effect (LATE) framework for Instrumental Variables (IV; Imbens and Angrist 1994; Angrist, Imbens and Rubin 1996). Also, Chernozhukov and Hansen’s (2001, 2004, 2005) IV strategy for the estimation of causal effects on quantiles allows for multinomial treatments but relies on stronger distributional assumptions than does the LATE framework.

This paper uses a combination of IV methods and a semiparametric framework to develop and implement two estimators of the causal effect of schooling (or any similar dichotomous causal variable) on the dispersion of earnings outcomes. Unlike most previous estimators in the LATE framework, however, our target is the overall average causal effect on variance, not an average effect on compliers. Our initial approach is based on a modification of Abadie (2002, 2003), using an idea suggested by Angrist (2004). In particular, we propose a weighted IV scheme that recovers population average causal effects on variance, using a symmetry assumption. Our second approach invokes somewhat stronger assumptions to develop a quantile-based method of estimating average treatment effects on inequality. Both approaches rely on symmetry assumptions similar to those discussed by Heckman and Vytlacil (2000) and used by Songnian Chen (1999) to estimate mean effects.

In addition to outlining asymptotic properties and presenting the results of a small Monte Carlo study, we apply our estimators to the causal link between education and inequality; we use the same data and identification strategy as Card (1995) and Kling (2001). In this case the instrument is a potential student’s proximity to college. Distance instruments of this sort also have been used by Kane and Rouse (1993, 1995), Rouse (1995), Currie and Moretti (2003) and Cameron and Taber (2004). In contrast with naive comparisons that do not adjust for selection bias, our IV and semiparametric estimates result in little evidence
of a significantly higher degree of inequality for the college educated than for high school graduates in 1976. However, the proposed estimates are less precise than the conventional OLS estimates. This semiparametric result is contrary to the OLS estimates, which show a significant contrast in within-group inequality between college and high school. Our findings suggest that the conventional estimates of the impact of education on inequality may be biased because of the endogenous schooling choice.

The paper is organized as follows: Section 2 introduces the basic model, establishes the key identification conditions, and describes the proposed estimators and corresponding procedures. Section 3 characterizes the asymptotic distributions of the estimators. Section 4 explores the finite sample properties of the estimators through a small-scale simulation study. Section 5 applies the estimation methods to measure the effect of college education on wage inequality. Section 6 concludes.

2 IDENTIFICATION

The data consist of \( n \) observations, indexed with subscript \( i \). The continuously distributed outcome variable is denoted by \( y_i \), a binary choice or treatment status by \( d_i \), an instrument by \( z_i \) and a vector of covariates by \( x_i \). In the application to college education and wage inequality, \( y_i \) is log wage, \( d_i \) is college attendance, \( z_i \) is an indicator of college proximity (e.g. Card 1995) or a continuously distributed variable of college cost in county (e.g. Cameron and Taber 2004). The covariates contain a set of individual demographic characteristics.

The causal effects of interest are characterized by potential outcomes to describe the counterfactual states of the world. Potential outcomes are indexed by possible values of \( d_i \), denoted by \( y_{di} \). Hence \( y_{1i} \) and \( y_{0i} \) indicate the potential outcomes with and without treatment, respectively. Consider the following description of \( y_{di} \):

\[
\begin{align*}
y_{1i} & = \mu_1(x_i) + \sigma_1(x_i)\epsilon_{1i}, \\
y_{0i} & = \mu_0(x_i) + \sigma_0(x_i)\epsilon_{0i},
\end{align*}
\]

where \( \mu_1(\cdot) \) and \( \mu_0(\cdot) \) are the means of potential outcomes; \( \sigma_1(x_i) \) and \( \sigma_0(x_i) \) are covariate-specific scale parameters, measuring the degree of within-group inequality; \( \epsilon_{1i} \) and \( \epsilon_{0i} \) are error terms independent of \( x_i \), normalized to have zero mean and unit variance. Treatment
status is determined by a binary choice model:

\[ d_i = I[m(x_i, z_i) \geq \eta_i] \tag{2.3} \]

where \( I[\cdot] \) is an indicator function; the selection index function \( m(x_i, z_i) \) is measurable but otherwise unspecified; and \( \eta_i \) denotes the latent error normalized to have unit variance, independent of \( x_i \) and \( z_i \).

Our goal is to identify the conditional average treatment effect (ATE) on the scale parameters in the covariate-specific population—i.e. \( \sigma_1(x_i)/\sigma_0(x_i) \) either for specific covariate values, or alternatively, an average over a subset of the covariate distribution. We propose two identification approaches. The first is based on IV methods, estimating the variance in potential outcomes for each treatment status. The second is a quantile-based procedure, estimating the ratio of the interquantile spreads for the treated versus the untreated.

### 2.1 Variance-Based IV Estimator

Our first approach builds on the IV estimator developed by Imbens and Angrist (1994), Abadie (2002, 2003) and Angrist (2004). In this subsection, we consider the case where there is a binary instrument \( z_i \) available to researchers. The dependence between the instrument and the treatment is recognized by using the potential treatment indicator, denoted by \( d_{zi} \) given \( z_i = z \). For instance, \( z_i \) can be an indicator of college proximity as an instrument for schooling in wage regressions (see e.g. Card 1995). The subgroup of individuals with \( d_{1i} > d_{0i} \) (or equivalently, \( d_{0i} = 0 \) and \( d_{1i} = 1 \)) is defined as compliers, who would attend college if living nearby a college at the end of high school but would not attend otherwise. Given the laten index assignment model (2.3), we have

\[ d_{1i} = I\{m(x_i, 1) \geq \eta_i\}, \quad d_{0i} = I\{m(x_i, 0) \geq \eta_i\}, \]

and compliers are those with \( m(x_i, 1) \geq \eta_i > m(x_i, 0) \).

The IV estimation method proposed by Imbens and Angrist (1994) has provided estimators of the average treatment effect for compliers, i.e. the so-called local average treatment effect (LATE), \( E[y_{1i} - y_{0i} | d_{1i} > d_{0i}] \), derived by regressing \( y_i \) on \( d_i \) with \( z_i \) as an instrument. Abadie (2002, 2003) extended this result to arbitrary functions of potential outcomes, \( h(y_{1i}) \) and \( h(y_{0i}) \). His methods suggest that the IV estimators of the conditional expectation
functions of $h(y_{1i})$ and $h(y_{0i})$ can be written as follows:

\[
E[h(y_{1i})|d_{1i} > d_{0i}, x_i = x] = \frac{\sum_{i=1}^{n} h(y_{1i})d_{i}(z_{i} - \hat{z}(x))}{\sum_{i=1}^{n} d_{i}(z_{i} - \hat{z}(x))},
\]

\[
E[h(y_{0i})|d_{1i} > d_{0i}, x_i = x] = \frac{\sum_{i=1}^{n} h(y_{0i})(1 - d_{i})(z_{i} - \hat{z}(x))}{\sum_{i=1}^{n} (1 - d_{i})(z_{i} - \hat{z}(x))},
\]

where $\hat{z}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{z_{K(x_i-x)}}{\sum_{i=1}^{n} K(x_i-x)}$, letting $K(\cdot)$ be a kernel function. For the problem studied in this paper, working with first and second moments, i.e. $h(y_{i}) = y_{i}$ and $h(y_{i}) = y_{i}^{2}$, one can estimate the variance of potential outcomes by using the treatment status of compliers.

When $h(y_{i})$ equals $y_{i}$, the Abadie approach estimates the average $y_{1i}$ and $y_{0i}$ for compliers. Using the notation established in (2.1)-(2.3), these parameters are

\[
E[y_{1i}|x_i, d_{1i} > d_{0i}] = \mu_{1}(x_i) + \sigma_{1i}(x_i)E[\epsilon_{1i}|m(x_i, 1) \geq \eta_{i} > m(x_i, 0)],
\]

\[
E[y_{0i}|x_i, d_{1i} > d_{0i}] = \mu_{0}(x_i) + \sigma_{0i}(x_i)E[\epsilon_{0i}|m(x_i, 1) \geq \eta_{i} > m(x_i, 0)],
\]

Using the Abadie approach to estimate effects on first and second moments, we obtain estimates of the (covariate-specific) variance in potential outcomes for compliers:

\[
V[y_{1i}|x_i, d_{1i} > d_{0i}] = \sigma_{1i}^{2}(x_i)V[\epsilon_{1i}|m(x_i, 1) \geq \eta_{i} > m(x_i, 0)],
\]

\[
V[y_{0i}|x_i, d_{1i} > d_{0i}] = \sigma_{0i}^{2}(x_i)V[\epsilon_{0i}|m(x_i, 1) \geq \eta_{i} > m(x_i, 0)].
\]

While these parameters are of value for predicting effects on variance for those whose schooling status is changed by the proximity instrument, they need not have predictive value for randomly chosen individuals with characteristics $x_i$.

### 2.1.1 IV Estimation of Variance – ATE

To go from LATE-type estimates of effects on variance to the estimates for an entire covariate-specific subpopulation, we invoke a symmetry assumption. Heckman and Vytlacil (2000) and Angrist (2004) have noted that when latent error distributions are symmetric, a sufficient condition for LATE to be equal to ATE is a first stage that has the following symmetry property:

\[
P(d_{i} = 1|z_{i} = 1) + P(d_{i} = 1|z_{i} = 0) = 1.
\]

This indicates that in the first stage, the binary instrument switches the probability of selection (i.e. propensity score) from $p$ to $1 - p$. 

5
Yet there is no reason for such an equality to hold for a given data-generating process. Angrist (2004) suggested the use of covariates for forming what Angrist called a symmetric subsample, referring to the notion that a first stage shifting of the probability of treatment through the origin reflects a kind of “observable symmetry.” Formally, we construct the symmetry subsample by finding regressors $x_i = x$ such that the sum of their propensity scores equals one, i.e:

$$P(d_i = 1|z_i = 1, x_i = x) + P(d_i = 1|z_i = 0, x_i = x) = 1.$$ (2.6)

Searching for regressor values that allow the above equality to hold provides us with the causal effect of interest conditional on $x_i = x$. Furthermore, one can average over those regressor values to get an average treatment effect over that subset of the covariate distribution. To find such regressor values, we adopt the kernel weighted method, assigning large weights and in effect “finding and keeping” relevant regressors that enable the above equality to hold, similar to what was informally done in Angrist (2004).\(^1\)

The resulting weighted IV estimators of effects on first and second moments produce the population average treatment effects. These can therefore be combined to generate estimates of $\sigma^2_1(x_i)$ and $\sigma^2_0(x_i)$, which are the parameters of our identification target.

The following assumptions formalize this identification strategy:

**Assumption 1**

(i) **Symmetry:** $f_d(\epsilon_d, \eta) = f_d(-\epsilon_d, -\eta)$ for $d = 0, 1$.

(ii) **First-stage:** The set of regressor values $x_i = x$ that satisfies the symmetric subsample condition has positive measure.\(^2\) That is, $P(x : p(x, 0) + p(x, 1) = 1) > 0$, where $p(x, z) \equiv P(d_i = 1|z_i = z, x_i = x)$ for $z = 0, 1$ denotes the propensity score for given $x$ and $z$.

The properties of the symmetric subsample under Assumption 1 motivate a simple IV-type estimation procedure. The idea is to estimate the variance by treatment status for compliers given different regressor values, and then “extract” the population average of the variance by finding a symmetric subsample. This leads to a two-stage search procedure.

---

\(^1\)When estimating the causal effect of child birth on mothers wages, Angrist found the “teen mothers” covariate to get the equality to hold, based on estimates of the probabilities with a flexible Probit model.

\(^2\)This positive measure condition will be relaxed when we study the quantile procedure described in next section.
In the second step, we use kernel weights to construct the symmetric subsample using the first-stage estimates of propensity scores. As mentioned earlier, symmetric pairs are found by assigning large weights to observations whose propensity scores sum close to one.

Here we use a kernel weighting scheme similar to the matching methods developed by Powell (1989), Ahn and Powell (1993) and Songnian Chen (1999). The weighting scheme in our case gives more weight to candidate observations if the sum of their propensity scores is closer to one as described in (2.6). Formally, letting $\hat{p}(x_i, z_i)$ be the nonparametric estimates of the propensity score for $z_i = 0, 1$ using standard kernel or other nonparametric methods, assign a kernel weight to each observation at $x_i = x$:

$$\hat{\omega}(x_i) = K_{h_1 n}(1 - \hat{p}(x, 1) - \hat{p}(x, 0)),$$

(2.7)

where $K_{h_1 n}(\cdot)$ is a kernel function given a bandwidth $h_1 n$. $h_1 n$ converges to zero as the sample size $n$ increases, so in the limit we only keep pairs where the sum of the propensity scores exactly equals one.

In the estimation stage, we estimate the conditional variance for compliers based on (2.4) and (2.5) at each covariate value, denoted by $\hat{\sigma}_1^2(x)$ and $\hat{\sigma}_0^2(x)$.

Using the estimates of the variance for compliers, weighted by the above kernel function, the estimates of conditional variance by treatment status for the subset of the population whose regressor values satisfy the symmetry property can be estimated as:

$$\hat{\sigma}_d^2 = \frac{\sum_{i=1}^{n} \hat{\omega}(x_i) \hat{\sigma}_d^2(x_i)}{\sum_{i=1}^{n} \hat{\omega}(x_i)}.$$

Our proposed estimator takes a kernel weighted average of these values.

$$\hat{r}^2 = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}. $$

### 2.2 Quantile-Based Semiparametric Estimator

While simple and intuitive, the proposed IV estimator may perform poorly in finite samples if the distribution of potential outcomes have heavy tails, as the estimator was based on second moments. To deal with heavy-tailed distributions, we develop a quantile-based semiparametric estimator using additional assumptions. In addition to the symmetry conditions,

---

3The estimates $\hat{\sigma}_1^2(x)$ and $\hat{\sigma}_0^2(x)$ for compliers give a point estimate of the covariate-specific scale ratio for compliers, $\hat{r}_c = \hat{\sigma}_1^2(x)/\hat{\sigma}_0^2(x)$.
we assume that \( f_d(\epsilon_d, \eta) \) is the same for both treatment statuses \( d = 0, 1 \) (note that the conditional scale of outcomes still depends on treatment status). Then we show that the average scale ratio is identified. To see why, notice that the quantiles of \( y_i d \) are related to the quantiles of \( \epsilon_i d \) by equations (2.1) and (2.2), suggesting that for any interquantile spread from \( \tau \) to \( 1 - \tau \), denoted by \( IQ_\tau \), we have\(^4\)

\[
IQ_\tau(y_i | d_i = 1, x_i = x, z_i = z) = \sigma_1(x)IQ_\tau(\epsilon_{1i} | \eta_i \leq m(x, z), x, z), \tag{2.8}
\]

\[
IQ_\tau(y_i | d_i = 0, x_i = x, z_i = z) = \sigma_0(x)IQ_\tau(\epsilon_{0i} | \eta_i \geq m(x, z), x, z), \tag{2.9}
\]

for the treated and untreated groups, respectively. It is worth emphasizing that the instrument \( z_i \) can be discrete or continuous in the quantile-based estimation procedure, as we will see below.

The above equalities suggest identification of \( \sigma_1(x)/\sigma_0(x) \) when we invoke the following identification assumptions:

**Assumption 2**

(i) **Symmetry:** \( f_d(\epsilon_d, \eta) = f_d(-\epsilon_d, -\eta) \) for \( d = 0, 1 \).

(ii) **First-stage:** The set of pairs of observations that satisfies the symmetric subsample condition has positive density.

(iii) \( f_d(\epsilon_d, \eta) \) is the same for \( d = 0, 1 \).

By virtue of Assumption 2(i) and (iii), the following equalities will hold for any quantile \( \tau \in (0, 1) \) and real number \( m \):\(^5\)

\[
q_\tau(\epsilon_{1i} | \eta_i \leq m_i, x_i, z_i) = -q_{1-\tau}(\epsilon_{0i} | \eta_i \geq m_i, x_i, z_i),
\]

\[
q_{1-\tau}(\epsilon_{1i} | \eta_i \leq m_i, x_i, z_i) = -q_\tau(\epsilon_{0i} | \eta_i \geq m_i, x_i, z_i).
\]

The above immediately implies an equality of conditional interquantile spreads between symmetric quantiles \( \tau \) and \( 1 - \tau \):

\[
IQ_\tau(\epsilon_{1i} | \eta_i \leq m_i, x_i, z_i) = IQ_\tau(\epsilon_{0i} | \eta_i \geq m_i, x_i, z_i). \tag{2.10}
\]

\(^4\)Eq. (2.8) and (2.9) are suggested by a location-scale shift model for a given quantile \( \tau \), \( F_{y|x,d}^{-1}(\tau|x, d) = \mu_d(x) + \sigma_d(x)F_d^{-1}(\tau) \), where \( \mu_d(x) \) and \( \sigma_d(x) \) are location and scale parameters, respectively, for given treatment \( d \) and covariates \( x \).

\(^5\)To see how we derived these equalities, denoting \( w = (x, z) \), we have \( q_\tau(\epsilon_i | \eta_i \leq m, w) = q_{1-\tau}(-\epsilon_i | -\eta \leq m, w) \) given a real number \( m \). The first equality comes from the bivariate symmetry conditions, and the second equality is a result of the assumption that \( f_d(\epsilon_d, \eta) \) is the same for \( d = 0, 1 \).
This equality illustrates our strategy to identify the ratio of interest. Intuitively, given a pair of observations, if a variation in the instrument (or covariates) can switch the treatment status and change the selection index from \( m \) to \(-m\), the ratio of interquantile ranges can identify the parameter of interest. Note that under the symmetry assumption, any pair of observations satisfies the above condition must have their propensity scores \( p(x, z) \) sum to one. Precisely, for a pair \((i, j)\) where \( i \) selects and \( j \) does not, the symmetry assumption yields the following relation:

\[
m(x_i, z_i) = -m(x_j, z_j) \iff p(x_i, z_i) + p(x_j, z_j) = 1.
\]

The second equality is analogous to the first stage condition of the variance-based IV estimator in the case of a binary instrument (see (2.6)). This property suggests that we can use pairs of observations that satisfy the symmetry condition to identify the causal effect of interest. In addition, we note that the above relation applies for both cases with a binary or continuous instrument. Intuitively, a continuously distributed instrumental variable is more likely to provide “rich enough” support that will facilitate finding more pairs whose propensity scores sum to one.

The rest of this subsection translates the above identification result into a tractable estimation procedure. The quantile-based estimation procedure takes three stages. The first two stages involve searching symmetric pairs \((i, j)\) for \( i \) from the treated and for \( j \) from the untreated using the first-stage nonparametric estimates \( \hat{p}(x_i, z_i) \) and \( \hat{p}(x_j, z_j) \) of the propensity scores. To give a higher weight to pairs who have the sum of propensity scores closer to one, define the kernel weighing scheme as below:

\[
\hat{\omega}_{ij} = K_{h_{1n}}(1 - \hat{p}(x_i, z_i) - \hat{p}(x_j, z_j)),
\]

where \( K_{h_{1n}}(\cdot) \) is a kernel function for a given bandwidth \( h_{1n} \).

In the estimation stage, we calculate the interquantile spread between quantiles \( \tau \) and \( 1 - \tau \) based on the estimates of quantile functions for the treated and the untreated. Here, we estimate quantile functions using the local polynomial procedure proposed in Chaudhuri (1991a, 1991b), which provides pointwise estimates for nonparametric quantile functions. An advantage of working with the local polynomial procedure is that the quantile estimates can be derived nonparametrically, without assuming a linear parametric model for the outcome.
equation. In addition, the local polynomial procedure gives a local linear representation for estimation errors in interquantile functions, which facilitates the analysis of the asymptotic properties of the proposed estimator as the appendix shows.

We illustrate how the local polynomial procedure works in practice when estimating \( q_{\tau}(y_i|d_i = 1, (x_i, z_i) = w) \). We first construct a cell centered at \( w \), denoted by \( C_n(w) \). For any other covariates \( w_i \) in \( C_n(w) \), we have \( \|w - w_i\| \leq h_{2n} \), where \( \| \cdot \| \) denotes the Euclidean norm and \( h_{2n} \) is a bandwidth sequence used in the second stage, decreasing to zero as the sample size \( n \) increases. Our estimator for the quantile function of the treated group can be defined as:

\[
\hat{q}_{\tau}(y_i|d_i = 1, (x_i, z_i) = w) = \arg \min_{\mu} \frac{1}{n} \sum_i I[w_i \in C_n(w)]d_i \rho_{\tau}(y_i - \mu),
\]

(2.12)

where \( \rho_{\tau}(u) = |\tau - I[u < 0]|u \). The estimator for the quantile function of the untreated group can be defined similarly. The conditional interquantile spreads \( \hat{I}_Q_{\tau}(y_i|\cdot) \) therefore are derived for each treatment status.

Finally, the proposed estimator of the average scale ratio is formulated as the ratio of the weighted average of the interquantile functions for the treated versus the untreated groups, using the first-stage kernel weights \( \hat{\omega}_{ij} \) defined in (2.11). Precisely,

\[
\hat{r} = \frac{\sum_{i,j} d_i (1 - d_j) \hat{\omega}_{ij} \hat{I}_Q_{\tau}(y_i|d_i = 1, x_i, z_i)}{\sum_{i,j} d_i (1 - d_j) \hat{\omega}_{ij} \hat{I}_Q_{\tau}(y_j|d_j = 0, x_j, z_j)}.
\]

(2.13)

The estimation procedure has the disadvantage of requiring three semiparametric procedures (propensity score estimation, quantile function estimation, and kernel matching). This requires the selection of the three smoothing parameters, making implementation in finite samples difficult. Unfortunately, this is a consequence of the generality of the model considered, where both of the location functions \( m(\cdot), \mu_1(\cdot) \) and \( \mu_0(\cdot) \) were left unspecified. Imposing parametric conditions on one of these functions will reduce the number of smoothing parameters needed, making implementation easier in finite samples. For example, if we imposed \( m(w) = w' \delta \), existing estimators could be used to estimate \( \delta \) and propensity scores. Examples of such estimators include Han (1987) and Powell, Stock and Stoker (1989).\(^6\)

\(^6\)Furthermore, if we assume the scale functions only depend on treatment status (and not the regressors), we note the conditional interquantile ranges will only depend on (one-dimensional) propensity score functions, reducing the dimensionality of the problem.


3 LARGE SAMPLE BEHAVIOR

In this section, we explore the asymptotic properties of the proposed estimators under regularity conditions. The asymptotic properties of the variance-based and quantile-based estimators are given as follows. Both theorems are proved in the appendix, which also contains the regularity conditions the proofs are based on. For the variance IV estimator, our theorem is as follows.

**Theorem 1** Under Assumption 1 and if regularity conditions RS, RD, MF, PS, KM1, KM2, and RSS in the Appendix hold then our variance-based IV estimator has the following linear representation:

\[
\hat{\sigma}^2 - \frac{\sigma^2_1}{\sigma^2_0} = P_{SS}^{-1} \frac{1}{n} \sum_{i=1}^{n} \sigma^{-2}_0 (\psi_{12i} - 2\mu_1 \psi_{11i} + \tilde{\psi}_{1si}) \quad (3.14)
\]

- \[
\hat{\sigma}^2 - \frac{\sigma^2_1}{\sigma^2_0} (\psi_{01i} - 2\mu_0 \psi_{02i} + \tilde{\psi}_{0si}) + o_p(n^{-1/2}),
\]

where for \(d = 0, 1\) and \(k = 1, 2\), \(\psi_{dki}\) is of the form

\[
\omega_i = \frac{1}{E[w_i E[I[d_i = d]|z_i = 1, x_i] - E[I[d_i = d]|z_i = 0, x_i = x]]^{-1} \times
\]

\[
\left( \frac{y^k_i I[d_i = d]|z_i = 1, x_i] - E[y^k_i I[d_i = d]|z_i = 1, x_i]|E[I[d_i = d]|z_i = 1, x_i] - E[I[d_i = d]|z_i = 0, x_i = x]\right) -
\]

\[
w_i(E[y^k_i I[d_i = d]|z_i = 1, x_i] - E[I[d_i = d]|z_i = 0, x_i = x])^{-2} \times
\]

\[
\left( \frac{1 - E[I[d_i = d]|z_i = 1, x_i] + I[d_i = d]|z_i = 0, x_i = x]}{1 - E[I[d_i = d]|z_i = 0, x_i = x]} \right)
\]

where \(\mu_d\) above is \(E[w_i E[I[d_i = d]|y_i|d_{0i} > d_{0i}, x_i]]\)

and \(\tilde{\psi}_{dsi} = \omega_i(\sigma^2_{d}(x_i) - \sigma^2_{d}), \omega_i = I[p(x_i, 1) + p(x_i, 0) = 1]\) and \(P_{SS} = E[w_i]\).

The root-\(n\) consistency and asymptotic normality of the estimator follows from this linear representation.

Turning attention to the quantile estimator, while the regularity conditions for the quantile-based estimator are standard when compared to existing work (e.g. Ahn and Powell 1993, Chen and Khan 2003), they are still quite detailed, particularly as multiple semiparametric steps are involved. To ease notational burdens, letting \(w = (x, z)\), we impose the
parametric restriction \( m(w) = w'\delta \equiv v \) and assume that the scale parameters depend on treatment but not covariates; i.e. \( \sigma_1(x_i) = \sigma_1 \) and \( \sigma_0(x_i) = \sigma_0 \). In addition, define \( q_{r}^{(d)}(w) = q_{r}(y_{i}|d_{i} = d, (x_{i}, z_{i}) = w) \), \( \Delta q_{r}^{(d)}(w) = IQ_{r}(y_{i}|d_{i} = d, (x_{i}, z_{i}) = w) \) and \( \Delta q_{r}^{(d)}(w) = E[\Delta q_{r}^{(d)}(w)|v] \) for \( d = 0, 1 \). The asymptotic property of the quantile-based estimator of the average scale ratio is stated below.

**Theorem 2** Under Assumption 2 and regularity conditions H1,S0, RD2, S2, KH1 and I, if the selection equation estimator has the following linear representation, composed by \( \psi_{i}^{+} \):

\[
\hat{\delta} - \delta = \frac{1}{n} \sum_{i=1}^{n} \psi_{i}^{+} + o_{p}(n^{-1/2}),
\]

then we have

\[
\sqrt{n}(\hat{r} - r) \Rightarrow N(0, \Sigma_{0}^{-2}E[\psi_{i}^{-} + \mathcal{M}\psi_{i}^{+}]^2),
\]

where \( \Sigma_{0} = E[1 - p(v_i)]^2E[\Delta q_{r}^{(1)} f_{V}(-v_{i})] \), \( p(v) \) is the propensity score and \( f_{V}() \) is the density of \( v \). Let \( \psi_{i}^{-} = (1 - p(v_i))^2f_{V}(v_{i})f_{W}(w_{i})[d_{i}\phi_{1i} - (1 - d_{i})(1 + r)\phi_{0i}] \), where we denote \( \phi_{di} \) as \( f_{U_{2d}|W}(0|w_{i})^{-1}(I[y_{i} \leq q_{2}^{(d)}(w_{i})] - \tau_{2}) - f_{U_{1d}|W}(0|w_{i})^{-1}(I[y_{i} \leq q_{1}^{(d)}(w_{i})] - \tau_{1}) \) for \( d = 0, 1 \). Finally, define \( \mathcal{M} \) as

\[
E \left[ (1 - p(v_i)) \cdot [\mathcal{G}_{2}(v_{i}, v_{j})p(-v_{i})f_{V}(-v_{i}) + \mathcal{G}(v_{i}, v_{j})p(-v_{i})f_{V}(-v_{i}) + \mathcal{G}(v_{i}, v_{j})p(-v_{i})f_{V}(-v_{i})] \right],
\]

(3.17)

where \( \mathcal{G}(v_{i}, v_{j}) \equiv \int \int [\Delta q_{r}^{(1)}(w_{i}) - \Delta q_{r}^{(0)}(w_{i})(1 + r)] (w_{i} + w_{j})dF(w_{i}|v_{i})dF(w_{j}|v_{j}), \) and \( \mathcal{G}_{2}(\cdot, \cdot) \) is the partial derivative of \( \mathcal{G}(\cdot) \) with respect to its second argument.

# 4 MONTE CARLO STUDIES

In the previous section we explored the conditions under which the proposed estimators are well-behaved in a large sample. In this section we assess the finite sample performance of the estimators through Monte Carlo studies. For each Monte Carlo iteration, a sample is drawn with size \( n = 100, 200, 400 \) and \( 800 \). We use each sample to calculate the average scale ratio, and then iterate the entire process 801 times.

We use the same set of simulation designs to study the proposed estimation methods. In particular, we simulate the error terms, \( \eta \) and \( \epsilon_{d} \) for \( d = 0, 1 \), using various bivariate
distributions. The common regressor in both outcome and selection equations is normally
distributed. The instrument in the selection equation is either a normal or binomial random
variable. We set the scale function as $\sigma_1 = \exp(2)$ and $\sigma_0 = \exp(0)$. Thus the true value of the
average scale ratio approximately equals 7.389. The following two sections report the simula-
tion results for the proposed variance-based IV estimator and quantile-based semiparametric
procedure, respectively.

4.1 Simulation Results of the Variance-Based IV Estimator

The proposed IV estimator requires a binary instrument, for which we use a binomial random
indicator with a 50-50 chance of being zero or one. In the two-stage search procedure, we
first construct a symmetric subsample using the first-stage propensity scores. Observations
are included in the subsample if the corresponding propensity scores switch from $p$ to $1 - p$
for $p \in (0, 1)$ when the instrument changes from zero to one. In the estimation stage, we use
the subsample to estimate the variance in potential outcomes by treatment status. We adopt
a normal kernel in the estimation stage and a uniform kernel in searching the symmetric
subsample. For both kernels, we set bandwidths equal to $n^{-1/5}$.

Table 1 reports the results using four distributional designs, including bivariate normal
and bivariate Student-t with three different degrees of freedom, so as to examine the estima-
tor’s performance when the tails of bivariate distributions are increasingly thick. For each
design, we report four assessment measures: mean bias, median bias, root mean-square error
(RMSE) and mean absolute deviation (MAD). Results show that RMSE and MAD in the
bivariate normal design converge at the rate of root-n. The patterns of RMSE and MAD in
Student $t(10)$ and $t(5)$ are similar to those in the normal design. It is worth noting that the
performance of the variance-based estimator in Student-$t(10)$ and $t(5)$ is almost as good as
that of the normal design, even with a sample size as small as 100. In addition, for distribu-
tions with very thick tails, such as Student-$t(3)$, both RMSE and MAD still converge to zero
although the convergence is slower than in the case of normality.

An important remark is that the variance-based estimator is consistent only when the
bivariate distribution has finite second moments. When the moment parameters are not
well-defined, such as bivariate Cauchy distributions, the variance-based estimator is no longer
consistent. Overall, results suggest good finite-sample performance of the proposed variance-
based estimator when outcome distributions have finite second moments.

4.2 Simulation Results of the Quantile-Based Estimator

Unlike the variance-based estimator, the quantile-based estimator works well even when the distributions of the error terms have extremely thick tails. In particular, the quantile-based estimator is consistent under bivariate Cauchy distributions.

Deriving the quantile-based estimates requires a three-stage procedure. First, estimate the probability of treatment conditional on observed covariates and the instrument. Second, generate a kernel weight for each pair of observations from the treated versus the untreated groups, and assign higher weights to pairs whose sum of the probabilities of treatment is closer to one. Third, estimate conditional quantile regressions to calculate interquantile spreads between symmetric quantiles (e.g., 75-25 in our design). Finally, identify the average scale ratio by the ratio of the weighted average of interquantile spreads, using the kernel weights derived in the second stage. We use uniform kernels in the entire estimation procedure and set bandwidths as $n^{-1/5}$.

Using a continuous instrument (generated by a normal random variable), results in Part (a) of Table 2 show that RMSE and MAD converge at the rate of root-$n$ in various distributions, including the bivariate normal and bivariate Student-t(5). In particular, the bivariate normal and bivariate Student-t(5) generate very similar Monte Carlo results. As expected, the simulated sample in the case of bivariate Cauchy contains a number of outliers due to heavy tails, making the mean and median biases and the RMSE converge slowly or sometimes increase with the sample size. Nevertheless, the MAD measure in the Cauchy design still converges rapidly toward zero. Because MAD is the most relevant assessment for the quantile-based estimator in the table, the convergence of the MAD suggests that the proposed estimator performs reasonably well in small samples even when the distribution of the errors terms has fat tails.

Furthermore, to make a fair comparison between the variance-based and quantile-based estimators, we simulate the quantile estimator in Part (b) of Table 2 using the binary instrument used to generate Table 1. Results show that the variance-based estimator outperforms the quantile-based estimator because all the assessment measures are smaller in magnitude. Yet, the quantile-based estimator converges faster than the variance-based estimator, espe-
cially in distributions with thick tails. This reflects two fundamental differences between the proposed estimation methods. First, the quantile-based estimator is less sensitive to fat-tailed distributions and thus converges faster than the variance-based estimator does. Second, the way observations that satisfy symmetry conditions are selected differs in both procedures. Unlike the variance-based estimator, the quantile-based estimator includes pairs of observations whose probabilities of treatment comply with a change in the instrument for given covariates or with a change in the covariates given the instrument. For instance, if a variation in covariates can switch the probability of treatment from \( p \) to \( 1 - p \) within a pair of observations given a value of the instrument, this pair will be included to generate the quantile-based estimates but be excluded from the variance-based procedure. Because of this difference, the quantile-based estimator is more precise but computationally cumbersome than the variance-based method because of the extensive number of symmetry pairs.

As Parts (a) and (b) of Table 2 show, a continuous instrument does not necessarily generate better simulation results than a binary instrument does, nor the other way around. Two factors are in effect in this comparison. On the one hand, because semiparametric estimators usually work better with discrete regressors with finite support, estimators using a binary instrument are more likely to outperform the estimator using a continuous instrument. On the other hand, the use of a continuous instrument provides more variation in the sample and thus enhances the identification. Combining these two effects, our results show that the use of the binary instrument somewhat improves the performance of the quantile-based estimator in the bivariate Cauchy design, but not necessarily in other designs such as bivariate normal and bivariate Student-t.

5 CAUSAL EFFECTS OF SCHOOLING ON INEQUALITY

We apply the proposed estimation methods to a study of the causal effects of schooling on wage inequality. Economists have been interested in understanding the link between college education and wage inequality at least since Juhn, Murphy and Pierce (1993).\(^8\) Yet few empirical studies in the literature address the endogeneity of the schooling choice. Endogenous schooling choice may bias the estimated effect of education on inequality because ability bias

may truncate the wage distribution for each schooling group, causing the degree of inequality at each schooling level to be underestimated. In turn, the contrast in inequality between schooling levels (measured by the average scale ratio) may not necessarily reflect the causal effect of education on inequality in potential earnings. Parametric estimates by Chen (2004) suggest that truncation bias caused by selection in schooling can be considerable under the assumption of normality. The proposed estimators in our paper provide flexible ways to address the selection issues using weak distributional and functional form assumptions.

In what follows, we lay out an empirical model, describe the data and summarize the empirical results.

5.1 Determination of Wage and Wage Inequality

Consider the log hourly wage (or briefly “wage”) as the outcome variable \( y \) and college attendance as the treatment variable \( d \). Our empirical application focuses on the following wage regression model:

\[
y = \alpha d + \mu(x) + \sigma_d(x)\epsilon_d,
\]

(5.18)
given schooling choice \( d = 0 \) or \( 1 \). \( \alpha \) is the coefficient on college attendance. The covariate-specific location and scale parameters, \( \mu(x) \) and \( \sigma_d(x) \), are unknown functions. Following the inequality literature, we use \( \sigma_d(x) \) to measure within-group inequality. We use a latent index model to characterize the schooling choice \( d = I\{m(x, z) \geq \eta\} \), where \( m(\cdot) \) remains unspecified and \( z \) indicates the college proximity instrument. \( \eta \) and \( \epsilon_d \) are random variables, independent of \( x \), normalized to have unit variance. Our empirical study aims to identify the average scale ratio \( \sigma_1^2(x)/\sigma_0^2(x) \) using variance-based and quantile-based inequality measures.

As an instrument for schooling, we use the college proximity indicator, which has been suggested by Card (1995), Kane and Rouse (1993, 1995), Rouse (1995), Kling (2001) and Currie and Moretti (2003). Using the same data that generated Card’s (1995) and Kling’s (2001) results on the return to education, we estimate the effect of college attendance on wage inequality. Parametric first-stage estimates (based on Probit and linear probability regressions) indicate that college proximity is associated with a 5 percent to 7 percent higher

\[^9\]Blau and Kahn (1994) and Lemieux (1998) have adopted similar location-scale shift models to study wage inequality: the former studies the gender difference in inequality, supposedly having no problems of selectivity, the latter estimates the effect of unionization on wage inequality, assuming selection entirely dependent on the history of observables.
probability of college attendance with standard errors as small as .02. In the following section, we describe the instrument and the set of covariates in the regressions.

5.2 Data and Replicated IV Estimates of Returns to Schooling

The data in this paper are drawn from the *National Longitudinal Survey for Young Men* (NLSYM), which began with 5525 youths aged 14-24 in 1966 and followed them until 1981. The NLSYM provides detailed individual controls, including the respondent’s demographics, parental education, and proximity to a four-year college in the county of residence. The NLSYM sample used by Kling (2001), which replicated Card’s (1995) results, can be downloaded from the *Journal of Business and Economic Statistics* web site.

Our empirical studies begin by reproducing Card’s (1995) and Kling’s (2001) results on the return to years of schooling. We use Kling’s data and programs. These are shown in the first three columns in Part (a) of Table 4, where the return to years of schooling increases to 13 percent from 7 percent after the problem of selectivity bias of schooling is addressed. Note that Kling’s estimate in column (3) excludes the quadratic term for work experience, generating almost the same results as Card.

Because our empirical goal is to estimate the effect of college attendance on wage inequality, we focus here on the comparison between two subgroups: high school graduates (with 12 years of schooling) and the college educated (with more than 12 years of schooling). This excludes 497 respondents with less than 12 years of schooling from Kling’s sample. Our sample contains 992 high school graduates and 1521 respondents with some college education.

Descriptive statistics from our sample are reported in columns (3)-(6) of Table 3. Average log wages in the sample are about 1.75 for college and 1.64 for high school. Standard deviations in log wages are one percentage point higher for high school than for college. As Card and Kling have noted, men who were raised in local labor markets with a nearby four-year college have significantly higher levels of education. This differential persists even after controlling for regional and family background variables, as shown in their studies.

Our parametric benchmark uses fewer regional background variables than Card’s and Kling’s studies, because of the limitation of the sample size. Columns (4) and (5) of Table 4 show that IV and OLS results based on the subset of covariates are similar to Card’s and Kling’s when we control for the full set of covariates. Our benchmark includes the following
covariates: work experience, a Black indicator, residence in Southern states in 1976, residence in a metropolitan area in 1976 and 1966, and college attendance of both parents. Following Kling’s (2001) suggestion, we take work experience as an endogenous variable in our model, using age as an excluded instrument.

5.3 Estimation Results on Inequality

This section presents the estimates of the educational effect on wage inequality using the conventional parametric methods and the proposed semiparametric procedures. These estimates have important causal interpretations for the relation between college attendance and wage inequality if the assumed identification conditions are valid. All of the estimates are conditioned on the set of observables summarized in Table 3.

5.3.1 Results Using Conventional Methods

As a benchmark, OLS estimates in Column (1) of Table 5 show a significant contrast in residual inequality between college and high school. If schooling is randomly assigned, the OLS estimates suggest that the average variance ratio equals 1.15 and the average standard deviation ratio equals 1.07, similar to the unconditional statistics in Table 3. In our sample, less than 10 percent of the inequality among the college educated is captured by the variation in covariates, while about one quarter of the variation among high school graduates is explained by the covariates. These results are in line with Juhn, Murphy and Pierce (1993), who noted that inequality between 1963 and 1989 resulted primarily from the variation in unobserved heterogeneity rather than from observed characteristics.

The unexplained component of the variation can be caused by wage shocks or unobserved heterogeneity, both of which constitute wage inequality. Because wage shocks and unobserved heterogeneity can be correlated with the schooling choice, selection bias of schooling needs to be addressed when measuring the degree of inequality; otherwise a simple comparison of inequality between schooling levels may not provide a useful causal interpretation.

Parametric methods, such as Heckman’s two-stage least squares, correct for selection bias in schooling, assuming a specific joint distribution for the error terms in the model. In the first stage of the procedure, selectivity adjustments (e.g. the inverse Mills ratios under the assumption of normality) are generated to identify the causal effect of education on
average wages. Less emphasized in the literature, however, is the adjustment for truncation required to estimate the causal effect of schooling on inequality. In Heckman’s framework, the truncation adjustment is determined by a function of first-stage estimates, including probabilities of treatment and inverse Mills ratios. As discussed earlier, truncation bias arises because latent ability and taste for education are correlated with wage residuals, making the observed wage distributions by schooling choice more concentrated than those generated by random assignment.10

Indeed, adjusting for truncation makes a considerable difference, as the Heckman two-stage estimates in column (2) show. The truncation adjustment increases the variance-based inequality measures by about 30 percent for college earnings and by about 40 percent for high school. In turn, the variance ratio decreases to 1.05 from 1.15, and the ratio of standard deviations decreases to 1.02 from 1.07. In summary, the parametric results suggest that conventional estimates omitting truncation adjustments overstate the educational differences in wage inequality in 1976. After accounting for selection and truncation biases, the educational difference in the degree of wage inequality is not statistically significant in the parametric models.

5.3.2 Semiparametric Estimates of the Causal Effect of Education on Inequality

The proposed variance- and quantile-based semiparametric estimators correct for possible truncation bias under symmetry conditions. Implementing the estimators requires first-stage estimates of propensity scores in order to select observations that satisfy symmetry conditions. As noted earlier, one important difference between these two estimators is in the way the observations that satisfy symmetry conditions are selected. In the variance-based estimation procedure, observations are selected if a change in the college-proximity indicator switches the propensity scores from \( p \) to \( 1 - p \) for \( p \in (0, 1) \). This creates what Angrist (2004) called a symmetric subsample, which accounts for 45 percent to 64 percent of observations in our sample, as shown in Part (a) of Table 6. In contrast, the quantile-based estimation procedure

---

10 Consider the conditional mean function for the college group, \( E[y|d = 1, x, z] = \alpha + \mu(x) + \beta_\lambda \lambda(x, z) \), where \( \lambda(x, z) = E[\epsilon_1|\eta \geq m(x, z)] \) is the inverse Mills ratio, the selectivity adjustment for mean effects. It is worth noting that the estimated residual variance \( V[y|d = 1, x, z] = \sigma_1^2(x) \delta(x, z) \) understates the conditional variance in potential outcomes \( \sigma_1^2(x) \) because \( \delta(x, z) = V[\epsilon_1|\eta \geq m(x, z)] \) ranges between 0 and 1. \( \delta(x, z) \) can be viewed as an adjustment for distributional truncation due to selection. Column (1) of Table 5 reports the residual variance without the truncation adjustment, and Column (2) reports the conditional variance in potential outcomes after the adjustment.
selects pairs of observations from college and high school if their sum of propensity scores equals one. This selection scheme creates an extensive sample, extracted from a collection of all the possible \(992 \times 1,521 = 1,508,832\) matches from college and high school. Part (b) shows that roughly between 43 percent and 67 percent of the matches are selected to identify the average interquantile treatment effect of college education. As expected, because of the extensive sample size, the quantile-based procedure provides much more precise estimates than the variance-based method.

We use uniform kernel functions to estimate propensity scores and construct pairwise matching kernel weights. Notably, the estimation results are robust for various specifications of kernel functions but are sensitive to the choice of the bandwidth. In particular, the number of observations selected by symmetry conditions increases with bandwidth for both estimators. This can be seen in Table 6, which reports results using \(n^{-1/5}, n^{-1/6}\) and \(n^{-1/7}\) bandwidths. For instance, when the size of the bandwidth increases by approximately 56 percent from \(n^{-1/5}\) to \(n^{-1/7}\) given \(n = 2513\), the quantile-based estimator selects up to a quarter more pairs of observations that satisfy symmetry conditions.

In addition to the distinct selection schemes, the proposed estimators differ in their ability to identify the degree of inequality for each level of schooling. The variance-based estimation method provides identification for the marginal distribution of potential earnings by schooling levels. In contrast, the quantile-based estimation procedure can identify the interquantile treatment effect but not the interquantile spreads of potential earnings for a given schooling level. The interquantile spreads generated in the procedure, using the method described in (2.12), have no causal interpretation although they are important for the identification of the average scale ratio. Thus, estimates of interquantile spreads by schooling levels are omitted in Part (b) of Table 6.

Contrary to the OLS benchmark, the proposed semiparametric estimates produce little evidence for the conclusion that the college educated experienced a higher degree of inequality than high school graduates in 1976, although the proposed estimates are less precise than the benchmark. This result is in line with the full-parametric estimates presented earlier. In particular, the proposed variance-based estimates cannot be distinguished from full-parametric estimation results because of extremely wide confidence intervals. On the other hand, the
quantile-based estimates are much more precise than the variance-based estimates because of the large number of symmetric pairs. The estimation results based on quantile regressions show no significant impact of college education on the 60-40 and 75-25 interquantile spreads, but these do indicate a negative impact on the 90-10 interquantile spread for the $n^{-1/5}$ bandwidth. When the bandwidth increases to $n^{-1/6}$ or $n^{-1/7}$, the educational contrast in the 90-10 interquantile spread is no longer statistically significant, suggesting that the quantile-based estimates are sensitive to the choice of bandwidths in the kernel functions.

In summary, as with Heckman’s two-stage procedure, the semiparametric estimates show little evidence of the causal impact of college education on wage inequality in 1976, except for a certain range of bandwidths. These results are contrary to the conventional OLS estimates, which generally indicate a positive and significant association between college education and wage inequality.

6 CONCLUDING REMARKS

This paper proposes IV-type and semiparametric estimators to test whether the contrast in wage inequality between the college educated and high school graduates is statistically significant. Using weak distributional and functional form assumptions, we provide two ways to assess the causal effect of schooling on wage inequality. One is based on conditional variance functions, building on the previous LATE results. The other is based on conditional interquantile spreads, using techniques of quantile regressions. A key ingredient of the identification strategy for both estimators is the kernel weighting scheme that selects observations satisfying symmetry conditions. For the variance-based estimator, the observations are selected if a binary instrument switches the probability of treatment symmetrically around zero. For the quantile-based estimator, college-high school pairs are assigned positive weights if the sum of their probabilities of treatment is close enough to one.

The simulation studies suggest that the proposed estimators perform well in finite samples with the number of observations as small as 100. Both estimators converge at the rate of root-n if the instrument is binary and if the bivariate distribution has a finite second moment. The variance-based procedure is computationally simpler and faster than the quantile-based estimator. Furthermore, in the case where the instrument is continuous or the bivariate
distribution has no finite second moment, the variance-based estimator does not apply, but the quantile-based estimator does, and converges at the rate of root-$n$.

We apply the proposed semiparametric procedures to the NLSYM66 sample in 1976, which was used by Card (1995) and Kling (2001) to instrument schooling with a college proximity indicator. After replicating their well-known results about the causal effect of schooling on average wages, we use the same sample to estimate the causal effect of college education on inequality. Our empirical results show that the decision to invest in college education did not increase the degree of wage inequality among workers in 1976. This result is contrary to the OLS benchmark, although the proposed estimates are less imprecise because of the limitation of the sample size. In light of our empirical results, we believe that the endogeneity issues caused by schooling choice can be important in estimating the educational contrast in inequality, especially in large samples such as the U.S. Census. As previous work by Autor, Katz and Kearney (2004) and Lemieux (2004) has noted, it is important to determine whether the recent growth of wage inequality can be explained by the growing contrast in inequality between the college educated and high school graduates. The semiparametric methods proposed in the current paper can potentially help us reassess the role of college education in rising inequality during the recent decades. We leave this empirical task for future research.

Although the proposed estimators prove consistent, possible improvements in efficiency could be achieved in several ways. For instance, while the current paper demonstrates how symmetry conditions identify the ATE on scale parameters, Angrist (2004), Songnian Chen (1999) and Heckman and Vytlacil (2000) have provided methods for estimating the ATE on location parameters using similar identification conditions. If the treatment effects on both location and scale parameters can be estimated simultaneously, then the efficiency of the proposed estimators can be improved with a combined estimation procedure. In addition, we note that the proposed quantile-based estimator over-identifies the average scale ratio for multiple pairs of symmetric quantiles (see eq. (2.13)). This suggests that the degree of efficiency can be improved further by averaging the proposed estimates over various pairs of symmetric quantiles. We leave the potential improvement in efficiency to future studies.
ACKNOWLEDGEMENTS

We would like to thank Josh Angrist, Jonah Gelbach and Ed Vytlacil for valuable suggestions. Thanks to seminar participants at MIT Economics, Queens University, University of Toronto, and the 2004 Evaluation Conference at the Centre for European Economic Research for helpful comments. All errors are ours.

APPENDIX A: REGULARITY CONDITIONS AND PROOFS

The distribution theory of our proposed IV estimator is based on the following regularity conditions:

**Assumption RS** (Random sampling) The vector \((y_i, z_i, d_i, x_i')\) is i.i.d.

**Assumption RD** (Regressor distribution) The regressor vector \(x_i\) has support which is a compact subset of \(\mathbb{R}^k\). \(x_i\) may have discrete and continuous components, and we let \(k_c\) denote the number of continuous components. We assume the conditional density function of the continuous components given the discrete components is continuously differentiable of order \(p\), where \(p > 5k_c/2\).

**Assumption MF** (Moment functions) Define the functions \(m_{kl}(\cdot) = E[y_i | x_i, d_i = m, z_i = 1]\). Then these functions are \(p\) times continuously differentiable.

**Assumption PS** (Propensity score functions) Define the functions \(p_m(\cdot) = E[d_i | x_i = \cdot, z_i = m]\) \(m = 0, 1\). Then these functions are \(p\) times continuously differentiable.

**Assumption KM1** The kernel function used to select which regressor values satisfy the symmetric subsample property is twice continuously differentiable with bounded derivatives. The bandwidth used in this stage, denoted here as \(h_m\), satisfies \(\sqrt{n}h_m \to 0\).

**Assumption KM2** The kernel function \(k(\cdot)\) used to estimate the propensity score and moment functions has bounded support, is \(p\) times continuously differentiable, and is of ”order” \(p\), i.e.

\[
\int k(u)u^l du = 0 \quad l = 1, 2, \ldots, p - 1
\]
The bandwidth used here, denoted by $h_n$, satisfies $nh_n^k \to \infty$, and $\sqrt{nh_n^p} \to 0$.

**Assumption RSS** The set of regressor values that satisfies the symmetric subsample condition has positive measure. That is,

$$P(x_i : p_0(x_i) + p_1(x_i) = 1) \equiv P_{SS} > 0$$

With these regularity conditions we will establish the following theorem:

**Theorem 3** Under the above Assumptions, our variance IV estimator has the following linear representation:

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2} - \frac{\sigma_1^2}{\sigma_0^2} = P_{SS}^{-1} \frac{1}{n} \sum_{i=1}^{n} \sigma_0^{-2} (\psi_{1i} + \psi_{12i}) - \frac{\sigma_0^2}{\sigma_0^2} (\psi_{01i} + \psi_{02i}) + o_p(n^{-1/2}) \quad (A-1)$$

where recall $\sigma_1^2$ here denotes $E[\sigma_1^2(x_i)|\omega_i = 1]$ and $\sigma_0^2 = E[\sigma_0^2(x_i)|\omega_i = 1]$, and the terms $\psi_{11i}, \psi_{12i}, \psi_{01i}, \psi_{02i}$ are defined as in the text in the statement of the theorem.

**Proof:** Turning attention to the proof we will derive the limiting distribution theory for $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}$ to which we can apply the delta method to get the distribution theory for the scale ratio- i.e. the ratio of standard deviations.

Note we can linearize the ratio as in e.g. page 2204 of Newey and McFadden. Therefore, it will suffice to derive linear representations of $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}$ and $\frac{\hat{\sigma}_0^2}{\hat{\sigma}_0^2}$. We will show the arguments for the former noting analogous results will follow for the latter.

We will establish a linear representation for

$$\sigma_1^2 - \sigma_1^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_i (\sigma_{1i}^2 - \sigma_{1i}^2) \quad (A-2)$$

Note by a mean value expansion of the matching kernel function of estimated propensity scores around true propensity scores, permitted by Assumption K1, and the uniform convergence of the propensity score estimators, which will follow from Assumptions PS and KM2, (see, e.g. Lemma 8.10 in Newey and McFadden(1994)), we may conclude that
\[
\frac{1}{n} \sum_{i=1}^{n} \omega_i \to P_{SS}
\]

Next, we turn attention to the linear term. Define

\[
\sigma^2_{1i} = E[d_i y_i^2 | x_i, d_{1i} > d_{0i}] - (E[d_i y_i | x_i, d_{1i} > d_{0i}])^2
\]

We will add and subtract \(\sigma^2_{1i}\) from the numerator term in (A-2), so we can derive linear representations for

\[
\frac{1}{n} \sum_{i=1}^{n} \omega_i (\hat{\sigma}^2_{1i} - \sigma^2_{1i})
\]

and

\[
\frac{1}{n} \sum_{i=1}^{n} \omega_i (\sigma^2_{1i} - \sigma^2_{1i})
\]

Turning attention to the first term (A-3), here we will establish a linear representation for

\[
\frac{1}{n} \sum_{i=1}^{n} \omega_i (\hat{\sigma}^2_{1i} - \sigma^2_{1i})
\]

Here we will only formally establish a linear representation for the second moment of the outcome variable component of the variance, noting the square of the first moment term can be handled very similarly. Recall the term \(E[y_i^2 | x_i, d_{1i} > d_{0i}]\) and its kernel estimator each involve the ratio of differences. Our first step will be to linearize the ratio. To ease notation, let the true values of \(E[y_i^2 | x_i, d_{1i} > d_{0i}]\) be denoted as \(\frac{a_1-a_2}{b_1-b_2}\), where \(a_1, a_2, b_1, b_2\) denote \(E[y_i d_i | z_i = 1, x_i], E[y_i d_i | z_i = 0, x_i], E[d_i | z_i = 1, x_i], E[d_i | z_i = 0, x_i]\) respectively. Let the kernel estimator be denoted as \(\frac{\hat{a}_1-\hat{a}_2}{\hat{b}_1-\hat{b}_2}\). By the fourth root consistency of the kernel estimators, which follows from Assumptions RS, RD, MF, PS, KM2 (see Chen and Khan(2003)) the linearization of ratio implies we can derive linear representations for

\[
\frac{1}{n} \sum_{i=1}^{n} \omega_i (b_1 - b_2)^{-1}(\hat{a}_1 - a_1 - \hat{a}_2 - a_2) - \frac{1}{n} \sum_{i=1}^{n} \omega_i (a_1 - a_2)(b_1 - b_2)^{-2}(\hat{b}_1 - b_1 - \hat{b}_2 - b_2)
\]

since the remainder term is \(o_p(n^{-1/2})\).

Here, we will focus on the term

\[
\frac{1}{n} \sum_{i=1}^{n} \omega_i (b_1 - b_2)^{-1}(\hat{a}_1 - a_1)
\]

(A-7)
as similar terms may be used for the other components. Here, we notice that \(\hat{a}_1\) is simply a kernel estimator for the regression function \(E[y_i d_i | z_i = 1, x_i = x_i]\). Consequently, we can apply results from Newey and McFadden (1994) or Chen and Khan (2003) to represent

\[ w_i (b_1 - b_2)^{-1} (\hat{a}_1 - a_1) \]

as (now expressing the definitions of \(a_1, a_2, b_1, b_2\))

\[ \frac{1}{n} \sum_{i=1}^{n} \omega_i (E[d_i | z_i = 1, x_i] - E[d_i | z_i = 0, x_i])^{-1} (y_i^2 d_i z_i - E[y_i^2 d_i | z_i = 1, x_i] + y_i^2 d_i (1 - z_i) - E[y_i^2 d_i | z_i = 0, x_i]) + o_p(n^{-1/2}) \]

which takes care of the first term in the linearization in (A-6). The second term is of the form:

\[ \frac{1}{n} \sum_{i=1}^{n} \omega_i (E[y_i^2 d_i | z_i = 1, x_i] (E[d_i | z_i = 1, x_i] - E[d_i | z_i = 0, x_i])^{-2} (d_i z_i - E[d_i | z_i = 1, x_i] + d_i (1 - z_i) - E[d_i | z_i = 0, x_i]) + o_p(n^{-1/2}) \]

and subtracting (A-10) from (A-9) establishes the form of the linearization for the second moment of the outcome variable. We will denote the term in the resulting summation (excluding \(\omega_i\)) by \(\psi_{12i}\). Turning attention to the square of the first moment, the linear representation for the first moment would be the same as above simply replacing \(y_i^2\) with \(y_i\). Let \(\psi_{11i}\) denote this term. Now let \(\mu_1\) denote \(E[\omega | E[d_i y_i | d_{1i} > d_{0i}, x_i]]\) then the linear representation for the square of the first moment can be denoted by

\[ \frac{1}{n} \sum_{i=1}^{n} 2\omega_i \mu_0 \psi_{11i} + o_p(n^{-1/2}) \]

which is a straightforward application of the delta method.

Now, for (A-4), we note if we replace \(\hat{\omega}_i\) with \(\omega_i\) the expression inside the summation is mean 0 with variance \(E[\sigma_1^4(x_i) | \omega_i] P_{SS} - P_{SS} \sigma_1^4\). Let \(\tilde{\psi}_{1si}\) denote the random variable inside the summation in (A-4) that results when we replace \(\hat{\omega}_i\) with \(\omega_i\). Therefore, expanding \(\hat{\omega}_i\) around kernel weights with the true propensity scores, and noting the linear term is asymptotically negligible (see, e.g. Khan and Powell (2001) Lemma A.9 for technical details in a similar setup), we can represent (A-4) as

\[ \frac{1}{n} \sum_{i=1}^{n} \omega_i \tilde{\psi}_{1si} + o_p(n^{-1/2}) \]
Collecting all our results we can conclude that we have the following linear representation for the variance of the treated group.

\[
\hat{\sigma}^2_1 - \sigma^2_1 = \frac{1}{n} \sum_{i=1}^{n} \omega_i (\psi_{12i} - 2\mu_0 \psi_{11i} + \tilde{\psi}_{1si}) + o_p(n^{-1/2}) \quad \text{(A-13)}
\]

Note we can derive an analogous linear representation for \(\hat{\sigma}^2_0 - \sigma^2_0\) with analogous terms in the summation, which we will denote here by \(\psi_{01i}, \psi_{02i}, \tilde{\psi}_{0si}\).

Next linearize the ratio to conclude

\[
\frac{\hat{\sigma}^2_1 - \sigma^2_1}{\hat{\sigma}^2_0 - \sigma^2_0} = P_{SS}^{-1} \frac{1}{n} \sum_{i=1}^{n} \sigma_0^{-2} (\psi_{12i} - 2\mu_1 \psi_{11i} + \tilde{\psi}_{1si}) - \frac{\sigma_1^2}{\sigma_0^2} (\psi_{01i} - 2\mu_0 \psi_{02i} + \tilde{\psi}_{0si}) + o_p(n^{-1/2}) \quad \text{(A-14)}
\]

where \(\mu_0\) above is \(E[\omega_i E[(1 - d_{1i})y_{1i}|d_{1i} > d_{0i}, x_i]]\).

This completes the proof for the ratio of variances- as mentioned to this we can apply the delta method to obtain a linear representation for the scale ratio.

**Regularity Conditions of Theorem 2**: Let \(h_{0n}\) and \(h_{1n}\) denote the bandwidths for the selection equation estimation and the pairwise matching kernel weighing scheme in the first stage, and let \(h_{2n}\) denote the bandwidth for the linear polynomial quantile regressions in the second stage. The regularity conditions for Theorem 2 is summarized below.

**Assumption I** (Identification) \(\Sigma_0 > 0\).

We next impose conditions on the kernel function used to match propensity score values and its bandwidth sequence:

**Assumption KH1** The kernel function \(K_{1n}(\cdot)\) is assumed to have the following properties:

i) \(K_{1n}(\cdot)\) is twice continuously differentiable with a bounded second derivative and has a compact support; (ii) symmetric about zero; and (iii) a fourth-order kernel with \(\int u^l K_{1n}(u)du = 0\) for \(l = 1, 2, 3\) and \(\int u^4 K_{1n}(u)du \neq 0\). The bandwidth sequence \(h_{1n}\) is of the form: \(h_{1n} = c_1 n^{-\gamma_1}\), where \(c_1\) is a constant and \(\gamma_1 \in (\frac{1}{5}, \frac{1}{6})\).
The following assumption characterizes the smoothness of the density of and the conditional expectation functions of the selection index:

**Assumption S0** The function $f_V(\cdot)$ has an order of differentiability of four, with the fourth-order derivative bounded.

We next impose three conditions associated with the estimation of interquartile spreads. This involves smoothness assumptions on the conditional quantile functions and on the distributions of $w_i = (x_i, z_i)$ and the residuals associated with the quantile functions. For notational convenience, we describe the conditions in terms of $w$, whose support is denoted by $\mathcal{W}$.

**Assumption RD2** (Distribution of regressors and instruments) The vector $w$ can be decomposed as $w = (w^{(c)}, w^{(ds)})'$ where the $k_c$-dimensional vector $w^{(c)}$ is continuously distributed, and the $k_{ds}$-dimensional vector $w^{(ds)}$ is discretely distributed. Letting $f_{W^{(c)}|W^{(ds)}}(\cdot|w^{(ds)})$ denote the conditional density function of $w^{(c)}$, we assume it is bounded away from zero and is Lipschitz continuous on $\mathcal{W}$. Letting $f_{W^{(ds)}}(\cdot)$ denote the mass function of $w^{(ds)}$, we assume that there is a finite number of mass points on $\mathcal{W}$. Finally, we let $f_{W}(\cdot)$ denote $f_{W^{(c)}|W^{(ds)}}(\cdot|\cdot)f_{W^{(ds)}}(\cdot)$.  

**Assumption S2** (Smoothness of conditional quantile functions) 

S2.1 The polynomial used for the second-stage quantile function estimators is of order $m$. 

S2.2 For all values of $w^{(ds)}$, the quantile functions $q_{\tau_1}^{(d)}(\cdot)$ and $q_{\tau_2}^{(d)}(\cdot)$, $d = 0,1$ are bounded and $m$ times continuously differentiable with bounded $m^{th}$ derivatives with respect to $w^{(c)}$ on $\mathcal{W}$.

**Assumption H1** (Second-stage bandwidth sequence for interquartile spread estimation). 

The bandwidth sequence used to estimate the conditional interquantile spread is of the form: $h_{2n} = c_2 n^{-\gamma_2}$, where $c_2$ is a constant, and $\gamma_2 \in \left(\frac{\gamma_1 + 1}{m}, \frac{1 - 4 \gamma_1}{3 k_c}\right)$, where $\gamma_1$, $m$ and $k_c$ are given in Assumptions KH1, S2 and RD2 respectively.

**Proof of Theorem 2:** The arguments used to derive the limiting distribution theory are very similar to those used in Chen and Khan (2003), hereafter referred to as CK. We thus
only provide a sketch of the main arguments, referring readers interested in technical details to CK.

We note we can write \( \hat{r} = \frac{\hat{\Sigma}_1}{\hat{\Sigma}_0} \), where
\[
\hat{\Sigma}_1 = \frac{1}{n(n-1)} \sum_{i \neq j} d_j(1 - d_i) \hat{\omega}_{ij} \Delta \hat{q}_r^{(1)}(w_i),
\]
\[
\hat{\Sigma}_0 = \frac{1}{n(n-1)} \sum_{i \neq j} d_j(1 - d_i) \hat{\omega}_{ij} \Delta \hat{q}_r^{(0)}(w_i).
\]

Our proof strategy will be to linearize the ratio of estimators as we did with the previous estimator. Our proof strategy is to establish the probability limit of the denominator and establish a linear representation for the numerator. The probability limit of the denominator follows from similar arguments used in proving Lemmas A.5 and A.6 in CK:
\[
\hat{\Sigma}_0 \xrightarrow{p} \Sigma_0,
\]

Turning attention to \( \hat{\Sigma}_1 - r\hat{\Sigma}_0 \), consider an expansion of \( \hat{\omega}_{ij} \) around \( \omega_{ij} = h^{-1}_n K_2 \left( \left( w'_i \delta_0 + w'_j \delta_0 \right) / h_2 n \right) \).

After using this expansion, \( \hat{\Sigma}_1 \) equals:
\[
\frac{1}{n(n-1)} \sum_{i \neq j} d_j(1 - d_i) \omega_{ij} \left[ \Delta \hat{q}_r^{(1)}(w_i) - r \Delta \hat{q}_r^{(0)}(w_i) \right].
\]

We note that if we replace \( \Delta \hat{q}_r^{(1)}(w_i), \Delta \hat{q}_r^{(0)}(w_i) \) with \( \Delta q_r^{(1)}(w_i), \Delta q_r^{(0)}(w_i) \) in the above expression, the term is \( o_p(\frac{n}{2}) \) by arguments similar to those used in the steps used to prove Lemma A.4 in CK. We can thus work with:
\[
\frac{1}{n(n-1)} \sum_{i \neq j} d_j(1 - d_i) \omega_{ij} \left[ \left( \Delta \hat{q}_r^{(1)}(w_i) - \Delta q_r^{(1)}(w_i) \right) + \left( r \left( \Delta \hat{q}_r^{(0)}(w_i) - \Delta q_r^{(0)}(w_i) \right) \right) \right].
\]

We establish a linear representation for the term involving \( (\Delta \hat{q}_r^{(1)}(w_i) - \Delta q_r^{(1)}(w_i)) \). It follows from the arguments used in Lemma A.4 in CK that
\[
\frac{1}{n(n-1)} \sum_{i \neq j} d_j(1 - d_i) \omega_{ij} \left[ \Delta \hat{q}_r^{(1)}(w_i) - \Delta q_r^{(1)}(w_i) \right]
\]
can be expressed as
\[
\frac{1}{n} \sum_{i=1}^{n} (1 - p(v_i))^2 f_V(v_i) d_i f_W(w_i) \left( f_{U_1 | W}(0 | w_i)^{-1}(I[y_i \leq q_r^{(1)}(w_i)] - \tau_2) - f_{U_1 | W}(0 | w_i)^{-1}(I[y_i \leq q_r^{(1)}(w_i)] - \tau_1) \right) + o_p(n^{-1/2}).
\]
An analogous linear representation can be derived for the term involving $(\Delta^q(w_i) - \Delta^{q(0)}(w_i))$, where we would replace $d_i$ with $(1 - d_i)$ in the above expression, and superscripts $(1)$ with superscripts $(0)$. Collecting both these terms we get this can be written as

$$\frac{1}{n} \sum_{i=1}^{n} \psi_i + o_p(n^{-1/2}).$$

We next consider the linear term of $\hat{\omega}_{ij}$ around $\omega_{ij}$. This is of the form

$$\frac{1}{n(n-1)} \sum_{i \neq j} d_j (1 - d_i) \omega'_{ij}(w_i + w_j)'(\hat{\delta} - \delta_0) \left[ \Delta^{q(1)}_r(w_i) - \Delta^{q(0)}_r(w_i)(r) \right].$$

where $\omega'_{ij} = \frac{\sum_{i \neq j} d_j (1 - d_i) \omega_{ij}(w_i + w_j)'}{h_{2n}}$.

Note we can replace $(\Delta^{q(1)}_r(w_i) - \Delta^{q(0)}_r(w_i)(r))$ with $(\Delta^{q(1)}_r(w_i) - \Delta^{q(0)}_r(w_i)(r))$ in the above expression. The resulting remainder term is $o_p(n^{-1/2})$ by the root-$n$ consistency of $\hat{\delta}$ and the uniform consistency of the quantile estimators. In what follows, we derive an expression for the probability limit of

$$\frac{1}{n(n-1)} \sum_{i \neq j} d_j (1 - d_i) \omega'_{ij} \left[ \Delta^{q(1)}_r(w_i) - r \Delta^{q(0)}_r(w_i) \right] (w_i + w_j)'.$$

Using standard U-statistic projection theorems and the change of variables, the above term converges in probability to $M$, defined in (3.17). Thus, the linear term in the expansion has the linear representation:

$$\frac{1}{n} \sum_{i=1}^{n} M \psi_i + o_p(n^{-1/2}).$$

Finally we note higher order terms in the expansion of $\hat{\omega}_{ij}$ around $\omega_{ij}$ are asymptotically negligible by the uniform rates of convergence of the quantile estimators and the root-$n$ consistency of $\hat{\delta}$. This completes the linear representation $\hat{r} - r$.

References


Table 1: Monte Carlo Results (I): Kernel-Weighted IV Estimators of the Average Scale Ratio.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Distribution and Sample Size</th>
<th>Mean Bias</th>
<th>Med. Bias</th>
<th>RMSE</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary IV:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bivariate Normal</td>
<td>100</td>
<td>0.1863</td>
<td>0.1196</td>
<td>0.4606</td>
<td>0.2212</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.0782</td>
<td>0.0477</td>
<td>0.2568</td>
<td>0.1465</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.0607</td>
<td>0.0536</td>
<td>0.1933</td>
<td>0.1220</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>0.0455</td>
<td>0.0331</td>
<td>0.1394</td>
<td>0.0824</td>
</tr>
<tr>
<td>Bivariate Student-t(10)</td>
<td>100</td>
<td>0.1331</td>
<td>0.0466</td>
<td>0.4781</td>
<td>0.2410</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.0466</td>
<td>0.0119</td>
<td>0.2913</td>
<td>0.1662</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>-0.0152</td>
<td>-0.0352</td>
<td>0.1914</td>
<td>0.1284</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>-0.0292</td>
<td>-0.0352</td>
<td>0.1469</td>
<td>0.0951</td>
</tr>
<tr>
<td>Bivariate Student-t(5)</td>
<td>100</td>
<td>0.0782</td>
<td>-0.0087</td>
<td>0.4654</td>
<td>0.2598</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.0213</td>
<td>-0.0313</td>
<td>0.3407</td>
<td>0.1885</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>-0.0408</td>
<td>-0.0639</td>
<td>0.2379</td>
<td>0.1583</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>-0.0611</td>
<td>-0.0726</td>
<td>0.1885</td>
<td>0.1255</td>
</tr>
<tr>
<td>Bivariate Student-t(3)</td>
<td>100</td>
<td>0.1059</td>
<td>0.0003</td>
<td>0.5979</td>
<td>0.3173</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-0.0049</td>
<td>-0.0819</td>
<td>0.4761</td>
<td>0.2654</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>-0.0624</td>
<td>-0.1251</td>
<td>0.3940</td>
<td>0.2394</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>-0.0970</td>
<td>-0.1410</td>
<td>0.3397</td>
<td>0.2076</td>
</tr>
</tbody>
</table>

Note: The number of replications is 801. RMSE stands for the root mean-square error, and MAD stands for the mean absolute deviation.
<table>
<thead>
<tr>
<th>Distribution and Sample Size</th>
<th>Distribution</th>
<th>Mean Bias</th>
<th>Med. Bias</th>
<th>RMSE</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bivariate Normality</td>
<td>Continuous IV</td>
<td>100</td>
<td>0.2468</td>
<td>-0.0169</td>
<td>1.9441</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.0948</td>
<td>-0.0084</td>
<td>1.2451</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>-0.0441</td>
<td>-0.0990</td>
<td>0.8307</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>-0.0020</td>
<td>-0.0248</td>
<td>0.5731</td>
</tr>
<tr>
<td>Bivariate Student-t(5)</td>
<td>Continuous IV</td>
<td>100</td>
<td>0.3053</td>
<td>0.0178</td>
<td>2.3147</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.1161</td>
<td>0.0176</td>
<td>1.3548</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>0.0038</td>
<td>-0.0234</td>
<td>0.8635</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>-0.0380</td>
<td>-0.0444</td>
<td>0.6149</td>
</tr>
<tr>
<td>Bivariate Cauchy</td>
<td>Continuous IV</td>
<td>100</td>
<td>11.7745</td>
<td>0.1225</td>
<td>110.0670</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>12.5100</td>
<td>0.9406</td>
<td>141.5662</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>8.2089</td>
<td>0.5693</td>
<td>136.8467</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>1.9160</td>
<td>0.3461</td>
<td>10.2839</td>
</tr>
<tr>
<td></td>
<td>Binary IV:</td>
<td>100</td>
<td>0.6575</td>
<td>0.3964</td>
<td>2.2403</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.3153</td>
<td>0.1625</td>
<td>1.3438</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>0.1889</td>
<td>0.1246</td>
<td>0.9159</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>0.1684</td>
<td>0.1290</td>
<td>0.6497</td>
</tr>
<tr>
<td>Bivariate Student-t(5)</td>
<td>Binary IV:</td>
<td>100</td>
<td>0.4464</td>
<td>0.1826</td>
<td>2.1636</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.2869</td>
<td>0.1864</td>
<td>1.4235</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>0.0667</td>
<td>0.0030</td>
<td>0.9772</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>0.0247</td>
<td>0.0030</td>
<td>0.6707</td>
</tr>
<tr>
<td>Bivariate Cauchy</td>
<td>Binary IV:</td>
<td>100</td>
<td>3.7260</td>
<td>-0.6895</td>
<td>32.5516</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>1.0291</td>
<td>-0.4128</td>
<td>8.4902</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>0.2468</td>
<td>-0.3072</td>
<td>5.5602</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>0.5676</td>
<td>-0.2702</td>
<td>11.2538</td>
</tr>
</tbody>
</table>

Note: The number of replications is 801. RMSE stands for the root mean-square error, and MAD stands for the mean absolute deviation. All the bandwidths equal $n^{-1/5}$. Use the scale ratio by 75-25 interquantile ranges.
Table 3: Mean and Standard Deviation

<table>
<thead>
<tr>
<th></th>
<th>All Schooling Levels</th>
<th></th>
<th>High School</th>
<th></th>
<th>Some College</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (1)</td>
<td>Std Dev (2)</td>
<td>Mean (3)</td>
<td>Std Dev (4)</td>
<td>Mean (5)</td>
<td>Std Dev (6)</td>
</tr>
<tr>
<td>Years of schooling</td>
<td>13.26</td>
<td>2.68</td>
<td>12.00</td>
<td>.00</td>
<td>15.37</td>
<td>1.64</td>
</tr>
<tr>
<td>(a) Outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log hourly earnings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1976</td>
<td>1.66</td>
<td>.44</td>
<td>1.64</td>
<td>.42</td>
<td>1.75</td>
<td>.43</td>
</tr>
<tr>
<td>(b) Instruments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-yr college in county</td>
<td>.68</td>
<td>.47</td>
<td>.66</td>
<td>.47</td>
<td>.73</td>
<td>.44</td>
</tr>
<tr>
<td>(c) Covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion black</td>
<td>.23</td>
<td>.42</td>
<td>.25</td>
<td>.43</td>
<td>.15</td>
<td>.35</td>
</tr>
<tr>
<td>Residence south 1976</td>
<td>.40</td>
<td>.49</td>
<td>.37</td>
<td>.48</td>
<td>.35</td>
<td>.48</td>
</tr>
<tr>
<td>Live in metropolitan 1976</td>
<td>.71</td>
<td>.45</td>
<td>.69</td>
<td>.46</td>
<td>.78</td>
<td>.41</td>
</tr>
<tr>
<td>Live in metropolitan 1966</td>
<td>.65</td>
<td>.48</td>
<td>.64</td>
<td>.48</td>
<td>.70</td>
<td>.46</td>
</tr>
<tr>
<td>Both parents attended college</td>
<td>.06</td>
<td>.24</td>
<td>.01</td>
<td>.10</td>
<td>.11</td>
<td>.32</td>
</tr>
<tr>
<td>(d) Size</td>
<td>3010</td>
<td>3010</td>
<td>992</td>
<td>992</td>
<td>1521</td>
<td>1521</td>
</tr>
</tbody>
</table>

Note: The first two columns are based on the NLSYM66 sample used in Card’s (1995) and Kling’s (2001) studies.

Table 4: Replicated IV Estimates of Return to Schooling

<table>
<thead>
<tr>
<th>Estimation method</th>
<th>OLS</th>
<th>IV</th>
<th>IV</th>
<th>IV</th>
<th>OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Coefficient on years of schooling</td>
<td>.073</td>
<td>.132</td>
<td>.133</td>
<td>.122</td>
<td>.075</td>
</tr>
<tr>
<td>Number of observations</td>
<td>3010</td>
<td>3010</td>
<td>3010</td>
<td>3010</td>
<td>3010</td>
</tr>
<tr>
<td>Quadratic experience</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Full set of demographics</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Card (1995): Table-column</td>
<td>T2-c2</td>
<td>T3-c5a</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kling (2001): Table-column</td>
<td>-</td>
<td>-</td>
<td>T1-c4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(b) Coefficient on college attendance

<table>
<thead>
<tr>
<th></th>
<th>.173</th>
<th>.473</th>
<th>.467</th>
<th>.447</th>
<th>.187</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>2513</td>
<td>2513</td>
<td>2513</td>
<td>2513</td>
<td>2513</td>
</tr>
<tr>
<td>Quadratic experience</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Full set of demographics</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

Note: Standard errors in parentheses. In columns (2)-(5), schooling and experience are treated as endogenous variables, with college proximity and age as excluded instruments. All the columns include indicators of black, south in 1976, residence in Standard Metropolitan Statistical Area (SMSA) in 1976, SMSA in 1966, and college attendance of both parents. In addition, Card (1995) and Kling control for the following variables that are included in columns (2)-(5): eight regional dummies, indicators for living with both parents and for living only with mother, and eleven interaction terms of parental education variables. We apply columns (4)-(5) to estimate the degree of dispersion in wages. See Tables 5 and 6 for parametric and semiparametric results.
Table 5: The Educational Effects on Wage Inequality and Average Scale Ratios, Using Conventional Estimation Methods

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>OLS</th>
<th>Heckman two stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Variance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>.157</td>
<td>.207</td>
</tr>
<tr>
<td>High School</td>
<td>.137</td>
<td>.198</td>
</tr>
<tr>
<td>College/HS Ratio</td>
<td>1.146</td>
<td>1.045</td>
</tr>
<tr>
<td></td>
<td>[.010]</td>
<td>(.948,1.166)</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>.396</td>
<td>.455</td>
</tr>
<tr>
<td>High School</td>
<td>.370</td>
<td>.445</td>
</tr>
<tr>
<td>College/HS Ratio</td>
<td>1.070</td>
<td>1.022</td>
</tr>
<tr>
<td></td>
<td>[.010]</td>
<td>(.974,1.080)</td>
</tr>
</tbody>
</table>

Note: NLSYM66 in 1976 using the models presented in columns (4)-(5) in Table 4. Footnote 10 discusses how these results are derived. P-values of F-tests are in [ ]; confidence intervals at the five percent significance level derived from bootstrap are in ( , ).

Table 6: Semiparametric Estimates of the Population Average Treatment Effects on Wage Inequality.

<table>
<thead>
<tr>
<th></th>
<th>Bandwidths</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n^{-1/6}$</td>
</tr>
<tr>
<td>(a) Variance-Based Estimator</td>
<td></td>
</tr>
<tr>
<td>Size of Symmetric Subsample</td>
<td>1144</td>
</tr>
<tr>
<td>Variance:</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>.1938</td>
</tr>
<tr>
<td>High School</td>
<td>.1806</td>
</tr>
<tr>
<td>College/High School Ratio</td>
<td>1.0729</td>
</tr>
<tr>
<td></td>
<td>(.0914,1.6953)</td>
</tr>
<tr>
<td>Standard Deviation:</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>.4402</td>
</tr>
<tr>
<td>High School</td>
<td>.4250</td>
</tr>
<tr>
<td>College/High School Ratio</td>
<td>1.0358</td>
</tr>
<tr>
<td></td>
<td>(.3024,1.3021)</td>
</tr>
<tr>
<td>(b) Quantile-Based Estimator</td>
<td></td>
</tr>
<tr>
<td>Number of Symmetric Pairs</td>
<td>646974</td>
</tr>
<tr>
<td>College/High School Ratio:</td>
<td></td>
</tr>
<tr>
<td>60-40 Interquantile Spread</td>
<td>1.0931</td>
</tr>
<tr>
<td></td>
<td>(.8639,1.1372)</td>
</tr>
<tr>
<td>75-25 Interquantile Spread</td>
<td>.9871</td>
</tr>
<tr>
<td></td>
<td>(.8902,1.0478)</td>
</tr>
<tr>
<td>90-10 Interquantile Spread</td>
<td>.9588</td>
</tr>
<tr>
<td></td>
<td>(.8772,.9806)</td>
</tr>
</tbody>
</table>

Note: Same as Table 5