Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market

James J. Heckman

University of Chicago

Guilherme Sedlacek

Carnegie-Mellon University

This paper presents an empirical equilibrium model of self-selection in the labor market that recognizes the existence of measured and unmeasured heterogeneous skills. We derive a model of the sectoral allocation of workers of different demographic types and present a new econometric procedure for combining micro and macro data to estimate supply and demand functions for unmeasured sector-specific productive attributes. Our model extends previous empirical work on wage equations by introducing determinants of aggregate market demand and supply into an explicit, economically interpretable estimating equation. These extensions are required to produce a model that fits the distribution of wages for the U.S. labor market.

This research was supported by NSF grants DAR 79-25924, SES-8107963, and SES 84-11242 to the Quantitative Methods Center at National Opinion Research Center. The first draft (entitled "The Impact of the Minimum Wage on the Employment and Earnings of Workers in South Carolina") was prepared in December 1980. That draft and subsequent drafts have been circulated since 1981 in Heckman's course, Labor Economics 442. A fourth draft (entitled "An Equilibrium Model of the Industrial Distribution of Workers and Wages") was presented to the summer meetings of the Econometric Society in Stanford, California, June 1984. Bo Honore, Ricardo Barros, Joe Hotz, Robert Michael, Sherwin Rosen, and José Scheinkman made valuable comments on this paper as did participants in seminars at Chicago, Columbia, Concordia (Montreal), Kentucky, and Penn. We are especially grateful to Ricardo Barros for valuable comments. We also would like to thank two anonymous referees. We thank Vicky Longawa for valuable editorial assistance.

[*Journal of Political Economy*, 1985, vol. 93, no. 6]
 0022-3808/85/9306-0012$01.50

Diversity in the amount and type of skills possessed by workers is a central feature of modern labor markets. Yet econometric analysis of aggregate labor market data either ignores such diversity entirely (e.g., Sargent 1978; Geary and Kennan 1982) or assumes homogeneous skills for workers classified by such criteria as age, race, education, and sex (e.g., Hamermesh and Grant 1979; Gollop and Jorgenson 1983; Jorgenson 1985). While the second approach to labor aggregation improves on the first by recognizing worker diversity, it still ignores plausible heterogeneity in skills within the available crude demographic categories. Moreover, it is not obvious that demographic categories define economically meaningful skill categories.

Welch (1969) recognizes the diversity of skills within crude demographic-education groups and uses the Lancaster (1966) and Gorman (1980) characteristics model to postulate that labor incomes are the sum of the incomes earned on distinct measured and unmeasured attributes owned by each worker with a *uniform* price per attribute across all market sectors. In his model, workers are indifferent among sectors of the economy (i.e., there is no scope for comparative advantage) because identical firms are able to repackage worker skill bundles costlessly.

Heckman and Scheinkman (1982), building on suggestions by Mandelbrot (1962), derive conditions under which prices for measured and unmeasured attributes are uniform across all market sectors. They present empirical evidence that rejects this description of the labor market and hence the Welch approach for U.S. data. Their evidence suggests that the pursuit of comparative advantage is an important feature of U.S. labor market data (see also Sattinger 1980).

This paper presents an empirical equilibrium model of comparative advantage or self-selection in the labor market that recognizes the existence of measured and unmeasured heterogeneous skills within even narrowly defined demographic groups. The points of departure for our work are the seminal Roy (1951) model of income distribution and later applications of the Roy model by Rosen (1978) and Willis and Rosen (1979). We derive a model of the sectoral allocation of workers of different demographic types. We also present a new econometric procedure for combining micro and macro data to estimate supply and demand functions for unmeasured sector-specific productive attributes.

Our methodology extends previous statistical work on self-selection to an explicit market setting in which the prices of attributes respond to changes in the determinants of aggregate demand and supply. Our model extends previous empirical work on wage equations by introducing determinants of aggregate market demand and supply into explicit, economically interpretable estimating equations. We extend Roy's model of self-selection by embedding it in a market setting and

by (*a*) introducing a nonmarket sector, (*b*) allowing workers to select their sector of employment on the basis of utility maximization rather than income maximization, and (*c*) permitting unmeasured attributes to be nonlognormally distributed. These extensions are required to produce a model that fits the distribution of wages for the U.S. labor market.

This study presents evidence that supports the commonly utilized practice of aggregating manufacturing into a single sector for the purpose of estimating labor demand functions. However, a new aggregate is required that recognizes both measured and unmeasured heterogeneity in skills in the population and accounts for self-selection decisions by agents.

We use our model to estimate the importance of aggregation bias in measured aggregate real wage rates. Aggregation bias reduces measured wage variability in manufacturing below what it would be if the quality of the manufacturing work force were held constant. However, for the economy as a whole, precisely the opposite effect occurs. Aggregation bias causes measured aggregate wage variability to overstate quality constant wage variability. Because of comparative advantage, workers who move from one sector to another in response to a macro disturbance lower the average quality of the work force in the sector to which they go and raise the average quality in the sector from which they depart. This phenomenon accentuates measured wage variability over what it would be if sectoral labor force quality were held constant.

This paper is in four sections. Section I presents a rigorous statement of Roy's model of self-selection and embeds it in a market setting. We present a new method for combining micro and macro data to estimate the demand and supply of unmeasured sector-specific productive attributes. Section II extends Roy's model. Our extended model nests Roy's as a special case and so is convenient for econometric testing. Unlike the Roy model, the proposed model can generate Pareto-like right tails that are claimed to be an essential feature of income and wage distributions by Mandelbrot (1962), Lydall (1968), and others. Section III reports empirical estimates and tests of the new model. We estimate the contribution of self-selection to income inequality and present empirical evidence on the importance of aggregation bias in measured aggregate real wage movements. The paper concludes with a brief summary (Sec. IV).

I. An Estimable General Equilibrium Roy Model

A. The Model

We begin the analysis by expositing the point of departure for our own work: the Roy model of self-selection for workers with heteroge-

neous skills. Following Roy, we assume that there are two market sectors in which income-maximizing agents can work. Agents are free to enter the sector that gives them the highest income. However, they can work in only one sector at a time.

Each sector requires a unique sector-specific task. Each agent is endowed with a J-dimensional skill vector $\mathbf{s}$ that enables him to perform sector-specific tasks. Vector $\mathbf{s}$ is continuously distributed with density $g(\mathbf{s}|\Theta)$, where Θ is a vector of parameters. The model is short run in that aggregate skill distributions are assumed to be given. There are no costs of changing sectors, and investment is ignored. Because of this assumption, the model presented here applies to environments with certain or uncertain prices for sector-specific tasks. For simplicity and without any loss of generality, we assume an environment of perfect certainty. We leave the development of a more dynamic model with investment and mobility costs for another occasion.

Let $t_i(\mathbf{s})$ be a nonnegative function that expresses the amount of sector i specific task a worker with skill endowment $\mathbf{s}$ can perform. This function is technologically determined. However, it may shift over time as technology changes. The task functions are assumed to be continuously differentiable in $\mathbf{s}$. The distinction between tasks as objects of firm demand and skills as endowments of workers captures the idea that packages of skills cannot be unbundled and that different skills are used in different tasks.[1]

The output of sector i, denoted Y_i, is assumed to depend on the sum of individual sector-specific tasks employed in the sector and not on its distribution. Denoting $\mathbf{A}_i$ as a vector of nonlabor inputs, the aggregate production function for sector i is assumed to be of the form

$$Y_i = F^{(i)}(T_i, \mathbf{A}_i), \quad i = 1, 2,$$

where T_i is the total amount of task i employed in sector i. Function $F^{(i)}$ is assumed to be twice continuously differentiable and strictly concave in all of its arguments, with positive inputs required for positive outputs.

For fixed output price P_i, the equilibrium price of task i in sector i, denoted π_i, is the value of the marginal product of a unit of the task

$$\pi_i = P_i \frac{\partial F^{(i)}}{\partial T_i}, \quad i = 1, 2. \tag{1}$$

An agent with endowment $\mathbf{s}$ works in sector i if his income is higher there, that is,

[1] This specification is sufficiently general that it permits the same skills to be equally productive in generating all tasks. Thus some of the skills may have the economic character of general human capital.

$$\pi_i t_i(\mathbf{s}) \geq \pi_j t_j(\mathbf{s}), \quad i \neq j,\ i, j = 1, 2. \tag{2}$$

Indifference between sectors is a negligible probability event since the t_i, $i = 1, 2$, are assumed to be continuous nondegenerate random variables.[2] Throughout we assume that prices are positive ($\pi_i > 0$).

Inequality (2) defines a set of $\mathbf{s}$ values, not necessarily connected, in which agents with values of $\mathbf{s}$ in the set earn their highest income by working in sector i. This set is defined as

$$\mathcal{S}_i = \{\mathbf{s}\colon \pi_i t_i(\mathbf{s}) \geq \pi_j t_j(\mathbf{s}),\ i \neq j\}.$$

If the t_i are linear functions of $\mathbf{s}$, (2) partitions the domain of $\mathbf{s}$ into two connected sets. For this specification of the $t_i(\mathbf{s})$ functions (and others as well) there is a market stratification of workers into tasks by their $\mathbf{s}$ type. Demographic groups differing in their distribution of skill endowment tend to specialize in different sectors. There may be "black" or "teenage" jobs, not because those demographic categories are of direct interest to employers, but because members of those groups possess skill endowments of special use in a particular sector.

The log wage in sector i of an individual with endowment $\mathbf{s}$ is

$$\ln w_i(\mathbf{s}) = \ln \pi_i + \ln t_i(\mathbf{s}). \tag{3}$$

Assuming that the function mapping skills to tasks does not change over time, (3) implies that log wage functions (expressed as functions of $\mathbf{s}$) have identical coefficients in successive cross sections except for their intercepts. This implication is termed the "proportionality hypothesis" in this paper.[3] Specification (3) rationalizes the "paradoxical" result that the rate of return to schooling (the coefficient of schooling in a log wage equation that is linear in schooling) has not changed over time despite expansion in the aggregate stock of schooling. In wage function (3) an exogenous increase in the supply of schooling affects only the intercept of the log wage equation.

Wage equation (3) is not a conventional hedonic function. In the hedonic models of Tinbergen (1951, 1956) and Rosen (1974), an implicit market prices out each component of $\mathbf{s}$. In the model of (3) the t_i are priced out, not (directly) the components of $\mathbf{s}$. Hedonic wage equations fit in separate market sectors could be interpreted as revealing "prices" for the attributes in each sector, but there would be no economic content in such an interpretation. In a single cross section of data, wage equation (3) is empirically indistinguishable from a conventional hedonic wage equation.

[2] More precisely, (t_1, t_2) is a nondegenerate, continuously distributed random vector.

[3] A better name would be "additivity hypothesis," but that term has special meaning in the theory of consumer demand. Clearly wage functions in levels are proportionately related across time if the hypothesis is correct.

The proportion of the population working in sector i is the proportion of the population whose skill endowments lie in $\mathcal{S}_i$:

$$\text{pr}(i) = \int_{\mathcal{S}_i} g(\mathbf{s}|\Theta)d\mathbf{s}.$$

The aggregate supply of task to sector i, $\tilde{T}_i$, is obtained by integrating the micro supply over $\mathcal{S}_i$:

$$\tilde{T}_i = \int_{\mathcal{S}_i} t_i(\mathbf{s})g(\mathbf{s}|\Theta)d\mathbf{s}, \quad i = 1, 2.$$

Both $\tilde{T}_i$ and pr(i) are homogeneous of degree zero functions of $\boldsymbol{\pi}$ and are monotone increasing (strictly nondecreasing) functions of π_i and monotone decreasing (strictly nonincreasing) functions of π_j ($i \neq j$).

An equilibrium exists for a given vector $\mathbf{A}_i$ of nonlabor inputs if, when the $\tilde{T}_i$ are inserted in (1), there are prices $\hat{\boldsymbol{\pi}}$ that exactly evoke supply $\tilde{T}_i$, $i = 1, 2$. Under standard conditions on the technology it is possible to establish that an equilibrium exists in the labor market.

Without further restrictions, the Roy model produces no interesting refutable empirical hypotheses.[4] To produce such hypotheses it is necessary to postulate specific functional forms.

Roy assumes that the density of skills $g(\mathbf{s}|\Theta)$ and the task functions $t_i(\mathbf{s})$ are such that (ln t_1, ln t_2) is normally distributed with mean (μ_1, μ_2) and covariance matrix Σ. Letting (u_1, u_2) be a mean zero normal vector, agents in the Roy model choose between two possible wages:

$$\ln w_1 = \ln \pi_1 + \mu_1 + u_1$$

or

$$\ln w_2 = \ln \pi_2 + \mu_2 + u_2.$$

Workers enter sector 1 if ln $w_1 >$ ln w_2. Otherwise they enter sector 2.

Letting $\sigma^* = \sqrt{\text{var}(u_1 - u_2)}$ and $c_i = [\ln(\pi_i/\pi_j) + \mu_i - \mu_j]/\sigma^*$, $i \neq j$,

$$\text{pr}(i) = P(\ln w_i > \ln w_j) = \Phi(c_i), \quad i \neq j, \; i, j = 1, 2,$$

where $\Phi(\ )$ is the cumulative distribution function of a standard normal variable. When standard sample selection bias formulae are used (see, e.g., Heckman 1976, 1979), the mean of log wages *observed* in sector i is

$$E(\ln w_i|\ln w_i > \ln w_j) = \ln \pi_i + \mu_i + \left(\frac{\sigma_{ii} - \sigma_{ij}}{\sigma^*}\right)\lambda(c_i), \quad i, j = 1, 2, \; i \neq j, \tag{4}$$

[4] The proportionality hypothesis is an implication of the assumption of the existence of sector-specific efficiency units that underlies wage specification (3) and not specifically the Roy model. For further discussion of the empirical content of the Roy model, see Heckman and Singer (1985).

where

$$\lambda(c) = \frac{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}c^2)}{\Phi(c)}$$

is a convex monotone decreasing function of c with $\lambda(c) \geq 0$,

$$\lim_{c \to \infty} \lambda(c) = 0, \lim_{c \to -\infty} \lambda(c) = \infty.$$

Convexity is proved in Heckman and Sedlacek (1986).

The variance of log wages observed in sector i is

$$\begin{aligned} \operatorname{var}(\ln w_i | \ln w_i > \ln w_j) = \sigma_{ii}\{\rho_i^2[1 - c_i\lambda(c_i) - \lambda^2(c_i)] \\ + (1 - \rho_i^2)\}, \end{aligned} \tag{5}$$

where $\rho_i = \operatorname{correl}(u_i, u_i - u_j)$, $i \neq j$, $i, j = 1, 2$. The variance of the log of observed wages never exceeds σ_{ii}, the population variance, because the term in braces in (5) is never greater than unity. In general, sectoral variances decrease with increased selection. For example, if ρ_1 and ρ_2 do not equal zero, as π_1 increases with π_2 held fixed so that people shift from sector 2 to sector 1, the variance in the log of wages in sector 1 increases while the variance in the log of wages in sector 2 decreases.

Using the fact that $w_i = \pi_i t_i$,

$$E(\ln t_1 | \ln w_1 > \ln w_2) = \mu_1 + \frac{\sigma_{11} - \sigma_{12}}{\sigma^*} \lambda(c_1), \tag{4a)'}$$

$$E(\ln t_2 | \ln w_2 > \ln w_1) = \mu_2 + \frac{\sigma_{22} - \sigma_{12}}{\sigma^*} \lambda(c_2). \tag{4b)'}$$

Focusing on (4a)′ and noting that λ is positive for all values of c_1 (except $c_1 = \infty$), we see that the mean of log task 1 used in sector 1 exceeds, equals, or falls short of the population mean endowment of log task 1 as $\sigma_{11} - \sigma_{12}$ is greater than, equal to, or less than zero. If endowments of tasks are uncorrelated ($\sigma_{12} = 0$), self-selection always causes the mean of $\ln t_1$ employed in sector 1 to be above the population mean μ_1. The opposite case occurs when $\sigma_{11} - \sigma_{12}$ is negative. This case can arise only when values of $\ln t_1$ and $\ln t_2$ are sufficiently positively correlated. If this occurs, the mean of log task 1 used in sector 1 falls below the population mean μ_1. Since covariance matrices must be positive semidefinite, $\sigma_{11} + \sigma_{22} - 2\sigma_{12} \geq 0$. Thus if $\sigma_{11} - \sigma_{12} < 0$, $\sigma_{22} - \sigma_{12} > 0$ so the mean of log task 2 employed in sector 2 necessarily lies above the population mean μ_2. In the Roy model the unusual case can arise in at most one sector. Notice from (5) that only if $\sigma_{11} - \sigma_{12} = 0$ (so $\rho_1^2 = 0$) is the variance of log task 1 employed in sector 1 identical to the variance of log task 1 in the population.

Otherwise, the sectoral variance of observed log task 1 is less than the population variance of log task 1.

To gain further insight into the effect of self-selection on the distribution of earnings for workers in sector 1, it is helpful to draw on some results from normal regression theory. The regression equation for $\ln t_2$ conditional on $\ln t_1$ is

$$\ln t_2 = \mu_2 + \frac{\sigma_{12}}{\sigma_{11}} (\ln t_1 - \mu_1) + \epsilon_2, \tag{6}$$

where $E(\epsilon_2) = 0$ and $\text{var}(\epsilon_2) = \sigma_{22}[1 - (\sigma_{12}^2/\sigma_{11}\sigma_{22})]$.

Figure 1 plots regression function (6) for the case $\sigma_{12} = \sigma_{11}$ and $\mu_2 > \mu_1 > 0$. For each value of $\ln t_1$, the population values of $\ln t_2$ are normally distributed around the regression line. Individuals with high values of $\ln t_1$ also tend to have a high value of $\ln t_2$. Assuming $\pi_1 = \pi_2$, individuals with ($\ln t_1$, $\ln t_2$) endowments above the 45-degree line of equal income shown in figure 1 choose to work in sector 2, while those individuals with endowments below this line work in sector 1. Because $\sigma_{12} = \sigma_{11}$, the regression function is parallel to the line of equal income.

The distribution of ϵ_2 about the regression line is the same for all values of $\ln t_1$. When individuals are classified on the basis of their $\ln t_1$ values, the same proportion of individuals work in sector 1 at all values of $\ln t_1$. For this reason the distribution of $\ln t_1$ employed in sector 1 is the same as the latent population distribution. If π_1 is raised (or π_2 is lowered) so that the 45-degree equal income line is shifted upward, the same proportion of people enter sector 1 at each value of t_1.

Figure 2 plots regression function (6) for the case $\sigma_{12} > \sigma_{11}$ and $\mu_2 > \mu_1 > 0$. As before we set $\pi_1 = \pi_2$. Individuals with endowments above the 45-degree line choose to work in sector 2, while those with endowments below this line work in sector 1. When individuals are classified on the basis of their t_1 values, the fraction of people working in sector 1 *decreases* the higher the value of t_1. Self-selection causes the mean of log task 1 employed in sector 1 to be less than the mean of log task 1 in the total population. People with high values of t_1 are underrepresented in sector 1 and low t_1 values are overrepresented. In the extreme, when $\ln t_1$ and $\ln t_2$ are perfectly correlated, all high-income individuals are in sector 2, while all the low-income individuals are in sector 1. The highest-paid sector 1 worker earns the same as the lowest-paid sector 2 worker.

If π_1 is raised (or π_2 is lowered) so that the line of equal income is shifted upward, the mean of $\ln t_1$ employed in sector 1 must rise. The only place left to get t_1 is from the high end of the t_1 distribution. Unlike the case of $\sigma_{12} = \sigma_{11}$, in which a 10 percent increase in π_1

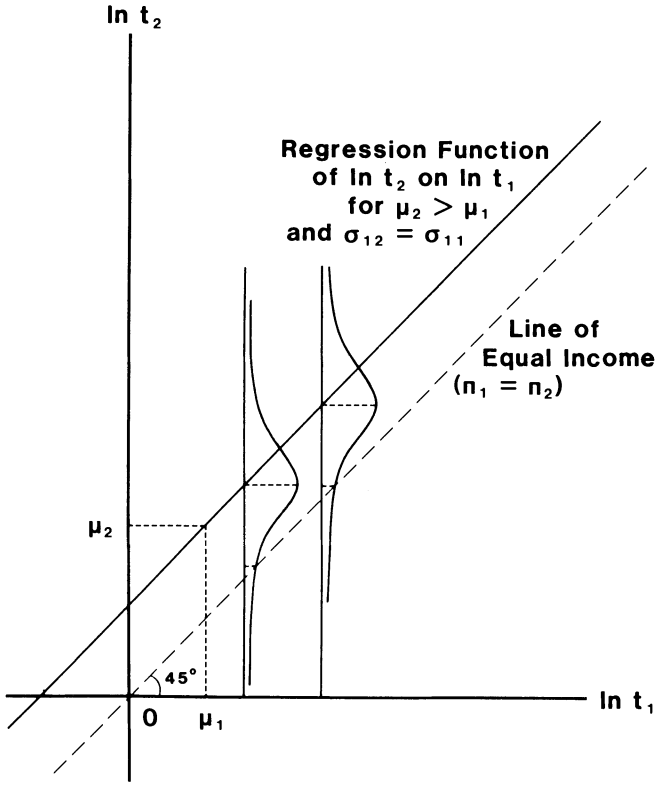

FIG. 1

results in a 10 percent increase in measured average earnings in sector 1, when $\sigma_{12} > \sigma_{11}$ a 10 percent increase in π_1 results in a greater than 10 percent increase in the measured average earnings in sector 1 as the average quality of the sector 1 work force increases. The variance of log wages in sector 1 *increases*.

If $\sigma_{11} < \sigma_{12}$, then $\sigma_{12} < \sigma_{22}$ in order for Σ to be a covariance matrix. In the population, log task 2 must have greater variability than log task 1. Individuals with high t_1 values tend to have high t_2 values. But the population distribution of log task 2 has more mass in the tails. The higher an agent's value of t_1, the more likely it is that he will be able to get higher income in sector 2. At the lower end of the distribution, the process works in reverse: lower t_1 individuals on average have poor t_2 values. Self-selection causes the ln t_1 distribution in sector 1 to have an evacuated right tail, an exaggerated left tail, and a lower mean than the population mean of ln t_1.

If $\sigma_{12} < \sigma_{11}$ (a case not depicted graphically), the proportion of each t_1 group working in sector 1 *increases*, the higher the value of t_1. The

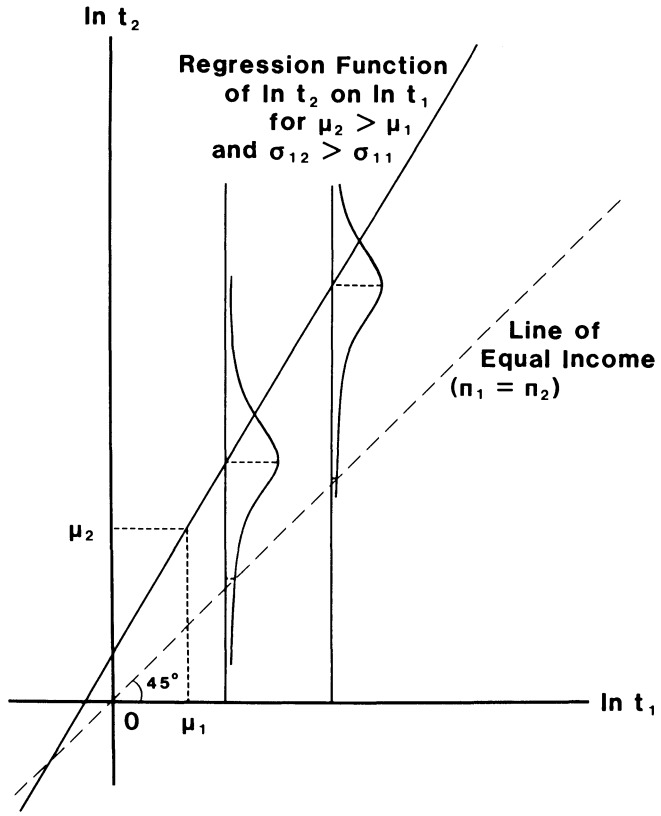

FIG. 2

mean of the log task employed in sector 1 exceeds μ_1. A 10 percent increase in π_1 produces an increase of less than 10 percent in the average earnings of workers in sector 1 as the mean of ln t_1 employed in sector 1 declines. In fact if $\sigma_{12} > \sigma_{22}$ it is possible for an increase in π_1 to cause measured sector 1 wages to decline.

B. Estimating the Model

We next propose a method for consistently estimating (*a*) the parameters of the distribution of tasks including the parameters of the functions relating skills to tasks and (*b*) the parameters of the sectoral demand functions for unmeasured tasks.

We assume access to the following commonly available data: (i) time-series data on the aggregate amount of compensation paid to workers in each sector; (ii) microeconomic repeated cross-section data on the wages of workers by sector and their associated demographic

and productivity characteristics; and (iii) time-series data on sectoral determinants of the demand for tasks. The most challenging aspect of our problem is that quantities of sector-specific tasks and their associated prices are not directly measured.

Assume that the functional form relating skills to tasks is $\ln t_i = \mathbf{c}_i\mathbf{s}$, $i = 1, 2$. Vector $\mathbf{s}$ is decomposed into measured and unmeasured components, $\mathbf{s}_o$ and $\mathbf{s}_u$. Their associated coefficients in $\mathbf{c}_i$ are $\mathbf{c}_{io}$ and $\mathbf{c}_{iu}$. Assuming that (*a*) $\mathbf{s}_u$ is distributed independently of $\mathbf{s}_o$ and (*b*) $E(\mathbf{c}_{iu}\mathbf{s}_u) = 0$ defines an operational task function. Let $\boldsymbol{\beta}_i = \mathbf{c}_{io}$, $\mathbf{c}_{iu}\mathbf{s}_u = u_i$, and $\mathbf{s}_o = \mathbf{x}$ so that the task function may be written as

$$\ln t_i = \boldsymbol{\beta}_i\mathbf{x} + u_i, \quad i = 1, 2 \tag{7}$$

and log real wages are

$$\ln w_i = \ln \pi_i + \ln t_i = \ln \pi_i + \boldsymbol{\beta}_i\mathbf{x} + u_i, \quad \mathrm{i} = 1, 2. \tag{8}$$

Unless $\sigma_{ii} - \sigma_{ij} = 0$, least-squares estimators of the parameters of equation (8) fit on sector i wage data are inconsistent because of selection bias. Empirical evidence of self-selection supports the model. It is necessary to control for selection bias in order to perform a proper test of the proportionality hypothesis. This hypothesis states that the slope coefficients of selectivity-corrected real wage equation (8) should be the same in all cross sections, but the intercept may vary if task prices change.

The intercept of real wage equation (8) combines two parameters: (*a*) the log of the real price of task i, $\ln \pi_i$, and (*b*) the intercept of the task function, denoted β_{0i}. Assuming a time-invariant distribution of unobservable u_i, sample selection bias corrected regressions of log wages on $\mathbf{x}$ consistently estimate $\ln \pi_i$ up to a constant (β_{0i}) from the intercept of the wage equation. Conventional methods are available to estimate consistently the slope coefficients of $\boldsymbol{\beta}_i$ and $\Sigma = \mathrm{var}(u_1, u_2)$.[5]

Estimating sample selection bias corrected versions of (8) for each sector for each cross section generates a time series on $\ln \pi_i + \beta_{0i}$. To obtain the quantities of log task employed in each sector in each period, subtract the estimated intercept from the log real wage bill in sector i, WB_i (the total labor compensation paid out in the sector denominated in constant dollars). This produces an estimate of labor aggregate $\ln T_i$ up to an unknown additive constant (β_{0i}). This labor aggregate is *not* a Divisia labor index. That index is constructed assuming homogeneous skills for measured demographic categories. Our index of labor skills recognizes that skills may be diverse within even narrowly defined demographic groups, that demographic

[5] See Heckman (1976) and Heckman and Sedlacek (1981, 1986). It is possible to estimate Σ and (μ_1, μ_2) with no regressors in the model.

groups are not necessarily economically meaningful skill groups, and that self-selection determines the supply of skills in the market.[6]

Let l denote a year subscript. Assuming that the aggregate derived demand for tasks is loglinear in aggregate tasks and real task prices, we write

$$\ln T_{il} = \delta_{0i} + \delta_{1i} \ln\left(\frac{\pi_{il}}{P_{il}}\right) + \boldsymbol{\delta}_{2i} \ln\left(\frac{\mathbf{P}_{Al}}{P_{il}}\right) + e_{il}, \quad l = 1, \ldots, L, \tag{9}$$

where e_{il} is a realization of a mean zero stationary stochastic process that shocks production technology (1), $\mathbf{P}_{Al}$ is a vector of real prices for other inputs, and P_{il} is the real price of output of sector i at time l. Economic theory predicts that δ_{1i} should be negative.

Setting π_{il}, $i = 1, 2$, equal to one in a benchmark year defines the units of tasks T_{il}. Using the definition of the real wage bill $T_{il}\pi_{il} = WB_{il}$, we may write equation (9) as

$$\ln\left(\frac{WB_{il}}{P_{il}}\right) = [\delta_{0i} - \beta_{0i}(\delta_{1i} + 1)] + (\delta_{1i} + 1)(\widehat{\ln \pi_{il}} - \ln P_{il}) + \boldsymbol{\delta}_{2i} \ln\left(\frac{\mathbf{P}_{Al}}{P_{il}}\right) + \tilde{e}_{il}, \quad i = 1, 2, \; l = 1, \ldots, L, \tag{10}$$

where $\widehat{\ln \pi_{il}}$ is the intercept estimated from the microeconomic log wage equation fit in sector i in year l (eq. [8]) and $\tilde{e}_{il}$ differs from e_{il} by the estimation error of $\widehat{\ln \pi_{il}}$ for $\ln \pi_{il}$; $\tilde{e}_{il} \equiv e_{il} + (\delta_{1i} + 1)(\beta_{0i} + \ln \pi_{il} - \ln \hat{\pi}_{il})$. Because it is plausible that aggregate shocks (e_{il}) determine deflated product price (P_{il}) as well as π_{il}, least squares does not in general consistently estimate the parameters of (10).

Potential instrumental variables for $\ln \pi_{il}$ and P_{il} include the determinants of the aggregate skill distribution such as government policy variables affecting labor supply.[7] The fact that the $\widehat{\ln \pi_{il}}$ are estimated from cross-section data does not create any econometric problem provided that in each cross section the u_i are distributed independently of each other, the number of cross-section observations used to estimate $\ln \pi_{il}$ becomes large relative to the number of time-series observations, and the numbers of both types of observations are assumed to become large.[8] Using standard instrumental variables methods, it is possible to estimate consistently the parameters of demand equation (10).

The model can be extended to let the population mean of the task

[6] For a discussion of Divisia indices of labor aggregates see Gollop and Jorgenson (1983).

[7] Jorgenson, Lau, and Stoker (1982) use such instruments in estimating general equilibrium models.

[8] In our case each cross section has around 3,200 observations whereas the time series has only 14 observations.

function shift over time. Defining β_{0il} as the intercept of task i function in year l, we write $\beta_{0il} = m_i(l) + \eta_{il}$, where $m_i(l)$ is a function of observed characteristics (e.g., polynomials in time) and η_{il} is a mean zero stationary stochastic process assumed to be distributed independent of $m_i(l)$. Noting that $\ln\hat{\pi}_{il}$ includes β_{0il} and substituting for β_{0i} in (10), we reach

$$\ln\left(\frac{WB_{il}}{P_{il}}\right) = \delta_{0i} + (\delta_{1i} + 1)(\ln\hat{\pi}_{il} - \ln P_{il}) - (\delta_{1i} + 1)m_i(l) + \boldsymbol{\delta}_{2i} \ln\left(\frac{\mathbf{P}_{Ai}}{P_{il}}\right) + [\tilde{e}_{il} - (\delta_{1i} + 1)\eta_{il}], \quad (11)$$

$$i = 1, 2, l = 1, \ldots, L,$$

where $\tilde{e}_{il}$ is redefined using $m_i(l)$ in place of β_{0i} in the previous expression. Provided that $m_i(l)$ is a low-dimensional function of l, instrumental variable methods still consistently estimate the parameters of (11). However, if the technology mapping skills to tasks is subject to distinct year-specific shocks, so $m_i(l)$ is a polynomial of degree L, there are no degrees of freedom in the time series and none of the parameters of the demand functions can be identified.

C. *Concluding Remarks on the Roy Model*

If only because the manufacturing sector as a whole has been the focus of so many empirical studies of the demand for labor, a natural starting point for our empirical analysis divides the economy into manufacturing and nonmanufacturing sectors. By dividing the data in this fashion, we can test for the existence of our proposed labor aggregate in either sector.

For the model to be empirically acceptable, it is required that (*a*) demand functions be downward sloping ($\delta_{1i} < 0, i = 1, 2$) and that (*b*) the proportionality hypothesis of the temporal stability of the wage equation (except for intercepts) not be rejected. In addition, since a normality assumption for (u_1, u_2) is not innocuous and sample selection bias corrections based on misspecified distributions produce biased estimates,[9] we require that fitted wage distributions accord with actual wage distributions in the sense of producing an acceptable χ^2 goodness-of-fit statistic.[10]

[9] However, they can still be used to test consistently for sample selection (see Heckman 1980). Goldberger (1983) and Heckman and MaCurdy (1985) discuss nonnormal models.

[10] A fourth test of the model examines evidence of sample selection bias (a nonzero coefficient on $\lambda[c_i]$ for $i = 1, 2$) in the wage equation in at least one of the two sectors. This test does not generalize to the model presented in the next section and so is not discussed further.

When the Roy model is fit on Current Population Survey earnings data disaggregated into manufacturing and nonmanufacturing sectors, it is rejected by these test criteria. The proportionality hypothesis is rejected and a χ^2 goodness-of-fit test strongly rejects the underlying distributional assumptions. (These estimates and their failure to account for the observed income distribution are discussed in Heckman and Sedlacek [1986].)

There are a number of possible responses to this rejection of the model. One possible reason for rejection is the highly aggregative nature of the manufacturing and nonmanufacturing sectors. By disaggregating the data into smaller, more economically well-defined sectors, we may be able to produce a model that survives our test criteria. The practical difficulty that arises in pursuing this avenue of investigation is that general multistate discrete data models are computationally very expensive to fit.

An alternative response to our rejection of the Roy model that is pursued in the rest of this paper preserves the two-market-sector split and generalizes the basic Roy model.

II. An Extended Roy Model

A. *The Model*

We extend the Roy model by (*a*) assuming that workers maximize utility and not just money income in making their sectoral choice decisions,[11] (*b*) decomposing earnings into hourly wage rates and hours of work and assuming that the latter are freely chosen, (*c*) developing a general nonnormal model for unmeasured tasks (u_1, u_2) that nests Roy's model as a special case, and (*d*) incorporating a nonmarket or household production sector as an alternative to market activity. All four extensions are required to produce a two-market-sector model of hourly wage rates that fits data from the U.S. labor market and survives the test criteria presented in Section I. We focus on explaining wage rates in our empirical analysis leaving the empirical analysis of hours of work and earnings for another occasion.

In place of task function (7), which maps skills to tasks, we utilize a more general Box-Cox model

$$\frac{t_i^{\lambda_i} - 1}{\lambda_i} = \boldsymbol{\beta}_i \mathbf{x} + u_i, \quad i = 1, 2. \tag{12}$$

Random variable u_i is equated to an underlying mean zero normal

[11] Lee (1978) was the first to make this extension in a model without a nonmarket sector.

random variable u_i^* for values of that variable that produce positive values of t_i, that is, $u_i = u_i^*$ if

$$1 + \lambda_i(\boldsymbol{\beta}_i \mathbf{x} + u_i^*) \geq 0. \tag{13}$$

Imposing a nonnegativity restriction on the admissible t_i avoids both mathematical and economic absurdities.[12] The joint density for (t_1, t_2) is presented in Appendix A. A convenient feature of our statistical model is that when $\lambda_i = 0$ equation (12) specializes to Roy's model (7), which always satisfies inequality (13) so u_i is normally distributed. By estimating λ_i we can determine whether or not the lognormal Roy model fits the data.

In our more general model, self-selection (with either income-maximizing or utility-maximizing selection rules) does not necessarily decrease the variance of $\ln t_i$ over what it would be in nonselected populations as is the case in the Roy model. In Heckman and Sedlacek (1986) we demonstrate that the more negative are the values of λ_i and the more negatively correlated are the latent normal random variables (u_1^*, u_2^*), the more likely it is that selection increases the variance of $\ln t_i$ and $\ln w_i$ for workers employed in sector i. In our more general model, self-selection can increase inequality (measured by the variance of logs) both within and between sectors over what it would be in the absence of self-selection, whereas in the Roy model selection must decrease within-sector inequality.

Our model can produce a Pareto tail for wage rates or tasks whereas the tails in the Roy model are thinner than Pareto tails. A Paretian tailed density $g(t_1)$ has a tail such that

$$\lim_{t_1 \to \infty} g(t_1) \sim c t_1^{-\alpha}, \quad \alpha > 1, c > 0.$$

Using the expression for the density of $t_1, f(t_1)$, given in equation (A2) of Appendix A and assuming $\lambda_1 = 0$, we get

$$\lim_{t_1 \to \infty} \frac{f(t_1)}{g(t_1)} \to 0$$

so that the lognormal has a thinner tail than a Pareto density.[13] For $\lambda_1 < 0$, our model has a Paretian tail in the sense that for each value of α it is possible to select $\lambda_1 = 1 - \alpha$ so that the density of t_1 has the same tail behavior as the selected member of the Pareto family. Our pro-

[12] Poirier (1978) and Amemiya and Powell (1981) have noted the importance of this restriction in applying the Box-Cox model.

[13] The same is true for a censored lognormal density where the censoring is due to self-selection decisions by agents (see Heckman and Sedlacek 1986). There we establish that the tail behavior of the censored normal and censored Box-Cox models is the same as the tail behavior of the uncensored models.

posed model can capture a feature of income and wage distributions claimed to be empirically important by Mandelbrot (1962) and Lydall (1968), whereas Roy's lognormal model cannot.

We extend Roy's model by including nonmarket participation as an option available to agents and by assuming that agents are utility maximizers rather than simple money income maximizers. The utility of participating in each sector is assumed to depend on sector-specific attributes such as wage rates, sector-specific consumption attributes (e.g., employment risk and job status), and the utility value of options that accrue to agents who participate in the sector (e.g., entitlement effects for social programs conditioned on sectoral participation as discussed in Mortensen [1977]). Letting V_i denote the utility of participating in sector i, where $i = 3$ designates the nonmarket sector, an agent chooses to participate in sector i if utility is maximized by doing so, that is,

$$V_i > V_j, \quad i \neq j, \; i, j = 1, 2, 3. \tag{14}$$

Let $\mathbf{z}_i$ denote a vector of measured sector-specific consumption attributes and household characteristics variables. Array all the $\mathbf{z}_i$, skill characteristics $\mathbf{x}$, and log task prices $\ln \pi_i$ into vector $\mathbf{f}$. Solving out for wages as a function of $\mathbf{x}$ and $\ln \pi_i$ in the utility functions, we reach a reduced-form linearized index function

$$\ln V_i = \boldsymbol{\gamma}_i \mathbf{f} + \upsilon_i, \quad i = 1, 2, 3. \tag{15}$$

We assume that $\mathbf{f}$ is distributed independently of all the υ_i and that $(\upsilon_1, \upsilon_2, \upsilon_3)$ is a mean zero multivariate normal random variable

$$(\upsilon_1, \upsilon_2, \upsilon_3) \sim N(0, \Sigma_{\upsilon}). \tag{16}$$

This specification produces the Thurstone multivariate probit model analyzed by Bock and Jones (1968) and Domencich and McFadden (1975).[14]

[14] A more explicit derivation of (15) and (16) from classical consumer choice theory adopts a loglinear specification for the mixed direct and indirect utility functions:

$$\ln V_i = \psi_{0i} + \psi_{1i} \ln w_i + \boldsymbol{\psi}_{2i} \mathbf{z}_i + \omega_i, \quad i = 1, 2, \tag{*}$$

$$\ln V_3 = \psi_{03} + \boldsymbol{\psi}_{23} \mathbf{z}_3 + \omega_3 \tag{**}$$

and assumes that $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3) \sim N(0, \Sigma_{\boldsymbol{\omega}})$ and that $\boldsymbol{\omega}$ is distributed independently of $\mathbf{z}_i$ for all i. Labor supply decisions within each sector are assumed to be optimally determined by agents. By permitting the coefficients in each sector to assume separate values for variables that are common to all $\mathbf{z}_i$ vectors, we recognize that sectors may differ in their consumption and investment possibilities. Substituting for $w_i = \pi_i t_i$ in (*) using (12) and assuming that λ_i is approximately but not exactly zero produces

$$\ln V_i = (\psi_{0i} + \psi_{1i} \ln \pi_i) + \psi_{1i} \boldsymbol{\beta}_i \mathbf{x} + \boldsymbol{\psi}_{2i} \mathbf{z}_i + (\psi_{1i} u_i + \omega_i),$$

where $\psi_{1i} u_i + \omega_i = \upsilon_i$, $i = 1, 2$, is approximately normally distributed because (13) is satisfied for all u_i^* in the neighborhood of $\lambda_i = 0$ for $i = 1, 2$, and u_i and ω_i are assumed

Since only sectoral choices and not the V_i are directly measured, it is possible to identify only parameters of the contrasts of utility evaluations among sectors. Without any loss of generality we normalize $V_3 = 1$ so $\boldsymbol{\gamma}_3 \equiv \mathbf{0}$ and $\upsilon_3 \equiv 0$. Using this convention, sector i is chosen if

$$\ln V_i - \ln V_j > 0, \quad \text{i.e., } (\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j)\mathbf{f} + \upsilon_i - \upsilon_j > 0 \text{ for all } j \neq i.^{15} \quad (17)$$

In Heckman and Sedlacek (1986) we prove that in a model with at least one nondegenerate regressor in $\mathbf{f}$, it is possible to identify $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$, $\text{var}(\upsilon_2)$, and $\text{cov}(\upsilon_1, \upsilon_2)$ from data on observed sectoral choices provided that $\text{var}(\upsilon_1)$ is normalized to unity.[16]

B. *The Statistical Model*

The statistical model used to generate the empirical estimates reported in this paper is presented in Appendix B. It joins reduced-form sectoral choice equations (17) with wage equations produced by the Box-Cox model (12), where the wage is $w_i = \pi_i t_i$ so

$$\frac{(w_i/\pi_i)^{\lambda_i} - 1}{\lambda_i} = \boldsymbol{\beta}_i \mathbf{x} + u_i, \quad i = 1, 2. \quad (18)$$

We adopt a reduced-form approach to estimation because all the determinants of market wages are plausible determinants of utility in their own right. No restrictions are imposed between the parameters of (18) and the parameters of sectoral choice equations (17). Estimating an unrestricted sectoral choice model yields an upper bound on the goodness of fit of a more restricted explicitly structural sectoral choice model.

Provided that prices (π_i) are normalized to unity in a year and that there is stability over time in some parameters on the right-hand side of (18) (i.e., in elements of $\boldsymbol{\beta}_i$ or the variance of u_i), it is possible from successive cross sections of data to estimate task prices π_{il}, $i = 1, 2$, $l = 1, \ldots, L$. The selected normalization defines the units in which task prices are measured. In addition, if $\lambda_i \neq 0$, it is possible to estimate year effects (shifts in the intercept) in wage equation (18). Denote these year effects by β_{0il}, $i = 1, 2$, $l = 1, \ldots, L$.[17]

---

to be independent of $\mathbf{x}$, $\mathbf{z}_i$, and $\ln \pi_i$. Set $\upsilon_3 = \omega_3$. The assumption that the λ_i are close to zero appears to be consistent with the data. In Sec. III we note that the estimated λ_i for manufacturing is 0.08 (λ_2) while the value for nonmanufacturing is -0.06 (λ_1).

[15] Indifference occurs on a set of measure zero by virtue of the normality assumption for $(\upsilon_1, \upsilon_2, \upsilon_3)$ assuming that these random variables are nondegenerate.

[16] One such normalization is required. The probabilities of the events described by inequalities (17) are unchanged if the inequalities are all divided by the standard deviation of υ_1.

[17] More precisely and in the notation for time-series task prices and year effects in

Because wage equation (18) contains year effects (the $\ln \pi_{il}$ and β_{0il}) and wages are an argument of the utility function, and to allow for year-specific shocks to preferences, year effects are introduced into the reduced-form sectoral choice functions (15). These year effects are denoted by γ_{0il}, $i = 1, 2$, $l = 1, \ldots, L$.

Using the likelihood presented in Appendix B it is possible in a single cross section of data to estimate consistently the λ_i, $\boldsymbol{\beta}_i$, and $\boldsymbol{\gamma}_i$, $i = 1, 2$, as well as the variances of the latent normal variables u_i^* (which generate u_i), their covariance with υ_1 and υ_2, and the covariance structure of (υ_1, υ_2) (setting the variance of υ_1 to one or adopting some other normalization). The covariance between u_1^* and u_2^* is not identified.[18] From repeated cross-section data, it is possible to estimate consistently the prices π_{il} and the year effects β_{0il}, γ_{0il}, $i = 1, 2$, $l = 1, \ldots, L$, given a conventional normalization (suppressing intercepts or setting one value of these parameters to a known constant). The

---

wage functions introduced in the text, if $\lambda_i = 0$ for $i = 1, 2$, eq. (18) specializes to (making an obvious change of notation)

$$\ln w_{il} = \ln \pi_{il} + \boldsymbol{\beta}_{il}\mathbf{x}_i + u_{il}, \quad i = 1, 2, \; l = 1, \ldots, L,$$

so that $\ln \pi_{il}$ is indistinguishable from the intercept term β_{0il}. Assuming that β_{0il} is constant in successive cross sections and letting $\pi_{il} = 1$ in one year, it is possible to estimate a time series of task prices from selection-corrected wage functions. If $\lambda_i \neq 0$, π_{il} enters the model as a scale parameter. Assuming that $\pi_{il} > 0$, (18) may be rewritten as

$$\frac{w_{il}^{\lambda_i} - 1}{\lambda_i} = \frac{\psi_{il} - 1}{\lambda_i} + \psi_{il}\boldsymbol{\beta}_{il}\mathbf{x}_i + u_{il}\psi_{il}, \tag{*}$$

where $\psi_{il} = (\pi_{il})^{\lambda_i}$ for $i = 1, 2$ and $l = 1, \ldots, L$. With no restrictions over time in the variances of u_{il} or the slopes or the intercepts of the wage function, it is not possible to estimate a time series of task prices π_{il} from selection-corrected wage functions (i.e., π_{il} can always be set to unity in each year without affecting the fit of the model). By assuming, e.g., that one slope coefficient remains constant over time or that the variance of u_{il} remains time invariant, it is possible to estimate π_{il} given one normalization ($\pi_{il} = 1$ for a particular year) from selection-corrected wage functions. Evidence in support of the proportionality hypothesis (invariance of the slope coefficients of selection-corrected wage functions) justifies the procedure used to estimate task prices. Notice that separate values of λ_i can be estimated in each cross section irrespective of whether or not π_{il} can be identified. Year effects in the wage equation (the β_{0il}) can be estimated along with task prices (π_{il}) if the latter are identified by assuming temporal invariance in slope or variance parameters. One year effect (β_{0il}) must be set to zero unless the intercept of (18) is deleted. Note further that for the case $\lambda_i \neq 0$ ($i = 1, 2$) the estimated π_{il} are indistinguishable from a very special type of technical change in the task functions that scales the slope coefficients and unobservables by a common parameter and shifts the intercept in a restricted way (see the ψ_{il} above in eq. [*]). The only way to determine if the estimated π_{il} are valid prices is to see whether or not they act like prices in a behavioral equation. The evidence presented in Sec. III suggests that they do.

[18] Lee (1978) demonstrates that this parameter is not identified in a two-market-sector *utility-maximizing* lognormal Roy model. This lack of identification is a consequence of the introduction of new unobservables in the sectoral choice functions that are not directly attributable to the unobservables in the wage equations.

maximum likelihood estimator of the parameters of this model is consistent and asymptotically normal.[19]

The estimated task prices (π_{il}) can be used as input to estimate consistently the aggregate demand for task functions following the general methodology outlined at the end of Section I. When $\lambda_i \neq 0$, it is possible under conditions stated in note 17 to estimate β_{0il} and π_{il} separately. In that case, equations (10) and (11) can be rewritten to account for the fact that π_{il} and β_{0il} are not confounded. This point is discussed further below.

III. Empirical Estimates

In this section we report empirical estimates and tests of the extended Roy model described in Section II and of the sectoral demand for aggregate task functions described in Section I. We use these estimates to explore the empirical importance of aggregation bias in obscuring aggregate real wage movements. We also assess the contribution of self-selection to inequality in the distribution of log wage rates.

One convenient feature of our model is that it is not necessary to estimate the extended Roy model for all demographic groups in order to estimate task prices, π_{il}, or sectoral task demand functions. Assuming that all units of task i are perfect substitutes irrespective of their demographic source, estimates of π_{il} for one demographic group suffice to identify market task prices.[20] Dividing the aggregate wage bill for all demographic groups by the estimated task price produces a consistent estimate of the total amount of the task supplied to the market that can be used in the estimation of aggregate demand for task functions.

Another convenient feature of our model is that it is not necessary to estimate the extended Roy model for the reference demographic group for each available cross section. Assuming that the proportionality hypothesis is not rejected and the estimated model passes a goodness-of-fit test, we can estimate the slope coefficients of the model in a single cross section and fix these coefficients in other cross sections using the rest of the data to estimate year effects (the γ_{0il} and β_{0il}) and the log task prices ($\ln \pi_{il}$).

[19] An anonymous referee suggested that inequality (13) leads to a violation of classical regularity conditions because the range of the random variable u_i^* depends on parameters of the model. Inequality (13) requires only that the t_i and w_i be nonnegative, and so no violation of classical regularity conditions is induced by this restriction.

[20] This assumes no market discrimination. By estimating the extended Roy model for separate demographic groups, we can test for market discrimination. If there is no market discrimination the estimated π_{il} should be the same across different demographic groups.

We exploit both features of the model to reduce the computational cost required to secure the empirical results reported below. We use prime age white males aged 18–65 as our reference demographic group. We test the proportionality hypothesis and perform goodness-of-fit tests for the model on two years of data (1976 and 1980). The evidence suggests that it is legitimate to constrain the slope coefficients to equality in all years, using the remaining cross sections of data to estimate year effects and task prices.

This empirical strategy substantially reduces the computational cost. However, this saving is secured by assuming what in principle can be tested: (*a*) that estimated task prices are identical across all demographic groups and (*b*) that proportionality and goodness-of-fit tests are passed for all demographic groups in all years. We leave the execution of such tests for another occasion, recognizing that the empirical results reported below may be overturned in a more extensive battery of tests.

A. *Estimates of the Extended Roy Model*

We estimate the extended Roy model on a 4 percent random sample of prime age white males taken from the annual March Current Population Survey (CPS) for the years 1968–81 inclusive.[21] These data are described in detail in Appendix C. When the extended Roy model is fit on the complete sample it is decisively rejected. The proportionality hypothesis is rejected, and goodness-of-fit tests indicate that the model does not fit the empirical log wage distributions. However, when low-wage observations (persons whose real wages are less than \$0.75 per hour) are deleted, the model is not rejected. The empirical tests reported in this paper are based on samples that exclude such observations. The likelihood function presented in Appendix B explicitly accounts for this sample selection criterion.

The estimated model parameters are presented in table 1. This table records estimates based on a pooled 1976 and 1980 sample. Individuals are classified into one of three sectors depending on their source of income for the year. Roughly 16 percent of the sample has no labor income in 1980. Individuals without labor earnings are defined as participants in the nonmarket sector for that year (sector 3). Following census definitions, individuals are defined to be in the manufacturing sector if their SIC three-digit industry code is between

[21] Lillard, Smith, and Welch (1982) note the high nonreporting rate for key economic variables in the CPS and discuss the imputation procedures used by the Census Bureau. They demonstrate that there is a potential for substantial bias in using imputed CPS data. We eliminate all imputed observations from our analysis.

TABLE 1

ESTIMATES OF THE MODEL PARAMETERS

	Estimated Coefficient	Standard Error*	Normal Statistic†
Utility function in the nonmanufacturing sector (γ_1):			
Intercept	4.238367	.469394	9.029442
Education	.338785	.042739	7.926800
Experience	.224682	.028620	7.850411
Experience squared/100	−.333751	.071232	−4.685396
South dummy	.282627	.136377	2.072390
Predicted nonlabor income/100	.242310	.033105	7.319353
1980 intercept (γ_{01} for 1980)	.113196	.094107	1.202837
Utility function in the manufacturing sector (γ_2):			
Intercept	3.103701	.565689	5.486586
Education	.285896	.053022	5.392017
Experience	.163867	.036530	4.485828
Experience squared/100	−.257929	.072256	−3.569655
South dummy	.019389	.106355	.182301
Predicted nonlabor income/100	.172409	.036337	4.744774
1980 intercept (γ_{02} for 1980)	.017729	.074623	.237583
Correlation coefficient between υ_1 and υ_2: correl(υ_1, υ_2)	.296560	.147650	2.008529
Standard deviation of υ_2: $[\text{var}(\upsilon_2)]^{1/2}$	.850640	.117044	7.267723
Parameters of the mapping of the observed skills to the nonmanufacturing task (β_1):			
Intercept	−.112678	.101883	−1.105953
Education	.040472	.007908	5.117798
Experience	.005979	.008301	.720287

TABLE 1 (*Continued*)

	Estimated Coefficient	Standard Error*	Normal Statistic†
Experience squared/100	.019015	.018805	1.011173
South dummy	.016770	.042527	.394325
1980 intercept (β_{01t} for 1980)	−.312877	.356679	−.877195
Parameters of the mapping of the observed skills to the manufacturing sector task ($\boldsymbol{\beta}_2$):			
Intercept	−.331493	.299324	−1.107471
Education	.082424	.010596	7.778808
Experience	.027506	.012970	2.120790
Experience squared/100	−.027446	.028786	−.953469
South dummy	−.102184	.060104	−1.700135
1980 intercept (β_{02t} for 1980)	.038270	1.152317	.033212
Covariance structure of the latent task distribution:			
$(\sigma_{11}^{1/2}) = [\text{var}(u_1^*)]^{1/2}$	.574169	.006098	94.159852
$(\sigma_{22}^{1/2}) = [\text{var}(u_2^*)]^{1/2}$	.486769	.081631	5.963048
$\rho_{12}^* = \text{correl}(u_1^*, v_2 - v_1)$	.241512	.029820	8.351013
$\rho_{11}^* = \text{correl}(u_1^*, v_1)$	.454436	.029116	15.607939
$\rho_{21}^* = \text{correl}(u_2^*, v_2 - v_1)$	.235583	.009276	25.397051
$\rho_{22}^* = \text{correl}(u_2^*, v_2)$	.159303	.004145	38.435299
1980 estimated log task price change where $\pi_1(1976) = \pi_2(1976) = 1$:			
Nonmanufacturing sector ($\ln\hat{\pi}_{1t}$ for 1980)	.216560	.003588	60.358733
Manufacturing sector ($\ln\hat{\pi}_{2t}$ for 1980)	−.225510	.005036	−44.777223

Task transformation parameter (λ_i):			
Nonmanufacturing sector (λ_1)	−.060494	.032839	−1.842148
Manufacturing sector (λ_2)	.082650	.008020	10.305595
Log-likelihood for the model −2,099.01			
Number of individuals in the sample 3,262			

	χ^2 Statistic for the Hypothesis	Number of Degrees of Freedom	Values of χ^2 Random Variables at 5 Percent Significance Level for Stated Number of Degrees of Freedom
Likelihood ratio test for restricted model, $\lambda_1 = \lambda_2 = 0$:			
1976 data	8.18	2	5.99
1980 data	7.46	2	5.99
Goodness-of-fit test‡ for the extended Roy model:			
Manufacturing	34.1	50	67.51
Nonmanufacturing	64.7	50	67.51
Goodness-of-fit‡ for the lognormal three-sector model with $\lambda_1 = \lambda_2 = 0$:			
Manufacturing	42.7	50	67.51
Nonmanufacturing	71.9	50	67.51
Strong proportionality hypothesis	15.7	26	38.89

* Standard errors are computed from the square root of the diagonal elements of minus the inverse of the Hessian of the log likelihood.

† The ratio of the estimated coefficient to the estimated standard error. This ratio, when multiplied by the square root of the sample size, is asymptotically normal under the null hypothesis that the corresponding population parameter is zero.

‡ The χ^2 goodness-of-fit statistics were computed for the conditional (on sectoral choice) log wage distributions in each sector using 51 equispaced log wage intervals starting from ln 0.75 in intervals of length 0.07535 and terminating at ln 35.0. The statistics compare predicted and actual log wage distributions in each interval, integrating out the regressor variables. In computing the χ^2 statistics we account for parameter estimation error following Moore (1977). We pool 1976 and 1980 data to perform the test.

107 and 398. Roughly 21 percent of the 1980 sample falls into this category. The rest of the sample (63 percent) is classified as working in nonmanufacturing.

The first two sets of rows of the table record the parameters of the contrast between the indicated sector reduced-form preference function and the nonmarket sector preference function (the γ in eq. [15]). The arguments include conventional determinants of wages (education, work experience, and work experience squared) plus a South dummy (= 1 if a person resides in the South and = 0 otherwise) to capture regional wage and amenity differences. In addition, predicted nonlabor income is assumed to enter the preference function. Nonlabor income consists of all nonemployment income including unemployment benefits and social transfers. Nonlabor income is predicted for each individual in each sector to account for the fact that entitlements to various social programs (e.g., unemployment insurance) are conditioned on sectoral participation and also to eliminate spurious correlation between nonlabor income and unobserved components of preferences. The predictor variables are presented in Appendix C. The 1980 intercept is a dummy variable that equals one if an observation comes from the 1980 sample and is zero otherwise. (Its coefficient estimates γ_{0il} for 1980.) The estimates reveal that education and work experience increase the probability of market participation. These variables have a slightly stronger effect on participation in the nonmanufacturing sector than on participation in the manufacturing sector. The South dummy has little effect on the nonmarket-manufacturing choice but a stronger effect on the nonmarket-nonmanufacturing choice.

The coefficients on predicted nonlabor income are *positive* for both estimated sectoral utility functions, and statistically significantly so. At first sight this result is counterintuitive and appears to indicate that leisure is an inferior good. Positive coefficients are consistent with Mortensen's (1977) entitlement effect in which individuals participate in a sector to collect sector-specific social benefits (e.g., unemployment benefits or workmen's compensation). They are also consistent with the hypothesis that individuals are willing to forfeit income to enjoy the training or consumption benefits that accrue to individuals working in specific sectors.[22]

The insignificant coefficients for the 1980 dummy variables indicate that 1980 reduced-form preferences do not differ in intercept

[22] Recall that predicted nonlabor income and not actual nonlabor income is used to avoid a spurious correlation between assets (and benefits) and sectoral preferences. In many honestly conducted and reported empirical studies of male labor supply, leisure is found to be "inferior." The argument in the text provides one rationale for this finding.

from 1976 preferences. The next parameters reported in table 1 are estimates of the covariance structure for the unobservables in the utility contrasts (υ_1, υ_2). These unobservables are positively related—correl(υ_1, υ_2) = .29—and the variability of the unobservables in the manufacturing sector contrast, var(υ_2), is smaller than the variability of the unobservables in the nonmanufacturing sector contrast (which is normalized to unity).

The next two blocks of rows reported in table 1 report the coefficients of the estimated task (eq. [12]) or wage (eq. [18]) functions (the $\boldsymbol{\beta}$). Except for predicted nonlabor income, the same variables that enter the sectoral choice equations enter the wage equations. The data on hourly wages are constructed by dividing annual labor income by estimated annual hours of work.

Education has a strong positive effect in both sectors, but its effect is twice as strong in manufacturing. Wages grow much more steeply with work experience in the manufacturing sector than in the nonmanufacturing sector over the empirically relevant range. The hypothesis of no wage growth with work experience cannot be rejected for nonmanufacturing wages. The South dummy is statistically insignificant in both task functions. The 1980 dummy variable is statistically insignificant for both sets of coefficients, indicating little difference in the estimated intercepts of the task functions between 1976 and 1980 (the β_{0il} for those respective years).

The variance of u_1^* is greater than the variance of u_2^*. This is consistent with greater heterogeneity among the group of industries classified in the nonmanufacturing sector.

The estimated log task price changes for 1980 indicate a 22 percent decline in the price of the manufacturing task from its 1976 level and a 21 percent increase in the price of the nonmanufacturing task from its 1976 level. The estimated transformation parameter for manufacturing (λ_2) is positive, indicating that manufacturing wage rates do not have a Paretian right tail. The estimated transformation parameter for nonmanufacturing wages (λ_1) is slightly negative, indicating a Paretian right tail for wages in that sector.[23]

Even though both values of λ_i are estimated to be close to zero, a likelihood ratio test of the hypothesis that $\lambda_1 = \lambda_2 = 0$ performed on the 1976 data rejects that hypothesis. The hypothesis is also rejected with the 1980 data.[24]

[23] Heckman and Polachek (1974) estimate a negative value of λ for hourly wages using a Box-Cox procedure fit on data aggregated over both sectors. Their procedure does not account for the truncation discussed in App. A or the censoring discussed in App. B.

[24] A direct test of the hypothesis $\lambda_1 = \lambda_2 = 0$ for pooled 1976–80 data with period-specific intercepts for the task function (β_{0il} for $i = 1, 2$ and $l = 1, \ldots, L$) and prices (π_{il}

Further evidence in support of our more general model comes from goodness-of-fit statistics for the extended Roy model. Using the fixed wage intervals described in the notes to table 1 to compute a χ^2 goodness-of-fit test, we do not reject the hypothesis that the model fits manufacturing log wage data using a 5 percent level of significance. Compare the 34.1 χ^2 statistic to the 42.7 χ^2 statistic reported in the next set of rows in table 1 that results when $\lambda_2 = 0$. Inspection of figure 3 reveals that the estimated model closely fits the pooled 1976 and 1980 data.

The performance of the extended Roy model in predicting log wages in nonmanufacturing is also satisfactory. At a 5 percent significance level we do not reject the hypothesis that the model fits nonmanufacturing log wage data. The χ^2 statistic for the extended Roy model is 64.7, and for the normal model ($\lambda_1 = 0$) it is 71.9. Figure 4 reveals that the fit of the model to the nonmanufacturing data is rather good.

Heckman and Sedlacek (1986) compare plots of the extended Roy model with lognormal models with and without the assumption of utility maximization and with and without the presence of a nonmarket sector. We note that it is the addition of the nonmarket sector that substantially improves the fit of the model.

The final row of table 1 reports the result of a test of a strengthened version of the proportionality hypothesis stated in Section I. Although the procedure used to estimate task prices requires only temporal invariance of some of the slope coefficients of the task function (12) or the variance of u_i^*, or some other restrictions across time in these parameters, an estimated model in which preferences shift about in each year for unexplained reasons would not be economically very interesting. Accordingly, we test for stability of the slope coefficients and covariance structure of the task functions and the preference functions in 1976 and 1980. (The intercepts of the preference functions might be expected to shift over time since they depend, inter alia, on task prices.) The restrictions tested here are thus much stronger than the ones required to identify π_{il}.

The statistic reported in the final row of table 1 is produced by comparing the likelihoods for the pooled 1976 and 1980 data with the sum of separate likelihoods fit for 1976 and 1980 separately. Using a 5 percent significance level, we do not reject the strong propor-

---

for $i = 1, 2$, and $i = 1, \ldots, L$) raises a messy statistical problem. As noted in n. 17, when $\lambda_1 = \lambda_2 = 0$, it is not possible to estimate separate β_{0il} and π_{il} parameters, and so some parameters of the model become unidentified. Conventional likelihood ratio tests do not possess classical limiting distributions. While it is possible to construct a test of the hypothesis in this case (see Davies 1977), we have not done so here.

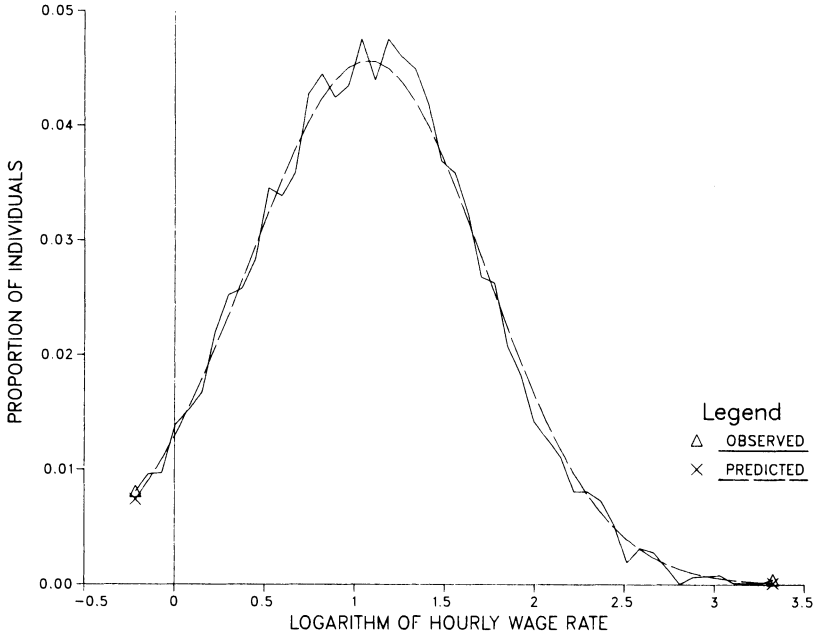

FIG. 3.—Nonmanufacturing sector: predicted versus observed log wage distribution

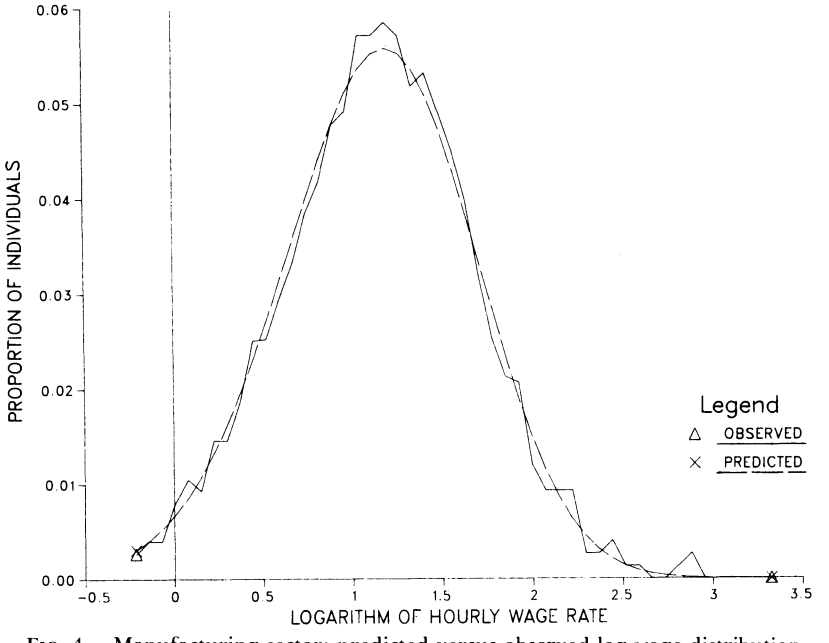

FIG. 4.—Manufacturing sector: predicted versus observed log wage distribution

tionality hypothesis.[25] When the model is restricted to a lognormal form ($\lambda_1 = \lambda_2 = 0$) we decisively reject the hypothesis (see Heckman and Sedlacek 1986).

Elsewhere (Heckman and Sedlacek 1986) we also decisively reject the proportionality hypothesis as a description of wage data aggregated over both market sectors (with self-selection ignored). An efficiency units assumption for labor quality does not describe the U.S. labor market taken as a whole. However, the empirical results just presented indicate that an efficiency units assumption for sector-specific tasks is valid within manufacturing and nonmanufacturing sectors.

B. *Estimating the Demand for Aggregate Sector-specific Tasks*

The extended Roy model could be fit to each available year of the CPS data (1968–81) to estimate the log task prices, $\ln \pi_{il}$, $i = 1, 2$ and $l = 1, \ldots, L$. To do so would be prohibitively expensive. Because the strong proportionality hypothesis is not rejected, it appears reasonable to fix all the parameters of the model at the values reported in table 1 except for the intercepts of the wage (β_{0il}) and preference functions (γ_{0il}) and the log task prices ($\ln \pi_{il}$) and to estimate the intercepts and log task prices from each year of the available CPS data. Estimates obtained from this procedure are reported in Appendix D.[26]

Given a time series of $\ln \hat{\pi}_{il}$ it is possible to estimate the parameters of the sectoral demands for aggregate tasks using a modification of the procedure described in Section I. In making the modification note that when $\lambda_i \neq 0$ it is possible to separate $\ln \pi_{il}$ from the intercepts of the task functions (the β_{0il}, $i = 1, 2$, $l = 1, \ldots, L$; see n. 17). Making this change we modify equation (10) to read

$$\ln\left(\frac{WB_{il}}{P_{il}}\right) = \delta_{0i} + (\delta_{1i} + 1)(\ln \hat{\pi}_{il} - \ln P_{il}) + \boldsymbol{\delta}_{2i} \ln\left(\frac{\mathbf{P}_{Al}}{P_{il}}\right) + \tilde{e}_{il}, \tag{19}$$

[25] When a separate model is fit in each year, the π_{il} are not identified (see n. 17). In each cross section 32 parameters are estimated (the number of parameters reported in table 1 less the 1980 dummy variables and the π_{il} terms). Thus 64 parameters are estimated in the unrestricted version. When all but the intercepts and task prices are constrained to equality in the pooled 1976 and 1980 sample, 38 parameters are estimated. Consequently, there are 26 degrees of freedom reported in table 1.

[26] One referee objected that the log task prices reported in App. D show too much temporal variability to be believed. This is an unusual argument in that neither tasks nor their prices are directly observed. Surely the only way to judge whether or not an estimate of a price is valid is to see if the estimated price acts like a price in an estimated behavioral relationship.

$i = 1, 2$ and $l = 1, \ldots, L$, where $\ln \hat{\pi}_{il}$ is the intercept estimated from the microeconomic log wage equation fit in sector i in year l (eq. [8]) and $\tilde{e}_{il}$ differs from $\mathbf{e}_{il}$ by the estimation error of $\ln \hat{\pi}_{il}$ for $\ln \pi_{il}$; $\tilde{e}_{il} \equiv e_{il} + (1 + \delta_{1i})(\ln \pi_{il} - \ln \hat{\pi}_{il})$. We assume that the task price for white males is the task price for all demographic groups.

Estimates of the sectoral aggregate demand for task functions (eq. [19]) are presented in table 2. The sectoral wage bill data come from U.S. Commerce Department data on total wages paid. (For further information on these data, see App. C.) The log of the real wage bill divided by real product price for each sector is regressed on the logs of (1) the estimated task price, (2) an index of energy prices, (3) an index of intermediate goods prices for the sector, and (4) the user cost of capital. Each of these prices is deflated by the real product price. Definitions and data sources for these variables are presented in Appendix C.

As noted in Section I, it is implausible that the $\ln \pi_{il}$ and P_{il} are exogenous variables in the aggregate task demand functions. Instrumental variable estimates based on the set of instruments recorded in the notes to the table are given in the right-hand-side columns of table 2. Not surprisingly, the instrumental variable estimates are less precisely determined than are the least-squares estimates. Note that the instrumental variable estimates of the elasticity of demand for unmeasured aggregate tasks are very close to the ordinary least squares (OLS) estimates, indicating that simultaneous equations bias is not present.[27] This empirical result is robust to a variety of choices of the set of instrumental variables. For this reason we focus our discussion on the OLS estimates.

The estimated elasticities of demand are negative and statistically significantly different from zero and thus are in accord with the predictions of economic theory.[28] The Durbin-Watson statistics indicate that serial correlation is not a problem and the R^2's are high.

It is important not to make too much out of these estimated demand functions. After all, there are only 14 time-series observations for each sector, and the number of degrees of freedom in the time

[27] A Durbin (1954) test does not reject this hypothesis.

[28] The estimated nonmanufacturing sector elasticity is also significantly different from -1, although this is not the case for the manufacturing elasticity. Thus the normalized wage bill in nonmanufacturing is significantly related to $\ln \pi_{il}$ (i.e., the estimated value of $\delta_{11} + 1$ is statistically significantly different from zero). At least for nonmanufacturing we can reject the argument that our estimated demand elasticity is the spurious product of a procedure that subtracts 1 from a coefficient that is not statistically different from zero (the estimated value of $\delta_{1i} + 1$ in eq. [19]) and finds that the insignificant coefficient minus 1 is not statistically significantly different from -1. Note, however, that we cannot reject this argument for the estimates for the manufacturing sector. Of course, it is possible that the true elasticity of demand for manufacturing is -1. There is no way to use these data to determine whether or not the estimated relationship is spurious.

TABLE 2

DEMAND FUNCTIONS FOR AGGREGATE TASKS (Eq. [19])

	ORDINARY LEAST SQUARES ESTIMATES		INSTRUMENTAL VARIABLE ESTIMATES*	
	Estimated Coefficient	Standard Error	Estimated Coefficient	Standard Error
	Nonmanufacturing Sector			
Constant (δ_{01})	12.119010	.11277737	11.900640	1.6234258
Log task price (δ_{11})†	−.951021	.02161820	−.934039	.3674947
Log energy price index (δ_{21})	.394647	.07575938	1.120513	1.0200879
Log intermediate goods price (δ_{31})	−.488665	.49281826	−.116150	6.4352536
Log user cost of capital (δ_{41})	−.099360	.05669152	.7744651	.7744651
R^2	.9958		...	
Number of observations (1968–81)	14		14	
Durbin-Watson statistic‡	1.447		1.462	
	Manufacturing Sector			
Constant (δ_{02})	11.057958	.11730702	10.797219	1.8079848
Log task price (δ_{12})	−.977697	.02421021	−.974127	.4916065
Log energy price index (δ_{23})	.162611	.09507995	.925919	1.2423610
Log intermediate goods price (δ_{32})	−.706052	.51737473	−.345029	6.8214180
Log user cost of capital (δ_{42})	−.045386	.06814409	.099210	1.1129916
R^2	.9905		...	
Number of observations (1968–81)	14		14	
Durbin-Watson statistic	1.966		2.200	

NOTE.—For the definitions of these variables see App. C.

* The instruments are: log energy price index, log intermediate goods price index, log user cost of capital, total population, total population squared, average weekly hours worked in the nonmanufacturing sector, unemployment rate in the United States. For further discussion see App. C. The regression results are unaffected when the hours worked variable is not used as an instrument.

† The reported coefficients are the estimated coefficients on log task prices from regression equations of the form (19) minus one.

‡ The lower limit for the Durbin-Watson test for a 5 percent significance level with five regressors (including an intercept) and 15 observations is 0.69. The upper limit is 1.97. The limits for 14 observations are wider.

series is small. The specification adopted here abstracts from the dynamic costs of adjustment that have been found to be important in other studies of labor demand. Nonetheless, our simple model appears to be consistent with the limited time-series data at our disposal. It is possible that in a longer time series with more degrees of freedom a more dynamic model of factor demand would be required to produce an acceptable fit of the data.

C. *Exploring the Importance of Aggregation Bias in Aggregate Wages*

The apparent lack of aggregate wage variability over the cycle for U.S. data may be a consequence of aggregation bias (see, e.g., Stockman 1983; Bils 1985). Since low-wage workers are the "first to go" in response to a downturn in demand, the lack of variability in measured average wages may partly reflect an employed worker quality composition effect.

Our model can be used to investigate the empirical importance of aggregation bias. For the U.S. manufacturing sector we find strong evidence of aggregation bias leading to an attenuation of measured average wage movements in relationship to the true "quality constant" movement in task prices. However, for the economy as a whole, just the opposite effect occurs. Aggregation bias *increases* measured wage variability in relationship to the underlying movement in quality constant task prices.

The manufacturing sector is harder hit by an aggregate disturbance such as an oil price increase than is the nonmanufacturing sector. Employment declines in the manufacturing sector. Some of the former manufacturing workers enter the nonmanufacturing sector rather than drop out of the work force altogether. The former manufacturing workers turn out to be at the bottom of the manufacturing task quality distribution, and their exit raises the average quality of the remaining manufacturing work force and hence attenuates the decline in measured average wages. This is the conventional aggregation bias effect discussed in the literature.

However, the former manufacturing workers who enter the nonmanufacturing sector turn out to be at the bottom of the task quality distribution in that sector. The new entrants lower the average quality of the work force in nonmanufacturing. The reduction in quality more than offsets the increase in task price in that sector. On net, aggregation bias exaggerates the aggregate decline in real wages over what it would be if task quality were held fixed. This effect is ignored in macroeconomic studies that neglect labor heterogeneity and self-selection.

Table 3 presents estimates of the impact of a 1 percent increase in the price of energy on employment, average sector task quality, task price, and average wages for each sector and for the economy as a whole. If estimates of the supply functions for all demographic groups were available, it would be straightforward to simulate the model. Recall, however, that we have estimated supply functions for only one demographic group—white males—and so we cannot use a direct simulation approach. Instead, we estimate a reduced-form equation relating log task prices to energy prices and other determinants of aggregate task demand and supply. These reduced-form equations are presented in Appendix E. Assuming structural invariance of the parameters of the economy, we can use our estimated log task price equation to estimate the effect of an energy price change on sectoral task prices. The numbers reported in table 3 are based on such reduced-form equations.

The numbers reported in the first column (for the manufacturing sector) and the first panel (for 1972) indicate the following response to a 1 percent increase in energy prices: (1) employment in manufacturing decreases by 1.854 percent; (2) the average task quality of workers employed in the sector rises by 0.919 percent; (3) the task price declines by 1.48 percent. Adding effects 2 and 3, we would observe average manufacturing wages to decline by only 0.561 percent. Two-thirds of the decline in the manufacturing task price is offset by a change in the quality of the work force. The composition bias effect in manufacturing is roughly of the same order of magnitude for the two other years (1976 and 1980).

For the nonmanufacturing sector in 1972, a 1 percent oil price increase raises employment, lowers average employed task quality (by 1.49 percent), and raises task price (by 0.471 percent).[29] The predicted *decline* in average wages in the sector is to be compared with the forecast *increase* in the nonmanufacturing task price. Similar results are found for 1976 and 1980.

For the economy as a whole (the third col. of the table), the task quality constant wage change is defined to be a weighted sum of the sectoral task price changes, where the weights are the employment proportions in each sector in the appropriate year. In 1972 the simulated aggregate wage decline (-0.950) is much larger than the skill constant wage change (-0.062). These simulations suggest that aggregation bias may be empirically important. However, its effect on

[29] Recall that an increase in the price of energy increases the demand for nonmanufacturing tasks (see table 2). These demand functions reflect a shift in relative demand from manufacturing to nonmanufacturing in response to a change in energy prices.

TABLE 3

SIMULATION OF A 1 PERCENT INCREASE IN THE ENERGY PRICE INDEX

	Manufacturing Sector	Nonmanufacturing Sector	U.S. Aggregate
Year: 1972:			
1. Percentage change in persons employed	−1.854	1.320	...
2. Percentage change in mean task or quality level for the employed population	.919	−1.496	...
3. Percentage change in task price	−1.480	.471	−.062*
4. Percentage change in observed average wage (2 + 3)	−.561	−1.025	−.950
Year: 1976:			
1. Percentage change in persons employed	−2.007	1.371	...
2. Percentage change in mean task or quality level for the employed population	.886	−1.461	...
3. Percentage change in task price	−1.480	.471	−.063*
4. Percentage change in observed average wage (2 + 3)	−.594	−.990	−.939
Year: 1980:			
1. Percentage change in persons employed	−1.993	1.244	...
2. Percentage change in mean task or quality level for the employed population	.953	−1.568	...
3. Percentage change in task price	−1.480	.471	−.034*
4. Percentage change in observed average wage (2 + 3)	−.527	−.997	−.949

NOTE.—The data sets on which the simulations are performed are defined in App. C.

* This is a weighted average of the task price change in each sector using the relative proportions employed in the sector in the year.

aggregate wage variability is opposite to that conjectured in the recent literature, which ignores the effect of self-selection or the pursuit of comparative advantage. Aggregation bias *increases* measured wage variability.[30]

D. *Assessing the Impact of Self-Selection on Inequality in Log Wages*

In this subsection we use our estimates of the extended Roy model to assess the impact of self-selection on inequality in market wage rates for employed white males. A commonly used measure of inequality—the variance in the natural logarithm of wages—and a prototypical year, 1980, are selected to make this assessment. We compare the observed variance in log wages (which is close to the variance predicted by the model) with the variance in log wages that would result if people were randomly assigned to manufacturing, nonmanufacturing, or nonmarket activity in a sense to be made precise below.

The first column of table 4 presents predicted values of sectoral and economywide means and variances of log wage rates. Actual values from the March 1980 CPS data are given in the second column. Notice that there is close agreement between actual and predicted values. The economywide variance is broken down into two components: (*a*) variability within sectors and (*b*) variability between sectors. The formula for the variance decomposition is given in the notes of the table. Note that virtually all of the total variance in log wage rates is due to within-sector variability (.99 = .288/.291).

The final column of table 4 presents values of sectoral and economywide means and variances of log wage rates for the random assignment economy. This economy is constructed by randomly assigning people so that (*a*) the proportions employed in each sector are set to be the same as those predicted in the sectors in 1980 by our equilibrium model and (*b*) sectoral task prices for the hypothetical economy are set at 1980 values, an assumption that is strictly defensible only if the aggregate task demand functions are perfectly elastic as in the

[30] Our conclusions appear to be at odds with those of Stockman (1983) and Bils (1985). Both conclude that there is little evidence of aggregation bias in aggregate wages. Stockman excludes nonworkers from his sample and thus induces a sample selection bias problem, which he notes but does not solve. Bils includes nonworkers in his analysis and corrects for selection bias assuming that wages are lognormal. Recall, however, that our tests reject the lognormal model. Neither author corrects for the effect of sectoral self-selection decisions. Our procedure, which adjusts for selection bias in a nonnormal model and accounts for the effect of comparative advantage on measured wages, produces much stronger evidence of aggregation bias than do these other studies.

TABLE 4

ASSESSING THE IMPACT OF SELF-SELECTION ON THE MEANS AND VARIANCES OF LOG WAGE RATES FOR WHITE MALES, 1980

	Prediction of Extended Roy Model	Actual 1980 Value	Random Assignment Economy Using 1980 Equilibrium Task Prices
		Nonmanufacturing Sector	
Mean of log wages (M_1)	1.054	1.040	.651
Variance of log wages (σ_1)	.319	.323	.344
Proportion of population in sector (P_1)	.619*	.630	.619*
		Manufacturing Sector	
Mean of log wages (M_2)	1.199	1.202	.968
Variance of log wages (σ_2)	.192	.201	.211
Proportion of population in sector (P_2)	.200*	.206	.200*
		Economywide	
Mean of log wages $\left(\frac{P_1M_1 + P_2M_2}{P_1 + P_2}\right)$	1.089	1.079	.728
Sum of within-sector variance $\left(\frac{P_1\sigma_1 + P_2\sigma_2}{P_1 + P_2}\right)$	.288	.293	.311
Between-sector variance $\left[\frac{P_1P_2(M_1 - M_2)^2}{(P_1 + P_2)^2}\right]$	.003	.004	.018
Total variance†	.291	.297	.329

* The random assignment economy is restricted to have the proportion of people in each of the three sectors predicted by our model using 1980 equilibrium values.

† Total variance = within-variance + between-variance

$$= \left(\frac{P_1\sigma_1 + P_2\sigma_2}{P_1 + P_2}\right) + \left[\frac{P_1P_2(M_1 - M_2)^2}{(P_1 + P_2)^2}\right].$$

original Roy model. While this definition of a random assignment of workers is arbitrary, so is any other candidate definition. The virtue of the definition we select is that it takes as its point of departure the configuration of data actually observed in 1980.[31] Note that assumption *b* is not strictly required for computing within-sector variances since task prices do not affect within-sector variances of log wage rates.

In the random assignment economy, the difference between the sectoral means of log wages is much greater than it is in the self-selection economy (0.317 vs. 0.145). Self-selection *decreases* the variance of log wages within each sector over the case of random selection (by 8.3 percent in nonmanufacturing and 9.1 percent in manufacturing). Recall that a reduction in sectoral variances is predicted by the Roy model but is not imposed on the data by our more general model. It is interesting that this qualitative feature of the Roy model is consistent with the data.

For the economy as a whole, self-selection *reduces* inequality. Within-sector inequality (summed over both sectors) declines by 7.4 percent. Because of the dramatic compression in the means of sectoral log wages, self-selection reduces the between-sector variance by 83 percent. Overall, self-selection reduces inequality (the variance in log wages) by 11.5 percent (from 0.329 to 0.291).

IV. Summary

This paper derives and estimates an empirical equilibrium model of self-selection in the labor market that recognizes the existence of measured and unmeasured heterogeneous skills within even narrowly defined demographic groups. We derive a model of the sectoral allocation of workers of different demographic types. We present a new econometric procedure that combines micro and macro data to estimate supply and demand functions for unmeasured productive attributes. Our estimated demand equations are downward-sloping functions of task prices.

Our methodology extends previous statistical work on self-selection to an explicit market setting in which the prices of attributes respond to changes in the determinants of aggregate demand and supply. Our

[31] If we had estimated the labor supply functions for all demographic groups it would be possible to compute equilibrium prices for tasks given a particular allocation of workers across sectors. However, since we estimate the supply function for only one demographic group, this procedure is not available to us.

model extends previous empirical work on wage equations by introducing determinants of aggregate market demand and supply into an explicit, economically interpretable estimating equation. We extend Roy's model of self-selection by embedding it in a market setting and by (*a*) introducing a nonmarket sector, (*b*) allowing workers to select their sector of employment on the basis of utility maximization rather than income maximization, and (*c*) permitting unmeasured attributes to be nonlognormally distributed. These extensions are required to produce a model that fits data on wage distributions from the U.S. labor market.

This study presents empirical evidence that justifies the commonly utilized practice of aggregating manufacturing into a single sector for the purpose of estimating labor demand functions. However, a new aggregate is required that recognizes both measured and unmeasured heterogeneity in skills in the population and that accounts for self-selection decisions by agents.

We use our model to estimate the importance of aggregation bias in measured aggregate real wage rates. Aggregation bias reduces measured wage variability in manufacturing below what it would be if the quality of the manufacturing work force were held constant. However, for the economy as a whole, precisely the opposite effect occurs. Aggregation bias causes measured aggregate wage variability to overstate quality constant wage variability. Because of comparative advantage, workers who move from one sector to another in response to a macro disturbance lower the average quality of the work force in the sector to which they go and raise the average quality in the sector from which they depart. This phenomenon accentuates measured wage variability over what it would be if sectoral labor force quality were held constant.

We also use our model to assess the contribution of self-selection (or the pursuit of comparative advantage) to inequality in log wage rates. We find that self-selection reduces aggregate wage inequality by more than 10 percent.

Appendix A

The Box-Cox Transformed Truncated Normal Model

The joint density of (t_1, t_2) is derived from equations (12) and (13) assuming that (u_1^*, u_2^*) are joint normal random variables. Define $(u_1^*, u_2^*) \sim N(\mathbf{0}, \Sigma_{\mathbf{u}^*})$. Let $\Phi(a_1, a_2; \rho)$ be the cumulative standardized bivariate normal with correlation coefficient ρ, where a_1 and a_2 are upper limits of integration.

The joint density of (t_1, t_2) given $\mathbf{x}$ is

$$f(t_1, t_2) =$$

$$\frac{t_1^{\lambda_1 - 1} t_2^{\lambda_2 - 1} |\Sigma u^*|^{-1/2}}{2\pi} \exp\left[-\tfrac{1}{2}\left(\frac{t_1^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1 \mathbf{x}, \frac{t_2^{\lambda_2} - 1}{\lambda_2} - \boldsymbol{\beta}_2 \mathbf{x} \right) \cdot \Sigma_{\mathbf{u}^*}^{-1} \left(\frac{t_1^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1 \mathbf{x}, \frac{t_2^{\lambda_2} - 1}{\lambda_2} - \boldsymbol{\beta}_2 \mathbf{x} \right)' \right] \Bigg/ \Phi\left[(\operatorname{sgn} \lambda_1) \frac{\boldsymbol{\beta}_1 \mathbf{x} + \frac{1}{\lambda_1}}{(\sigma_{11})^{1/2}}, (\operatorname{sgn} \lambda_2) \frac{\boldsymbol{\beta}_2 \mathbf{x} + \frac{1}{\lambda_2}}{(\sigma_{22})^{1/2}}; (\operatorname{sgn} \lambda_1)(\operatorname{sgn} \lambda_2) \rho_{12} \right], \quad \text{(A1)}$$

where ρ_{12} is the correlation coefficient between u_1^* and u_2^*.

The marginal density of t_1 given $\mathbf{x}$ is

$$f(t_1) = \frac{1}{\sqrt{2\pi\sigma_{11}}} \frac{t_1^{\lambda_1 - 1} \exp\left[-\tfrac{1}{2}\left(\frac{t_1^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1 \mathbf{x} \right)^2 \Big/ \sigma_{11} \right]}{\Phi\left[(\operatorname{sgn} \lambda_1) \frac{\boldsymbol{\beta}_1 \mathbf{x} + \frac{1}{\lambda_1}}{(\sigma_{11})^{1/2}} \right]}. \quad \text{(A2)}$$

For $\lambda_1 = 0$, this density specializes to the lognormal (we adopt the convention throughout these apps. that $\operatorname{sgn} \lambda_1 = 0$ when $\lambda_1 = 0$). For further discussion see Heckman and Sedlacek (1986).

Appendix B

The Generalized Roy Model

This appendix presents the generalized Box-Cox model. We also write out the likelihood function for the model. Let $d_i = 1$, $i = 1, \ldots, 3$, if the appropriate inequality in (17) is satisfied and zero otherwise.

Define

$$\rho_1 = \frac{\operatorname{cov}(\upsilon_1 - \upsilon_2, \upsilon_1)}{[\operatorname{var}(\upsilon_1)\operatorname{var}(\upsilon_1 - \upsilon_2)]^{1/2}}, \qquad \rho_2 = \frac{\operatorname{cov}(\upsilon_2 - \upsilon_1, \upsilon_2)}{[\operatorname{var}(\upsilon_2)\operatorname{var}(\upsilon_1 - \upsilon_2)]^{1/2}},$$

$$\rho_3 = \frac{\operatorname{cov}(\upsilon_1, \upsilon_2)}{[\operatorname{var}(\upsilon_1)\operatorname{var}(\upsilon_2)]^{1/2}}.$$

In the notation for the bivariate probit introduced in Appendix A,

$$\operatorname{pr}(d_1 = 1 | \mathbf{f}) = \Phi\left\{ \frac{\boldsymbol{\gamma}_1 \mathbf{f} - \boldsymbol{\gamma}_2 \mathbf{f}}{[\operatorname{var}(\upsilon_1 - \upsilon_2)]^{1/2}}, \frac{\boldsymbol{\gamma}_1 \mathbf{f}}{[\operatorname{var}(\upsilon_1)]^{1/2}}; \rho_1 \right\}, \quad \text{(B1}a\text{)}$$

$$\operatorname{pr}(d_2 = 1 | \mathbf{f}) = \Phi\left\{ \frac{\boldsymbol{\gamma}_2 \mathbf{f} - \boldsymbol{\gamma}_1 \mathbf{f}}{[\operatorname{var}(\upsilon_1 - \upsilon_2)]^{1/2}}, \frac{\boldsymbol{\gamma}_2 \mathbf{f}}{[\operatorname{var}(\upsilon_2)]^{1/2}}; \rho_2 \right\}, \quad \text{(B1}b\text{)}$$

$$\text{pr}(d_3 = 1|\mathbf{f}) = \Phi\left\{\frac{-\boldsymbol{\gamma}_1\mathbf{f}}{[\text{var}(\upsilon_1)]^{1/2}}, \frac{-\boldsymbol{\gamma}_2\mathbf{f}}{[\text{var}(\upsilon_2)]^{1/2}}; \rho_3\right\}. \tag{B1c}$$

Throughout we assume that $\text{var}(\upsilon_1) = 1$.

1. *The Density of Accepted Wages in the Box-Cox Model*

We derive the density of accepted wages in sector 1. The density of sector 2 accepted wages can be derived by a parallel argument.

Agents enter sector 1 provided that inequalities (17) for $i = 1$ are satisfied. These inequalities restrict the range of normal variates $(\upsilon_1 - \upsilon_2, \upsilon_1)$. The underlying u_1^* is restricted by the inequality (13) presented in the text. Let

$$\text{correl}(u_1^*, \upsilon_1 - \upsilon_2) = \rho_{12}^*, \text{correl}(u_1^*, \upsilon_1) = \rho_{11}^*, A = \frac{\boldsymbol{\gamma}_1\mathbf{f} - \boldsymbol{\gamma}_2\mathbf{f}}{[\text{var}(\upsilon_1 - \upsilon_2)]^{1/2}},$$

$$B = \frac{\boldsymbol{\gamma}_1\mathbf{f}}{[\text{var}(\upsilon_1)]^{1/2}}, R_1 = \frac{\rho_1 - \rho_{12}^*\rho_{11}^*}{\sqrt{[1 - (\rho_{11}^*)^2][1 - (\rho_{12}^*)^2]}},$$

$$C = -\frac{1}{(\sigma_{11})^{1/2}}\left(\boldsymbol{\beta}_1\mathbf{x} + \frac{1}{\lambda_1}\right).$$

In this notation, the conditional density of w_1, using $w_1 = \pi_1 t_1$ and (12), is

$$g(w_1|\boldsymbol{\gamma}_1\mathbf{f} - \boldsymbol{\gamma}_2\mathbf{f} + \upsilon_1 - \upsilon_2 > 0, \boldsymbol{\gamma}_1\mathbf{f} + \upsilon_1 > 0, \lambda_1 u_1^* > -\lambda_1\boldsymbol{\beta}_1\mathbf{x} - 1)$$

$$= \frac{\left(\frac{w_1}{\pi_1}\right)^{\lambda_1 - 1}}{\sqrt{2\pi\sigma_{11}}} \frac{1}{\pi_1} \exp\left\{-\frac{1}{2\sigma_{11}}\left[\frac{\left(\frac{w_1}{\pi_1}\right)^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1\mathbf{x}\right]^2\right\} \cdot$$

$$\Phi\left\{\frac{A + \rho_{12}^*\left[\frac{\left(\frac{w_1}{\pi_1}\right)^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1\mathbf{x}\right]\left(\frac{1}{\sigma_{11}}\right)^{1/2}}{\sqrt{1 - (\rho_{12}^*)^2}}, \right. \tag{B2}$$

$$\left. \frac{B + \rho_{11}^*\left[\frac{\left(\frac{w_1}{\pi_1}\right)^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1\mathbf{x}\right]\left(\frac{1}{\sigma_{11}}\right)^{1/2}}{\sqrt{1 - (\rho_{11}^*)^2}}; R_1 \right\}$$

$$\Phi[-(\text{sgn }\lambda_1)C, A, B; (\text{sgn }\lambda_1)\rho_{12}^*, (\text{sgn }\lambda_1)\rho_{11}^*, \rho_1]$$

where $\Phi(a, b, c; d, e, f)$ is a trivariate normal integral with upper limit a, b, c and correlation structure d, e, f. Letting

$$\text{correl}(u_2^*, \upsilon_2 - \upsilon_1) = \rho_{21}^*, \text{correl}(u_2^*, \upsilon_2) = \rho_{22}^*, \bar{A} = \frac{\boldsymbol{\gamma}_2\mathbf{f} - \boldsymbol{\gamma}_1\mathbf{f}}{[\text{var}(\upsilon_2 - \upsilon_1)]^{1/2}},$$

$$\bar{B} = \frac{\boldsymbol{\gamma}_2\mathbf{f}}{[\text{var}(\upsilon_2)]^{1/2}}, \; R_2 = \frac{\rho_2 - \rho_{21}^*\rho_{22}^*}{\sqrt{[1 - (\rho_{22}^*)^2][1 - (\rho_{21}^*)^2]}},$$

$$\bar{C} = \frac{-1}{(\sigma_{22})^{1/2}}\left(\boldsymbol{\beta}_2\mathbf{x} + \frac{1}{\lambda_2}\right),$$

the density of accepted wages in sector 2 is

$$g(w_2|\boldsymbol{\gamma}_2\mathbf{f} - \boldsymbol{\gamma}_1\mathbf{f} + \upsilon_2 - \upsilon_1 > 0, \boldsymbol{\gamma}_2\mathbf{f} + \upsilon_2 > 0, \lambda_2 u_2^* > -\lambda_2\boldsymbol{\beta}_2\mathbf{x} - 1)$$

$$= \frac{\left(\frac{w_2}{\pi_2}\right)^{\lambda_2 - 1}}{\sqrt{2\pi\sigma_{22}}} \frac{1}{\pi_2} \exp\left\{ - \frac{1}{2\sigma_{22}} \left[\frac{\left(\frac{w_2}{\pi_2}\right)^{\lambda_2} - 1}{\lambda_2} - \boldsymbol{\beta}_2\mathbf{x} \right]^2 \right\} \cdot$$

$$\Phi\left\{ \frac{\bar{A} + \rho_{21}^* \left[\frac{\left(\frac{w_2}{\pi_2}\right)^{\lambda_2} - 1}{\lambda_2} - \boldsymbol{\beta}_2\mathbf{x} \right] \left(\frac{1}{\sigma_{22}} \right)^{1/2}}{\sqrt{1 - (\rho_{21}^*)^2}}, \right.$$

$$\left. \frac{\bar{B} + \rho_{22}^* \left[\frac{\left(\frac{w_2}{\pi_2}\right)^{\lambda_2} - 1}{\lambda_2} - \boldsymbol{\beta}_2\mathbf{x} \right] \left(\frac{1}{\sigma_{22}} \right)^{1/2}}{\sqrt{1 - (\rho_{22}^*)^2}}; R_2 \right\} \Big/ \Phi[-(\text{sgn } \lambda_2)\bar{C}, \bar{A}, \bar{B}; (\text{sgn } \lambda_2)\rho_{21}^*, (\text{sgn } \lambda_2)\rho_{22}^*, \rho_2) \tag{B3}$$

2. *Accepted Wage Distributions When Wages Are above a Threshold*

We modify the densities in Section 1 to account for a sampling rule that $w_i > \tau$, where τ is a minimum threshold value. We derive the sector 1 accepted wage distribution. The derivation of the sector 2 accepted wage distribution follows by a parallel argument.

The requirement $w_1 > \tau > 0$ translates into the restriction

$$w_1 = \pi_1(\lambda_1\boldsymbol{\beta}_1\mathbf{x} + \lambda_1 u_1^* + 1)^{1/\lambda_1} > \tau$$

or

$$\frac{u_1^*}{(\sigma_{11})^{1/2}} > \left[-\left(\frac{1}{\lambda_1} + \boldsymbol{\beta}_1\mathbf{x}\right) + \frac{1}{\lambda_1}\left(\frac{\tau}{\pi_1}\right)^{\lambda_1} \right] \frac{1}{(\sigma_{11})^{1/2}} \tag{B4}$$

for all values of λ_1 not equal to zero. Combining restriction (13) with (B4) and assuming that $\lambda_1 < 0$ implies that

$$\left[-\left(\frac{1}{\lambda_1} + \boldsymbol{\beta}_1\mathbf{x}\right) + \frac{1}{\lambda_1}\left(\frac{\tau}{\pi_1}\right)^{\lambda_1} \right] \frac{1}{(\sigma_{11})^{1/2}} < \frac{u_1^*}{(\sigma_{11})^{1/2}} < -\left(\frac{1}{\lambda_1} + \boldsymbol{\beta}_1\mathbf{x}\right) \frac{1}{(\sigma_{11})^{1/2}}.$$

Notice that as $\lambda_1 \to 0$ from below this inequality becomes

$$\left[-\boldsymbol{\beta}_1\mathbf{x} + \ln\left(\frac{\tau}{\pi_1}\right)\right]\left(\frac{1}{\sigma_{11}}\right)^{1/2} < \frac{u_1^*}{(\sigma_{11})^{1/2}}.$$

When $\lambda_1 > 0$, (B4) will be the appropriate inequality; that is, (13) imposes no extra restrictions on the range of u_1^* beyond the one already imposed by (B4).

Letting

$$D = \frac{1}{\lambda_1}\left(\frac{\tau}{\pi_1}\right)^{\lambda_1}\frac{1}{(\sigma_{11})^{1/2}},$$

the density of accepted wages in sector 1 given $w_1 > \tau$ and $\lambda_1 > 0$ is

$$g(w_1|\boldsymbol{\gamma}_1\mathbf{f} - \boldsymbol{\gamma}_2\mathbf{f} + \upsilon_1 - \upsilon_2 > 0, \boldsymbol{\gamma}_1\mathbf{f} + \upsilon_1 > 0, w_1 > \tau, \lambda_1 > 0)$$

$$= \frac{\left(\frac{w_1}{\pi_1}\right)^{\lambda_1 - 1}}{\sqrt{2\pi\sigma_{11}}}\frac{1}{\pi_1}\exp\left\{-\frac{1}{2\sigma_{11}}\left[\frac{\left(\frac{w_1}{\pi_1}\right)^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1\mathbf{x}\right]^2\right\} \cdot$$

$$\Phi\left\{\frac{A + \rho_{12}^*\left[\frac{\left(\frac{w_1}{\pi_1}\right)^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1\mathbf{x}\right]\left(\frac{1}{\sigma_{11}}\right)^{1/2}}{\sqrt{1 - (\rho_{12}^*)^2}}, \right. \tag{B5}$$

$$\left. \frac{B + \rho_{11}^*\left[\frac{\left(\frac{w_1}{\pi_1}\right)^{\lambda_1} - 1}{\lambda_1} - \boldsymbol{\beta}_1\mathbf{x}\right]\left(\frac{1}{\sigma_{11}}\right)^{1/2}}{\sqrt{1 - (\rho_{11}^*)^2}}; R_1\right\}$$

$$\Phi[-(C + D), A, B; \rho_{12}^*, \rho_{11}^*, \rho_1].$$

The density of accepted wages in sector 1 given $w_1 > \tau$ and $\lambda_1 < 0$ is (B5) multiplied by

$$\frac{\Phi[-(C + D), A, B; \rho_{12}^*, \rho_{11}^*, \rho_1]}{\Phi(C, A, B; -\rho_{12}^*, -\rho_{11}^*, \rho_1) - \Phi(C + D, A, B; -\rho_{12}^*, -\rho_{11}^*, \rho_1)}.$$

Recall that $D < 0$ if $\lambda_1 < 0$. The density of $g(w_2|\boldsymbol{\gamma}_2\mathbf{f} - \boldsymbol{\gamma}_1\mathbf{f} + \upsilon_2 - \upsilon_1 > 0, \boldsymbol{\gamma}_2\mathbf{f} + \upsilon_2 > 0, w_2 > \tau, \lambda_2 < 0)$ is derived by a parallel argument.

3. *The Likelihood Function for Our Model for a Sample Consisting of All Nonworkers plus Workers with Wage Rates above a Threshold*

In this section we utilize results derived in Sections 1 and 2 to write out the likelihood function used to estimate our model. An individual is in our sample if he chooses to go to sector 1 and his wages are above τ, if he chooses to go to sector 2 and his wages are above τ, or if he chooses to go to sector 3. Denote

by CP the probability that an individual with characteristics $\mathbf{f}$ satisfies one of these three criteria. Then

$$CP = \text{pr}(d_1 = 1, w_1 > \tau|\mathbf{f}) + \text{pr}(d_2 = 1, w_2 > \tau|\mathbf{f}) + \text{pr}(d_3 = 1|\mathbf{f}). \quad \text{(B6)}$$

The probability that an agent chooses sector 1 and has a wage above threshold τ given that $\lambda_1 > 0$ is

$$\text{pr}(d_1 = 1, w_1 > \tau|\mathbf{f}, \lambda_1 > 0) = \Phi(A, B, \rho_1) \cdot \frac{\Phi[-(C + D), A, B; \rho^*_{12}, \rho^*_{11}, \rho_1]}{\Phi[-C, A, B; \rho^*_{12}, \rho^*_{11}, \rho_1]}. \quad \text{(B7)}$$

For $\lambda_1 < 0$, the desired probability is

$$\text{pr}(d_1 = 1, w_1 > \tau|\mathbf{f}, \lambda_1 < 0) = \Phi(A, B, \rho_1) \cdot \frac{\Phi(C, A, B; -\rho^*_{12}, -\rho^*_{11}, \rho_1) - \Phi(C + D, A, B, -\rho^*_{12}, -\rho^*_{11}, \rho_1)}{\Phi(C, A, B; -\rho^*_{12}, -\rho^*_{11}, \rho_1)}. \quad \text{(B7)}'$$

For sector 2

$$\text{pr}(d_2 = 1, w_2 > \tau|\mathbf{f}, \lambda_2 > 0) = \Phi(\bar{A}, \bar{B}, \rho_2) \cdot \frac{\Phi[-(\bar{C} + \bar{D}), \bar{A}, \bar{B}; \rho^*_{21}, \rho^*_{22}, \rho_2]}{\Phi(-\bar{C}, \bar{A}, \bar{B}; \rho^*_{21}, \rho^*_{22}, \rho_2)} \quad \text{(B8)}$$

and

$$\text{pr}(d_2 = 1, w_2 > \tau|\mathbf{f}, \lambda_2 < 0) = \Phi(\bar{A}, \bar{B}, \rho_2) \cdot \frac{\Phi(\bar{C}, \bar{A}, \bar{B}; -\rho^*_{21}, -\rho^*_{22}, \rho_2) - \Phi(\bar{C} + \bar{D}, \bar{A}, \bar{B}; -\rho^*_{21}, -\rho^*_{22}, \rho_2)}{\Phi(\bar{C}, \bar{A}, \bar{B}; -\rho^*_{21}, -\rho^*_{22}, \rho_2)}. \quad \text{(B8)}'$$

The probability that an individual will be observed in sector 1 given that he is in the sample, defined as $P1$, is

$$P1 = \frac{\text{pr}(d_1 = 1, w_1 > \tau|\mathbf{f})}{\text{pr}(d_1 = 1, w_1 > \tau|\mathbf{f}) + \text{pr}(d_2 = 1, w_2 > \tau|\mathbf{f}) + \text{pr}(d_3 = 1|\mathbf{f})}. \quad \text{(B9)}$$

The contribution to likelihood function L of an individual with characteristics $\mathbf{f}$ observed in sector 1 is

$$L_1 = g(w_1|\mathbf{f}, d_1 = 1, w_1 > \tau) \cdot P1, \quad \text{(B10)}$$

where $g(w_1|\mathbf{f}, d_1 = 1, w_1 > \tau)$ is defined in (B5).

By a parallel argument the probability that an individual will be observed in sector 2 given that he is in the sample, defined as $P2$, is

$$P2 = \frac{\text{pr}(d_2 = 1, w_2 > \tau|\mathbf{f})}{\text{pr}(d_1 = 1, w_1 > \tau|\mathbf{f}) + \text{pr}(d_2 = 1, w_2 > \tau|\mathbf{f}) + \text{pr}(d_3 = 1|\mathbf{f})}. \quad \text{(B11)}$$

The contribution to likelihood function L of an individual with characteristics $\mathbf{f}$ observed in sector 2 is

$$L_2 = g(w_2|\mathbf{f}, d_2 = 1, w_2 > \tau) \cdot P2, \quad \text{(B12)}$$

where $g(w_2|\mathbf{f}, d_2 = 1, w_2 > \tau)$ is defined analogously to (B5) and $P2$ is defined in (B11).

The probability that an agent with characteristics $\mathbf{f}$ chooses sector 3 (the nonmarket sector) conditional on being in the sample is

$$P3 = \frac{\text{pr}(d_3 = 1|\mathbf{f})}{\text{pr}(d_1 = 1, w_1 > \tau|\mathbf{f}) + \text{pr}(d_2 = 1, w_2 > \tau|\mathbf{f}) + \text{pr}(d_3 = 1|\mathbf{f})}. \quad \text{(B13)}$$

The contribution to likelihood function L of an individual with characteristics $\mathbf{f}$ observed in sector 3 is

$$L_3 = P3, \quad \text{(B14)}$$

where $P3$ is defined in (B13).

The likelihood satisfies classical regularity conditions because it is twice continuously differentiable in the parameters of the model, and the range of each random variable does not depend on the parameters of the model.

Appendix C

Description of the Data

1. *CPS Data*

The sample utilized to estimate the extended Roy model is derived from the March Current Population Survey (CPS). From the CPS data base of 1968–81 we randomly select a 4 percent subsample of civilian white males between the ages of 18 and 65. In constructing our sample we eliminate any observation with imputed data for any of the variables utilized in the analysis.

The following variables are extracted from the CPS data file: annual labor income last year, hours worked last week, number of weeks in the labor force last year, total income last year, years of schooling, age of the person, three-digit industry code of last year's job, and current state of residence. Total income and labor income variables are transformed into real variables by dividing by the CPI; we use 1967 dollar constant values.

We construct two variables: hourly wage rate and income from nonlabor sources. The hourly wage rate is obtained by dividing the labor income the respondent obtained in the year prior to the interview by the product of the number of weeks he was in the labor force in that year and the number of hours he worked in the week prior to the interview. The income from nonlabor sources is obtained by subtracting the labor income from the total income, both defined for the year prior to the interview. The sectoral nonlabor income obtained is then regressed on the following exogenous variables: age, education, state of residence, and polynomials of these variables. The predicted value from this regression is then utilized as a regressor to avoid spurious correlation between assets and the unobservables in the choice equations.

As noted in the text, we exclude all individuals whose real hourly wages are below \$0.75. The lower tail of the hourly wage distribution is excluded to minimize the effects of measurement error.

2. *Industry Data*

The following data series are utilized to estimate the industry task demand functions: industrial commodity price index, farm products price index, in-

termediate goods price index, energy price index, nonresidential fixed investment price deflator, corporate bond (Moody's) Aaa yields, total population by demographic group, average hours worked in each sector, and number of workers and hourly wages in the manufacturing and nonmanufacturing sectors. These series are obtained from the *Statistical Abstracts of the United States* (1968–81) and *Historical Statistics of the United States—Colonial Times to 1970*, both published by the U.S. Department of Commerce, Bureau of the Census.

The information required to estimate the industry demand equation described in the text is data on output price indices for both the manufacturing and nonmanufacturing sectors. The industrial commodities price index and the farm output price index are used as output price indices for the manufacturing and nonmanufacturing sectors, respectively.

The total wage bills in the manufacturing and nonmanufacturing sectors are obtained as the product of total number of employees times the average hourly wage they receive times the average number of hours worked in each sector.

3. *Data Sets for the Simulations Reported in Sections IIIC and IIID*

The simulation results reported in the text require the empirical distribution of exogenous characteristics in the population. These distributions are obtained from a 20 percent random sample derived from the CPS data file for the period 1968–81. The variables selected are age, years of schooling, and state of residence. Individuals with missing data for any of these three variables are excluded from the sample. The estimate of income from nonlabor sources is obtained by regressing nonlabor income in the population on the exogenous variables described above and polynomials of those variables.

4. *Definitions of the Variables Utilized in the Analysis*

Hourly wage rate = total labor income/(weeks × hours).
Weeks = weeks in the labor force in the previous year.
Hours = hours worked in the week prior to the interview.
Nonlabor income = total income − total labor income.
Total income = total income of the respondent in the previous year.
Total labor income = wage and salary income + nonfarm self-employment income + farm self-employment income (all in the previous year).
Education = years of schooling.
Experience = age − education − 6.
South = 1 if the respondent was living in the U.S. Census South at the time of the interview, 0 otherwise.
Sector choice = 1, working in a nonmanufacturing industry;
= 2, working in the manufacturing sector, last year three-digit industry code falls between 107 and 398;
= 3, not working, has zero total labor income in the previous year.
Energy price index = producer price index for energy.
Intermediate goods price = intermediate goods price index.
User cost of capital = nonresidential fixed investment price deflator times the corporate bond (Moody's) Aaa yields.

TABLE C1

DESCRIPTIVE STATISTICS FOR THE SAMPLE DATA IN 1976 AND 1980

	MANUFACTURING		NONMANUFACTURING		NOT WORKING	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Year: 1976 (sample size: 1,407):						
Age	38.778	12.616	36.967	13.358	29.018	18.661
Education	11.637	3.087	12.438	3.165	9.462	3.003
Nonlabor income	499.424	431.407	513.860	450.698	451.207	521.874
South	.225	...	.305	...	.297	...
Sectoral choice	.224	...	.584	...	.192	...
Hours worked last week	42.040	9.345	41.316	14.921	...	...
Weeks in the labor force	48.459	8.842	47.081	10.895	...	...
Real hourly wage rate	3.661	2.350	3.746	4.032	...	...
Year: 1980 (sample size: 1,855):						
Age	38.121	12.891	36.434	13.105	33.866	19.781
Education	12.304	2.792	12.940	3.029	10.270	3.138
Nonlabor income	449.724	501.727	460.821	500.187	632.843	609.917
South	.249	...	.366	...	.336	...
Sectoral choice	.206	...	.630	...	.164	...
Hours worked last week	42.673	9.677	41.739	14.416	...	...
Weeks in the labor force	48.935	8.346	47.433	10.410	...	...
Real hourly wage rate	3.326	2.104	3.325	3.899	...	...

NOTE.—These samples exclude workers with hourly wages less than $0.75 per hour. Standard deviations of dummy variables are not reported.

Appendix D

TABLE D1
ESTIMATED YEAR EFFECTS FOR REDUCED-FORM CHOICE EQUATIONS AND TASK FUNCTIONS; ESTIMATED TASK PRICES

	YEAR EFFECTS IN CHOICE EQUATIONS		YEAR EFFECTS IN TASK FUNCTIONS		ESTIMATED TASK PRICES	
YEAR (l)	Manufacturing (γ_{02l})	Nonmanufacturing (γ_{01l})	Manufacturing (β_{02l})	Nonmanufacturing (β_{01l})	Manufacturing ($\widehat{\ln \pi}_{2l}$)	Nonmanufacturing ($\widehat{\ln \pi}_{1l}$)
1968	.0218225	−.0220290	−.646926	1.227792	.577112	−1.617393
1969	−.0756869	−.1263055	−1.200297	.695159	1.153830	−.913847
1970	−.1034628	−.1500157	−.866472	.474126	.847478	−.660620
1971	−.0476136	.0496184	−1.197267	1.524989	1.181526	−1.733567
1972	−.0259010	−.0099231	1.169631	−.117819	−1.034553	.088555
1973	−.0118258	.0184051	−1.216458	.537582	1.147806	−.611408
1974	−.0586111	.0170755	−.807416	.158533	.735355	−.234924
1975	−.0295666	−.0008927	.129516	1.457207	−.114545	−1.653301
1976	.0000000	.0000000	.000000	.000000	.000000	.000000
1977	−.1478383	−.2038792	.621883	1.235021	−.584197	−1.436771
1978	−.1989508	−.1893151	−.781377	.200122	.634622	−.265950
1979	.1150525	.2075400	−1.414388	−.503048	1.250973	.438861
1980	.0177292	.1131955	.038270	−.312876	−.225510	.216559
1981	.1799860	.2688506	−.916524	−.062969	.730783	−.060345

NOTE.—1976 is the benchmark year.

Appendix E

Notes on the Computation of the Aggregation Bias Simulations Reported in Table 3

The simulations reported in Section IIIC in the text are performed using the parameters estimated for the extended Roy model (reported in table 1) with task prices and the intercepts of the utility functions adjusted to take into account the energy price increase.

The impact of an energy price increase on the supply of tasks is assumed to operate only through its effect on the task prices. Estimates of the intercepts of the reduced-form utility functions include the effect of task prices on the sector-specific utility of agents.

To compute the response of task prices to changes in the energy price we regress estimated log task prices on log energy prices and other determinants of the equilibrium market price. The estimated price equations (with standard errors in parentheses below the coefficients) are:

$$\widehat{\ln \pi_{1l}} = \underset{(.350)}{-.0736} + \underset{(1.145)}{.471} \cdot (\text{log energy price index}) - \underset{(7.232)}{2.021}$$

$$\cdot (\text{log intermediate goods price}) + \underset{(.811)}{.8349} \cdot (\text{log user cost of capital}),$$

$$R^2 = .2810, \text{ D-W} = 3.046;$$

$$\widehat{\ln \pi_{2l}} = \underset{(.350)}{.18611} - \underset{(1.09)}{1.4800} \cdot (\text{log energy price index}) - \underset{(6.912)}{2.934}$$

$$\cdot (\text{log intermediate goods price}) + \underset{(.795)}{1.6894} \cdot (\text{log user cost of capital}),$$

$$R^2 = .3939, \text{ D-W} = 2.91.$$

(The variables are defined in App. C.) Essentially the same empirical results are obtained if time trends are included in the regression.

In order to estimate the effects of task price changes on sectoral choices it is necessary to decompose the estimated year effects in the utility functions (the γ_{0il}) into two components: the contribution of log task price and the contribution of unobserved supply characteristics. We approximate the latter by time-trended variables: a time trend and the unemployment rate in the United States. The regression of the estimated intercepts on the estimated log task prices, a time trend, and the unemployment rate (standard errors in parentheses) are:

$$\hat{\gamma}_{01l} = \underset{(2.14)}{.543} + \underset{(.153)}{.154} \cdot (\widehat{\ln \pi_{1l}}) - \underset{(.0256)}{.0051} \cdot (\text{time trend})$$

$$+ \underset{(.0649)}{.0515} \cdot (\text{unemployment rate}),$$

$$R^2 = .3280, \text{ D-W} = 2.46;$$

$$\hat{\gamma}_{02l} = \underset{(.732)}{-.712} + \underset{(.0703)}{.0502} \cdot (\widehat{\ln \pi_{2l}}) + \underset{(.009)}{.009} \cdot (\text{time trend})$$

$$+ \underset{(.0384)}{.0204} \cdot (\text{unemployment rate}),$$

$$R^2 = .2852, \text{ D-W} = 1.66.$$

The estimated coefficients imply that a 1 percent increase in the price of energy decreases the manufacturing sector task price by 1.48 percent and increases the nonmanufacturing sector task price by 0.47 percent, and that the intercepts in the utility function will shift by 0.072 in the nonmanufacturing sector ($= 0.471 \times 0.154$) and by -0.074 in the manufacturing sector ($= -1.48 \times 0.0502$).

References

Amemiya, Takeshi, and Powell, James L. "A Comparison of the Box-Cox Maximum Likelihood Estimator and the Non-linear Two-Stage Least Squares Estimator." *J. Econometrics* 17 (December 1981): 351–81.

Bils, Mark J. "Real Wages over the Business Cycle: Evidence from Panel Data." *J.P.E.* 93 (August 1985): 666–89.

Bock, R. Darrell, and Jones, Lyle V. *The Measurement and Prediction of Judgment and Choice*. San Francisco: Holden-Day, 1968.

Davies, R. B. "Hypothesis Testing When a Nuisance Parameter Is Present Only under an Alternative." *Biometrika* 64 (August 1977): 257–54.

Domencich, Thomas A., and McFadden, Daniel. *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North-Holland, 1975.

Durbin, James. "Errors in Variables." *Rev. Internat. Statis. Inst.* 22 (1954): 23–32.

Geary, Patrick T., and Kennan, John. "The Employment–Real Wage Relationship: An International Study." *J.P.E.* 90 (August 1982): 854–71.

Goldberger, Arthur S. "Abnormal Selection Bias." In *Studies in Econometrics, Time Series, and Multivariate Statistics*, edited by Samuel Karlin, Takeshi Amemiya, and Leo A. Goodman. New York: Academic Press, 1983.

Gollop, Frank M., and Jorgenson, Dale W. "Sectoral Measures of Labor Cost for the United States, 1948–1978." In *The Measurement of Labor Cost*, edited by Jack E. Triplett. Chicago: Univ. Chicago Press (for N.B.E.R.), 1983.

Gorman, W. M. "A Possible Procedure for Analysing Quality Differentials in the Egg Market." *Rev. Econ. Studies* 47 (October 1980): 843–56.

Hamermesh, Daniel S., and Grant, James H. "Econometric Studies of Labor–Labor Substitution and Their Implications for Policy." *J. Human Resources* 14 (Fall 1979): 518–42.

Heckman, James J. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Ann. Econ. and Soc. Measurement* 5 (Fall 1976): 475–92.

———. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (January 1979): 153–61.

———. "Addendum to Sample Selection Bias as a Specification Error." In *Evaluation Studies Review*, vol. 5, edited by E. Stromsdorfer and G. Farkas. New York: Sage, 1980.

Heckman, James J., and MaCurdy, Thomas E. "Labor Econometrics." In *Handbook of Econometrics*, vol. 3, edited by Zvi Griliches and Michael Intriligator. Amsterdam: North-Holland, 1985.

Heckman, James J., and Polachek, Solomon. "Empirical Evidence on the Functional Form of the Earnings-Schooling Relationship." *J. American Statis. Assoc.* 69 (June 1974): 350–54.

Heckman, James J., and Scheinkman, José A. "The Importance of Bundling in a Gorman-Lancaster Model of Earnings." Mimeographed. Chicago: Univ. Chicago, 1982; rev. June 1984; presented at world meetings of the Econometric Society, Boston, 1985.

Heckman, James J., and Sedlacek, Guilherme. "The Impact of the Minimum Wage on the Employment and Earnings of Workers in South Carolina." In *Report on Minimum Wage Study Commission*, vol. 4. Washington: U.S. Government Printing Office, 1981.

———. "Econometric Models of Self-Selection and Income Distribution." In *Innovations in Quantitative Economics: Essays in Honor of Robert L. Basmann*, edited by Daniel J. Slottje. 1986 (in press).

Heckman, James J., and Singer, Burton. "The Identifiability of a Nonparametric Roy Model." Manuscript. Chicago: Univ. Chicago, 1985.

Jorgenson, Dale W. "Econometric Methods for Modeling Producer Behavior." In *Handbook of Econometrics*, vol. 3, edited by Zvi Griliches and Michael Intriligator. Amsterdam: North-Holland, 1985.

Jorgenson, Dale W.; Lau, Lawrence J.; and Stoker, Thomas M. "The Transcendental Logarithmic Model of Aggregate Consumer Behavior." In *Advances in Econometrics*, vol. 1, edited by Robert O. Basmann and G. Rhodes. Greenwich, Conn.: JAI, 1982.

Lancaster, Kelvin J. "A New Approach to Consumer Theory." *J.P.E.* 74 (April 1966): 132–57.

Lee, Lung-fei. "Unionism and Wage Rates: A Simultaneous Model with Qualitative and Limited Dependent Variables." *Internat. Econ. Rev.* 19 (June 1978): 415–33.

Lillard, Lee; Smith, James P.; and Welch, Finis. "What Do We Really Know about Wages: The Importance of Nonreporting and Census Imputation." Manuscript. Santa Monica, Calif.: Rand, 1982.

Lydall, Harold F. *The Structure of Earnings*. Oxford: Oxford Univ. Press, 1968.

Mandelbrot, Benoit. "Paretian Distributions and Income Maximization." *Q.J.E.* 76 (February 1962): 57–85.

Moore, David S. "Generalized Inverses, Wald's Method, and the Construction of Chi-squared Tests of Fit." *J. American Statis. Assoc.* 72 (March 1977): 131–37.

Mortensen, Dale T. "Unemployment Insurance and Job Search Decisions." *Indus. and Labor Relations Rev.* 30 (July 1977): 505–17.

Poirier, Dale J. "The Use of the Box-Cox Transformation in Limited Dependent Variable Models." *J. American Statis. Assoc.* 73 (June 1978): 284–87.

Rosen, Sherwin. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *J.P.E.* 82 (January/February 1974): 34–55.

———. "Substitution and Division of Labour." *Economica* 45 (August 1978): 235–50.

Roy, Andrew D. "Some Thoughts on the Distribution of Earnings." *Oxford Econ. Papers* 3 (June 1951): 135–46.

Sargent, Thomas J. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86 (December 1978): 1009–44.

Sattinger, Michael. *Capital and the Distribution of Labor Earnings*. Amsterdam: North-Holland, 1980.

Stockman, Alan C. "Aggregation Bias and the Cyclical Behavior of Real Wages." Manuscript. Rochester, N.Y.: Univ. Rochester, 1983.

Tinbergen, Jan. "Some Remarks on the Distribution of Labour Incomes." *Internat. Econ. Papers* 1 (1951): 195–207.

———. "On the Theory of Income Distribution." *Weltwirtschaftliches Archiv* 77, no. 2 (1956): 155–73.

Welch, Finis. "Linear Synthesis of Skill Distribution." *J. Human Resources* 4 (Summer 1969): 311–27.

Willis, Robert J., and Rosen, Sherwin. "Education and Self-Selection." *J.P.E.* 87, no. 5, pt. 2 (October 1979): S7–S36.

References

A Possible Procedure for Analysing Quality Differentials in the Egg Market

W. M. Gorman

The Review of Economic Studies, Vol. 47, No. 5. (Oct., 1980), pp. 843-856.

Stable URL:

http://links.jstor.org/sici?sici=0034-6527%28198010%2947%3A5%3C843%3AAPPFAQ%3E2.0.CO%3B2-L

Sample Selection Bias as a Specification Error

James J. Heckman

Econometrica, Vol. 47, No. 1. (Jan., 1979), pp. 153-161.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28197901%2947%3A1%3C153%3ASSBAAS%3E2.0.CO%3B2-J

Substitution and Division of Labour

Sherwin Rosen

Economica, New Series, Vol. 45, No. 179. (Aug., 1978), pp. 235-250.

Stable URL:

http://links.jstor.org/sici?sici=0013-0427%28197808%292%3A45%3A179%3C235%3ASADOL%3E2.0.CO%3B2-6