# Estimation of an Occupational Choice Model when Occupations are Misclassified

**Paul Sullivan**

ABSTRACT

*This paper develops an empirical occupational choice model that corrects for misclassification in occupational choices and measurement error in occupation-specific work experience. The model is used to estimate the extent of measurement error in occupation data and quantify the bias that results from ignoring measurement error in occupation codes when studying the determinants of occupational choices and estimating the effects of occupation-specific human capital on wages. The parameter estimates reveal that 9 percent of occupational choices in the 1979 cohort of the NLSY are misclassified. Ignoring misclassification leads to biases that affect the conclusions drawn from empirical occupational choice models.*

## I. Introduction

Occupational choices have been the subject of considerable research interest by economists because of their importance in shaping employment outcomes and wages over the career. Topics of study range from the analysis of job search and occupational matching (McCall 1990; Neal 1999) to studies of the determinants of wage inequality (Gould 2002) to dynamic human capital models of occupational choices (Keane and Wolpin 1997). Despite the large amount of research into occupational choices and evidence from validation studies such as Mellow and Sider (1983) which suggests that as many as 20 percent of one-digit occupational choices are misclassified, it is surprising that existing research has not corrected for

*Paul Sullivan is a research economist at the U.S. Bureau of Labor Statistics.* *The author thanks John Bound, two anonymous referees, and seminar participants at Virginia, the BLS, and the Annual Meeting of the Society of Labor Economists for comments that greatly improved this paper. Financial support from a National Institute on Aging Post-Doctoral Fellowship (grant number AG00221-14) at the University of Michigan is gratefully acknowledged. The views and opinions expressed in this paper are those of the author and do not necessarily reflect the views of the Bureau of Labor Statistics. The data used in this article can be obtained beginning October 2009 through September 2012 from Paul J. Sullivan, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Suite 3105, Washington, D.C. 20212-0001 <Sullivan.Paul.Joseph@bls.gov>.*
[Submitted December 2006; accepted November 2007]

classification error in occupations when estimating models of occupational choice. The existence of classification error in occupations is a serious concern because in the context of a nonlinear discrete choice occupational choice model, measurement error in the dependant variable results in biased parameter estimates.

The goal of this paper is to estimate a model of occupational choices that corrects for classification error in occupation data when direct evidence on the validity of individuals' self-reported occupations is unavailable. The approach taken in this paper is to specify a model of occupational choices that incorporates a parametric model of occupational misclassification. The parameters of the occupational choice model and the parameters that describe the extent of misclassification in occupation data are estimated jointly by simulated maximum likelihood. As is the case in all structural models, a limitation of this approach is that it requires the researcher to make parametric assumptions about objects in the model such as the functional form of the wage equation, the distribution of random variables that affect occupational choices, and the process by which occupations are misclassified.

The classification error literature consists of two broadly defined approaches to estimating parametric models in the presence of classification error.[1] One approach uses assumptions about the measurement error process along with auxiliary information on error rates, which typically takes the form of validation or reinterview data, to correct for classification error. Examples of this approach to measurement error are found in work by Abowd and Zellner (1985); Chua and Fuller (1987); Poterba and Summers (1995); Magnac and Visser (1999); and Chen, Hong, and Tamer (2005). The second approach to estimating models in the presence of misclassified data corrects for misclassification without relying on auxiliary information by estimating parametric models of misclassification. Examples of this approach are found in Hausman, Abrevaya, and Scott-Morton (1998); Dustmann and van Soest (2001); and Li, Trivedi, and Guo (2003).

The occupational choice model developed in this paper combines features of the two existing approaches to misclassification. Instead of relying on the availability of auxiliary information that provides direct evidence on misclassified occupational choices, information about misclassification is derived from observed wages. This approach takes advantage of the fact that observed wages provide information about true occupational choices because wages vary widely across occupations. Intuitively, the occupational choices identified by the model as likely to be misclassified are the ones where the observed wage is unlikely to be observed in the reported occupation. Also, the model developed in this paper uses additional information provided by the fact that true occupational choices are strongly influenced by observable variables, such as education, to draw inferences about the extent of misclassification in the data.

One methodological contribution of this work is that it develops a method of dealing with the problems created in panel data models when misclassification in the dependant variable creates measurement error in the explanatory variables in the model. Misclassification in occupation codes creates measurement error in lagged

1. An alternative approach to dealing with misclassification derives nonparametric bounds under relatively weak assumptions about misclassification. See, for example, Bollinger's (1996) study of mismeasurered binary independent variables in a linear regression, and Kreider and Pepper's (2007A, 2007B) work on misclassification in disability status.

occupational choices and occupation-specific work experience variables, so the true values of these variables are unobserved state variables. Existing research into occupational choices and misclassification in general has not addressed this problem.[2] This work addresses the problem by using simulation methods to approximate the otherwise intractable integrals over the unobserved state variables that appear in the likelihood function.[3] The simulation algorithm developed in this paper is applicable in a wide range of settings beyond occupational choice models.

The parameter estimates provide evidence that a substantial fraction of occupational choices are misclassified in the NLSY data, and suggest that ignoring misclassification leads to biases that affect the qualitative and quantitative conclusions drawn from estimated occupational choice models. Estimates of the transferability of human capital across occupations appear to be particularly sensitive to the false occupational transitions created by misclassification. The results also suggest that the extent of misclassification varies widely across occupations, and that observed wages provide a large amount of information about which occupational choices in the data are likely to be affected by misclassification. For example, the model predicts that high wage workers who are observed as professionals are very likely to be correctly classified, but low wage workers observed as professionals are likely to be misclassified.

The remainder of the paper is organized as follows. Section II describes the data and discusses the possible sources of measurement error in occupation codes. Section III presents the model of occupational choices and misclassification and discusses how the model is estimated. Section IV presents the parameter estimates, and Section V analyzes the patterns in misclassification predicted by the model using simulated occupational choice data. Section VI concludes.

## II. Data

The National Longitudinal Survey of Youth (NLSY) is a panel data set that contains detailed information about the employment and educational experiences of a nationally representative sample of young men and women who were between the ages of 14 and 22 when first interviewed in 1979. The employment data contain information about the durations of employment spells along with the wages, hours, and three-digit 1970 U.S. Census occupation codes for each job.

This analysis uses only white men ages 18 or older from the nationally representative core sample of the NLSY. The weekly labor force record found in the work

2. The only other paper to examine the connection between misclassification in the dependant variable and measurement error in explanatory variables is Keane and Sauer's (2006) study of female labor supply that examines misclassification in reported labor force status. Keane and Sauer (2006) estimate their model using the simulation procedure developed by Keane and Wolpin (2001) to deal with the problem of unobserved state variables in dynamic models.

3. This application of simulation methods adds to a growing literature that uses simulation methods to solve problems created by missing data and measurement error. For example, Lavy, Palumbo, and Stern (1998) and Stinebrickner (1999) use simulation methods to solve estimation problems created by missing data, and Stinebrickner and Stinebrickner (2004) develop a model of college outcomes that uses simulation methods to correct for measurement error in self-reported study time.

history files is aggregated into a yearly employment record for each individual. First, a primary job is assigned to each month based on the number of weeks worked in each job reported for the month. An individual's primary job for each year is defined as the one in which the most months were spent during that year. The yearly employment record is used to create a running tally of accumulated work experience in each occupation for each worker. This analysis considers only full-time employment, which is defined as a job where the weekly hours worked are at least 20.

Descriptions of the one-digit occupation classifications along with average wages are presented in Table 1. The highest paid workers are professional and managerial workers, while the lowest paid workers are found in the service occupation. Descriptive statistics are presented in Table 2. There are 954 individuals in the estimation sample who contribute a total of 10,573 person-year observations to the data. On average, each individual contributes approximately 11 observations to the data. Appendix 1 contains further details about the data used to estimate the model, including the details of how the sample is selected, and a discussion of the representativeness of the final sample.

### *A. Measurement Error in Occupation Codes*

The NLSY provides the U.S. Census occupation codes for each job. Interviewers question respondents about the occupation of each job held during the year with

**Table 1**
*Description of Occupations*

| One-Digit Occupation | Mean Wage | Example Three-Digit Occupations |
|---|---|---|
| Professional, technical, and kindred workers | \$11.19 | Accountants, chemical engineers, physicians, social scientists |
| Managers and administrators | \$12.89 | Bank officers, office managers, school administrators |
| Sales workers | \$9.05 | Advertising salesmen, real estate agents, stock and bond salesmen, salesmen and sales clerks |
| Clerical and unskilled workers | \$7.48 | Bank tellers, cashiers, receptionists, secretaries |
| Craftsmen and kindred workers | \$8.53 | Carpenters, electricians, machinists, stonemasons, mechanics |
| Operatives | \$7.20 | Dry wall installers, butchers, drill press operatives, truck drivers |
| Laborers | \$7.01 | Garbage collectors, groundskeepers, freight handlers, vehicle washers |
| Service workers | \$6.34 | Janitors, child care workers, waiters, guards and watchmen |

Notes: Based on the U.S Census occupation codes found in the 1979 cohort of the NLSY. Wages are in 1979 dollars.

**Table 2**
*Descriptive Statistics*

| Variable | NLSY Estimation Sample | Broader Sample from NLSY (for comparison) |
|---|---|---|
| Age | 24.4 | 25.5 |
| Years of high school | 3.5 | 3.7 |
| Years of college | 1.2 | 1.0 |
| Log wage | 1.95 | 1.98 |
| North Central | 0.32 | — |
| South | 0.30 | — |
| West | 0.17 | — |
| Professional | 0.14 | 0.12 |
| Managers | 0.11 | 0.11 |
| Sales | 0.05 | 0.06 |
| Clerical | 0.08 | 0.08 |
| Craftsmen | 0.25 | 0.24 |
| Operatives | 0.17 | 0.20 |
| Laborers | 0.10 | 0.11 |
| Service | 0.09 | 0.10 |
| Number of observations | 10,573 | 20,073 |
| Number of individuals | 954 | 1,932 |

Note: log wages in 1979 dollars. The NLSY estimation sample is described in Appendix 1. The broader sample relaxes an age restriction imposed on the estimation sample so it contains more individuals.

the following two questions: What kind of work do you do? That is, what is your occupation? Coders use these descriptions to classify each job using the three-digit Census occupation coding scheme. Misclassification of occupation codes may arise from errors made by respondents when describing their job, or from errors made by coders when interpreting these descriptions. Evidence on the extent of misclassification is provided by Mellow and Sider (1983), who perform a validation study of occupation codes using occupation codes found in the CPS matched with employer reports of their employee's occupation. They find agreement rates for occupation codes of 58 percent at the narrowly defined three-digit level and 81 percent at the more broadly defined one-digit level. Additional evidence on measurement error in occupation codes is presented by Mathiowetz (1992), who independently creates one and three-digit occupation codes based on occupational descriptions from employees of a large manufacturing firm and job descriptions found in these worker's personnel files. The agreement rate between these independently coded one-digit occupation codes is 76 percent, while the agreement rate for three-digit codes is only 52 percent. In addition to comparing the three- and one-digit occupation codes produced by independent coding, Mathiowetz (1992) also conducts a direct comparison of the company record with the employee's occupational description to see if the two sources could be classified as same three-digit occupation. This direct comparison results in an agreement rate of 87 percent at the three-digit level.

In general, papers examining occupational choices and the returns to occupation-specific work experience have not dealt with the difficult issues raised by measurement error in occupation codes even though it is widely believed that occupation codes are quite noisy. Work by Kambourov and Manovskii (2007) is a notable exception to this trend. They exploit the fact that the Panel Study of Income Dynamics (PSID) originally coded occupations using an approach similar to the NLSY in which occupation coders translated worker's verbatim descriptions of their occupation into an occupation code separately in each survey year. The PSID later released retrospective occupation data files where occupation coders were instead given access to a worker's complete sequence of occupational descriptions over his career. Kambourov and Manovskii (2007) show that occupational mobility is lower in the retrospective files, which is consistent with the hypothesis that coders introduce measurement error into occupation codes when they interpret workers' verbatim job descriptions. However, it is important to note that while this type of retrospective coding is likely to reduce the number of false occupational transitions found in the data, it does not provide any additional information about a worker's true occupation. Given this limitation of the PSID data, Kambourov and Manovskii (2007) estimate the returns to occupation-specific work experience, but they are not able to allow the wage equation to vary across occupations, or to estimate the importance of cross-occupation experience effects.

## III. Occupational Choice Model with Misclassification

### *A. A Baseline Model of Misclassification*

The model of occupational choices developed in this paper builds on previous models of sectoral and occupational choices such as Heckman and Sedlacek (1985, 1990) and Gould (2002). These models are all based on the framework of self-selection in occupational choices introduced by Roy (1951). Let $V^*_{iqt}$ represent the utility that worker $i$ receives from working in occupation $q$ at time period $t$. Let $N$ represent the number of people in the sample, let $T(i)$ represent the number of time periods that person $i$ is in the sample, and let $Q$ represent the number of occupations. Assume that the value of working in each occupation is the following function of the wage and nonpecuniary utility,

$$(1) \quad V^*_{iqt} = w_{iqt} + H_{iqt} + \varepsilon_{iqt},$$

where $w_{iqt}$ is the log wage of person $i$ in occupation $q$ at time $t$, $H_{iqt}$ is the deterministic portion of the nonpecuniary utility that person $i$ receives from working in occupation $q$ at time $t$, and $\varepsilon_{iqt}$ is an error term that captures variation in the utility flow from working in occupation $q$ caused by factors that are observed by the worker but unobserved by the econometrician.

The log wage equation is

$$(2) \quad w_{iqt} = \mu_{iq} + Z_{it}\beta_q + \sum_{k=1}^{Q} \delta_{qk} Exp_{ikt} + e_{iqt},$$

where $\mu_{iq}$ is the intercept of the log wage equation for person $i$ in occupation $q$, $Z_{it}$ is a vector of explanatory variables, and $Exp_{ikt}$ is person $i$ 's experience at time $t$ in

occupation $k$. This specification allows for a full set of cross-occupation experience effects, so the parameter estimates provide evidence on the transferability of skills across occupations. The final term, $e_{iqt,}$ represents a random wage shock. The deterministic portion of the nonpecuniary utility flow equation for person $i$ is specified as

$$(3) \quad H_{iqt} = X_{it}\pi_q + \sum_{k=1}^{Q} \gamma_{qk} Exp_{ikt} + \sum_{k=1}^{Q} \chi_{qk} Lastocc_{ikt} + \phi_{iq},$$

where $X_{it}$ is a vector of explanatory variables and $Lastocc_{ikt}$ is a dummy variable equal to one if person $i$ worked in occupation $k$ at time $t-1$. This variable allows switching occupations to have a direct impact on nonpecuniary utility, as it would if workers incur nonpecuniary costs when switching occupations. The final term, $\phi_{iq,}$ represents person $i$ 's innate preference for working in occupation $q$. In general, sectoral choice models of this type are identified even if the same explanatory variables appear in both the wage equation and the nonpecuniary utility flow equation. However, it is normally considered desirable to include a variable that impacts occupational choices but does not directly impact wages. In this application lagged occupational choice dummies and high school and college diploma dummies are included in the nonpecuniary equation but excluded from the wage equation. The exclusion of the lagged occupational choice dummies from the wage equation assumes that individuals incur psychic mobility costs when switching occupations, but there is no direct monetary switching cost. However, because occupation-specific experience effects vary across occupations, when an individual switches occupations his accumulated skills may be valued less highly in his new occupation.[4]

Let $O_{it}$ represent the occupational choice observed in the data for person $i$ at time $t$. This variable is an integer that takes a value ranging from one to $Q$. A person's true occupational choice may differ from the one observed in the data if classification error exists. Let $\hat{O}_{it}$ represent the true occupational choice, which is simply the occupation that yields the highest utility,

$$(4) \quad \hat{O}_{it} = q \quad if \quad V^*_{iqt} = \max\{V^*_{i1t}, V^*_{i2t}, \ldots, V^*_{iQt}\}$$

The model of misclassification allows the probability of misclassification to depend on the value of the latent variable $V^*_{iqt}$. The misclassification probabilities are denoted as

$$(5) \quad \alpha_{jk} = \Pr(O_{it} = j | \hat{O}_{it} = k), \quad for \quad j = 1, \ldots, Q \quad k = 1, \ldots, Q$$

That is, $\alpha_{jk}$ represents the probability that the occupation observed in the data is $j$, conditional on the actual occupational choice being $k$. The $\alpha$'s are estimated jointly along with the other parameters in the model. The $\alpha_{jj}$ terms are the probabilities that occupational choices are correctly classified. There are $Q \times Q$ misclassification probabilities, but there are only $[(Q \times Q) - Q]$ free parameters because the misclassification probabilities must sum to one for each possible occupational choice,

$$(6) \quad \sum_{j=1}^{Q} \alpha_{jk} = 1, \quad \text{for} \quad k = 1, \ldots, Q.$$

4. As in all selection models of this type, one can attempt to justify exclusion restrictions based on theory, but ultimately they are untestable assumptions.

Following existing parametric models of misclassification, the model assumes that the misclassification probabilities $\{\alpha_{jk} : k = 1, .., Q, j = 1, \ldots, Q\}$ depend only on $j$ and $k$, and not on the other explanatory variables in the model. One possible shortcoming of this baseline model of occupational misclassification is that it rules out person specific heterogeneity in the propensity to misclassify occupations that may be present in panel data such as the NLSY. Section D of this paper presents an extension of the model that allows for this type of within-person correlation in misclassification rates.

It is necessary to specify the distributions of the error terms in the model before deriving the likelihood function. Assume that $\varepsilon_{iqt} \sim$ iid extreme value and $e_{iqt} \sim N(0, \sigma^2_{eq})$. Let $\phi_i$ represent a $Q \times 1$ vector of person $i$'s preferences for working in each occupation, and let $\mu_i$ represent the $Q \times 1$ vector of person $i$ 's log wage intercepts in each occupation. Let $F(\mu, \phi)$ denote the joint distribution of the wage intercepts and occupational preferences.

Let $\theta$ represent the vector of parameters in the model, $\theta = \{\beta_k, \gamma_{kj}, \chi_{kj}, \alpha_{kj}, \pi_k, \delta_{jk}, \sigma_{ek}, F(\mu, \phi) : k = 1, \ldots, Q, j = 1, \ldots, Q\}$. For brevity of notation, when it is convenient I suppress some or all of the arguments $\{\theta, Z_{it}, X_{it}, Exp_{ikt}, Lastocc_{ikt}, w^{obs}_{it}\}$ at some points when writing equations for probabilities and likelihood contributions, even though the choice probabilities and likelihood contributions are functions of all of these variables. Define $\hat{P}_{it}(q, w^{obs}_{it})$ as the joint probability that person $i$ chooses to work in occupation $q$ in time period $t$ and receives a wage of $w^{obs}_{it}$. The outcome probability is

$$\hat{P}_{it}(q, w^{obs}_{it}|\mu, \phi) = \Pr(V^*_{iqt} = \max\{V^*_{i1t}, V^*_{i2t}, \ldots, V^*_{iQt}\}|w_{iqt} = w^{obs}_{it}) \times \Pr(w_{iqt} = w^{obs}_{it}). \quad (7)$$

There is no closed form solution for this probability, so it is approximated using simulation methods. This involves taking random draws from the distribution of the errors, and computing the mean of the simulated probabilities.[5] The likelihood function for the observed data is constructed using the misclassification probabilities and the true choice probabilities. Define $P_{it}(q, w^{obs}_{it})$ as the probability that person $i$ is *observed* working in occupation $q$ at time period $t$ with a wage of $w^{obs}_{it}$. This probability is the sum of the true occupational choice probabilities weighted by the misclassification probabilities,

$$P_{it}(q, w^{obs}_{it}|\mu, \phi) = \sum_{k=1}^{Q} \alpha_{qk} \hat{P}_{it}(k, w^{obs}_{it}|\mu, \phi). \quad (8)$$

Note that the outcome probability imposes the restriction that the observed wage is drawn from the worker's actual occupation, which rules out situations where a

5. A consequence of the extreme value assumption is that conditional on $\varepsilon$, this probability has a simple closed form solution. As a result it is straightforward to use a smooth simulator for the probabilities in the likelihood function. During estimation, 60 draws from the distribution of the errors are used to simulate the integral. Antithetic acceleration is used to reduce the variance of the simulated integral. As a check on the sensitivity of the estimates to the number of simulation draws the optimization routine was restarted using 600 draws. The parameter estimates (and value of the likelihood function at the maximum) were essentially unchanged by this increase in the number of simulation draws.

worker intentionally misrepresents his occupation and simultaneously provides a false wage consistent with the false occupation. The likelihood function is simply the product of the probabilities of observing the sequence of occupational choices observed in the data for each person over the years that they are in the sample,

$$(9)\quad L(\theta) = \prod_{i=1}^{N} \int \prod_{t=1}^{T(i)} \sum_{q=1}^{Q} 1\{O_{it} = q\} P_{it}(q, w_{it}^{obs} | \mu, \phi) dF(\mu, \phi)$$

$$(10)\qquad = \prod_{i=1}^{N} \int L_i(\theta | \mu, \phi) dF(\mu, \phi),$$

where $1\{\bullet\}$ denotes the indicator function which is equal to one if its argument is true and zero otherwise. The likelihood function must be integrated over the joint distribution of skills and preferences, $F(\mu, \phi)$. Following Heckman and Singer (1984), this distribution is specified as a discrete multinomial distribution.[6] Suppose that there are $M$ types of people, each with a $Q \times 1$ vector of wage intercepts $\mu^m$ and $Q \times 1$ vector of preferences $\phi^m$. Let $\omega_m$ represent the proportion of the $m$th type in the population. The unconditional likelihood function is simply a weighted average of the type specific likelihoods,

$$\begin{aligned}(11)\quad L(\theta) &= \prod_{i=1}^{N} \int L_i(\theta | \mu, \phi) dF(\mu, \phi) \\ &= \prod_{i=1}^{N} \sum_{m=1}^{M} \omega_m L_i(\theta | \mu_i = \mu^m, \phi_i = \phi^m) \\ &= \prod_{i=1}^{N} L_i(\theta)\end{aligned}$$

### *B. Evaluating the Likelihood Function*

The major complication in evaluating the likelihood function arises from the fact that classification error in occupation codes creates nonclassical measurement error in the observed occupation-specific work experience variables and previous occupational choice dummy variables that describe an individual's state. This implies that the true state of each agent is unobserved. Previous research into occupational choices has not addressed this issue. The key to understanding the solution to this problem is to realize that the model of misclassification implies a distribution of true values of occupation-specific work experience and lagged occupational choices for each individual in each time period. Estimating the parameters of the model by maximum likelihood involves integrating over the distribution of these unobserved state variables. However, there is no closed form solution for this integral, and, more importantly, the distribution is intractably complex. These problems are solved by

6. There is a large literature advocating the use of discrete distributions for unobserved heterogeneity. See, for example, Mroz (1999).

simulating the likelihood function. The algorithm involves recursively simulating $R$ sequences of occupation-specific work experience and lagged occupational choices that span a worker's entire career. The individual's likelihood contribution is computed for each simulated sequence, and the path probabilities are averaged over the $R$ sequences to obtain the simulated likelihood contribution. A detailed description of the simulation algorithm is presented in Appendix 2.

### *C. Identification*

This section presents the identification conditions for the occupational choice model with misclassification and discusses several related issues.

#### *1. Identification Conditions*

The identification conditions for a model of misclassification in a binary dependant variable are presented by Hausman, Abrevaya, and Scott-Morton (1998). This condition is extended to the case of discrete choice models with more than two outcomes by Ramalho (2002). The parameters of the model are identified if the sum of the conditional misclassification probabilities for each observed outcome is smaller than the conditional probability of correct classification. In the context of the occupational choice model presented in this paper this condition amounts to the following restriction on the misclassification probabilities,

$$(12) \quad \sum_{k \neq j} \alpha_{jk} < \alpha_{jj},\ j = 1, \ldots, Q.$$

This condition implies that it is not possible to estimate the extent of misclassification along with the rest of the parameter vector if the quality of the data is so poor that one is more likely to observe a misclassified occupational choice than a correctly classified occupational choice.

#### *2. Discussion*

Estimating the extent of classification error in the NLSY occupation data along with the parameters of the occupational choice model is only possible if one is willing to adopt a parametric model along with the associated functional form and distributional assumptions.[7] It is worthwhile to consider at an intuitive level how the parametric occupational choice model and misclassification model are linked together. Let $\tilde{\beta}$ represent the parameter vector for the occupational choice model, and let $\tilde{\alpha}$ represent the vector of misclassification parameters. Given $\tilde{\beta}$ the parametric model of occupational choices provides the probability that each occupational choice and wage combination observed in the NLSY is generated by the model. Taking $\tilde{\beta}$ as given, one could choose the value of $\tilde{\alpha}$ that maximizes the probability of observing the NLSY occupation and wage data. Broadly speaking, this will happen when the combinations of occupational choices and wages that are unlikely to be generated by the model at the parameter vector $\tilde{\beta}$ are assigned a relatively high probability of being affected by misclassification. During estimation, $\tilde{\beta}$ is not fixed, it is

7. It should be noted that as is the case with all parametric models of this type, if the model is misspecified, parameter estimates will be biased.

estimated simultaneously with $\tilde{\alpha}$, so estimating the model amounts to choosing the value of $\tilde{\beta}$ that best fits the data, with the added consideration that the chosen value of $\tilde{\alpha}$ allows misclassification to account for some of the observed patterns in the data.

Existing parametric models of misclassification estimate misclassification rates using discrete choice models, while in contrast this paper jointly models discrete occupational choices along with wages. The advantage of this approach is that to the extent that wages vary across occupations, observed wages provide information about which observed choices are likely to be affected by misclassification.[8] This approach uses information about the relationship between observable variables (such as education) and occupational choices, along with information about the consistency of observed wages with reported occupations to infer the extent of misclassification in the data. It should be noted that when occupations are measured with error, it is not possible to nonparametrically determine the exact relationship between true occupational choices, wages, and observable variables such as education. However, within a particular parametric model of occupational choices and wages, these parameters can be estimated.[9]

The availability of validation data on occupations from an outside data source would, in principle, allow one to relax some of the parametric assumptions adopted in this paper. For example, if another data set contained information about reported occupations, true occupations, and possibly other explanatory variables, this information could be used to integrate out the effect of measurement error. Of course, this approach relies on the assumption that the measurement error process is identical in the two sources of data. While this approach appears promising and is certainly worth pursuing in future research, on a practical level adopting this approach would most likely require additional data collection that was targeted specifically at validating occupation codes.[10] One possible approach would be to validate an individual's occupation by

8. In the extreme case where the wage distribution is identical across occupations observed wages do not provide any additional information about misclassification. However, even if the unconditional wage distribution is identical across occupations, if the wage distribution in each occupation is a function of observable characteristics (such as education and occupation-specific experience), and the effects of these variables on wages vary across occupations, then observed wages will still provide information about misclassification.

9. Although panel data is used to estimate the model, it is also possible to estimate this type of model using cross-sectional data. As an experiment, I randomly selected a cross-section of workers from the panel data NLSY sample and reestimated the model. The estimated level of misclassification in the cross-sectional version of the model was 8 percent, compared to 9 percent in the panel data version. The fact that these estimates are so close suggests that misclassification rates are primarily identified by the consistency of an individual's reported occupation with the cross-sectional distribution of choices, wages, and observable variables, rather than by the extent to which an observed occupational choice is consistent with an individual worker's observed sequence of career choices.

10. The major problem is that existing validation studies, such as the 1977 supplement to the CPS, question respondents about their occupation and then attempt to validate the reported occupations by surveying employers. In general there is no reason to be confident that the employer surveys provide occupation data that is free from error. Depending on the information contained in personnel files and the system that an employer uses to categorize employees, the responses provided by firms could in fact be noisier than those provided by individuals. As a result, it is generally accepted that these validation studies provide an upper bound on the extent of measurement error. In contrast, validating wage data appears to be a much simpler task, since one would expect that firms could normally provide accurate salary information from their payroll records.

questioning his supervisor, since presumably supervisors know the type of work performed by workers that they manage. This approach would circumvent some of the problems associated with validating occupation codes using personnel records, which may or may not contain job descriptions that accurately reflect occupations.

### *D. An Extended Model: Heterogeneity in Misclassification Rates*

The model of misclassification presented in Section III, Subsection A assumes that all individuals have the same probability of having one of their occupational choices misclassified. In a panel data setting such as the NLSY, it is possible that during the yearly NLSY interviews some individuals consistently provide poor descriptions of their jobs that are likely to lead to measurement error in the occupation codes created by the NLSY coders. On the other hand, some workers may be more likely to provide accurate descriptions of their occupations that are extremely unlikely to be misclassified. The remainder of this section extends the occupational choice model with misclassification to allow for time persistent misclassification by using an approach similar to the one adopted by Dustmann and van Soest (2001) in their study of misclassification of language fluency.

The primary goal of the extended model is to allow for person-specific heterogeneity in misclassification rates in a way that results in a tractable empirical model. Suppose that there are three subpopulations of workers in the economy, and that these subpopulations each have different probabilities of having their occupational choices misclassified. Define the occupational choice misclassification probabilities for Subpopulation $y$ as

$$(13) \quad \alpha_{jk}(y) = \Pr(O_{it} = j | \hat{O}_{it} = k), \quad j = 1, \ldots, Q \quad k = 1, \ldots, Q$$

$$(14) \quad \sum_{j=1}^{Q} \alpha_{jk}(y) = 1 \quad k = 1, \ldots, Q, \quad y = 1, 2, 3.$$

Denote the proportion of Subpopulation $y$ in the economy as $\xi(y)$,where $y = 1, 2, 3$ and $\sum_{y=1}^{3} \xi(y) = 1$. This specification of the misclassification rates allows for time-persistence in misclassification, since the $\alpha_{jk}(y)$'s are fixed over time for each subpopulation. During estimation the $\xi(y)$'s and $\alpha_{jk}(y)$'s of each subpopulation are estimated along with the other parameters of the model, so it is necessary to specify the misclassification model in such a way that the number of parameters in the model does not become unreasonably large. In order to keep the number of parameters at a tractable level, the number of subpopulations is set to a small number (3), and the misclassification probabilities are restricted during estimation so that the occupational choices of subpopulation one are always correctly classified.[11]

This model of misclassification incorporates the key features of heterogeneous misclassification rates in a fairly parsimonious way. Some fraction of the population ($\xi(1)$) is always correctly classified, and the remaining two subpopulations are allowed to have completely different misclassification rates, so that both the overall

11. This version of the model already has 421 parameters that must be estimated, so in order to keep the model tractable it was never estimated with more than three subpopulations.

level of misclassification and the particular patterns in misclassification are allowed to vary between subpopulations.

The likelihood function presented in Section III, Subsection A can be modified to account for person-specific heterogeneity in misclassification. The observed choice probabilities are easily modified so that they are allowed to vary by subpopulation,

$$(15)\quad P_{it}(q, w_{it}^{obs}|\mu, \phi, y) = \sum_{k=1}^{Q} \alpha_{qk}(y)\hat{P}_{it}(k, w_{it}^{obs}|\mu, \phi),$$

where $y = 1,\ldots,3$ indexes subpopulations. Conditional on subpopulations, the likelihood function is

$$(16)\quad L(\theta|y) = \prod_{i=1}^{N} \int \prod_{t=1}^{T(i)} \sum_{q=1}^{Q} 1\{O_{it} = q\} P_{it}(q, w_{it}^{obs}|\mu, \phi, y) dF(\mu, \phi)$$
$$= \prod_{i=1}^{N} \int L_i(\theta|\mu, \phi, y) dF(\mu, \phi)$$

The subpopulation that a particular person belongs to is not observed, so the likelihood function must be integrated over the discrete distribution of the type-specific misclassification rates,

$$(17)\quad L(\theta) = \prod_{i=1}^{N} \sum_{y=1}^{3} \sum_{m=1}^{M} \xi(y)\omega_m L_i(\theta|y, \mu_i = \mu^m, \phi_i = \phi^m) = \prod_{i=1}^{N} L_i(\theta).$$

## IV. Parameter Estimates

This section presents the simulated maximum likelihood parameter estimates for the occupational choice model. First, the parameters that reveal the extent of classification error in reported occupations are discussed, and then the parameter estimates from the occupational choice model that corrects for classification error and allows for person-specific heterogeneity in misclassification are compared to the estimates from a model that does not correct for measurement error. Next, the sensitivity of the estimates to measurement error in wages is examined. Finally, the model is used to simulate data that is free from classification error in occupation codes.

### *A. The Extent of Measurement Error in Occupation Codes*

The estimates of the misclassification probabilities for Subpopulations 2 and 3 along with the estimated proportions of each type in the population are presented in Panels A and B of Table 4. The bottom row of Panel A shows that correcting for classification error results in a large improvement in the fit of the model, since the likelihood function improves from $-18,695$ when classification error is ignored to $-17,821$ when classification error is corrected for. The probability in Row $i$, Column $j$ is

**Table 3**
*Occupational Transition Matrix—NLSY Data (top entry) and Simulated Data (bottom entry)*

| | Professional | Managers | Sales | Clerical | Craftsmen | Operatives | Laborers | Service |
|---|---|---|---|---|---|---|---|---|
| Professional | 74.7 | 6.9 | 2.3 | 4.5 | 5.0 | 3.0 | 1.3 | 2.2 |
| | 78.5 | 5.6 | 4.2 | 3.7 | 3.2 | 2.2 | 1.4 | 1.2 |
| Managers | 6.4 | 57.4 | 7.2 | 7.3 | 10.7 | 3.5 | 2.5 | 5.0 |
| | 6.6 | 58.5 | 9.4 | 7.4 | 10.3 | 2.9 | 2.6 | 2.3 |
| Sales | 8.0 | 14.9 | 53.5 | 7.7 | 5.4 | 5.2 | 2.2 | 3.2 |
| | 7.6 | 9.2 | 55.2 | 6.3 | 6.8 | 5.9 | 5.2 | 3.6 |
| Clerical | 10.3 | 12.4 | 5.9 | 44.8 | 6.8 | 7.0 | 8.3 | 4.6 |
| | 8.7 | 11.4 | 7.2 | 45.8 | 6.3 | 6.8 | 9.8 | 4.0 |
| Craftsmen | 2.9 | 5.3 | 1.0 | 2.2 | 66.6 | 11.1 | 8.1 | 2.6 |
| | 2.0 | 4.7 | 2.3 | 2.0 | 67.4 | 10.8 | 9.6 | 1.2 |
| Operatives | 2.4 | 2.2 | 2.1 | 3.1 | 18.4 | 56.8 | 10.1 | 4.9 |
| | 1.9 | 1.3 | 3.3 | 2.9 | 18.3 | 56.3 | 11.6 | 4.4 |
| Laborers | 2.7 | 3.3 | 1.8 | 7.9 | 23.2 | 18.6 | 36.2 | 6.1 |
| | 2.5 | 2.7 | 4.0 | 7.3 | 21.6 | 16.5 | 39.1 | 6.2 |
| Service | 3.9 | 7.8 | 1.5 | 3.5 | 8.4 | 6.8 | 8.6 | 59.5 |
| | 3.7 | 4.2 | 2.8 | 3.1 | 6.8 | 6.2 | 9.5 | 63.7 |
| Total | 14.0 | 11.5 | 5.3 | 7.6 | 25.8 | 16.9 | 9.6 | 9.4 |
| | 13.9 | 9.5 | 7.9 | 7.3 | 25.2 | 16.2 | 11.5 | 8.4 |

Entries are the percentage of employment spells starting in the occupation listed in the left column that ends in the occupation listed in the top row.

the estimate of $\alpha_{ij}(y)$, which is the probability that occupation $i$ is observed in the data conditional on occupation $j$ being the actual choice for a person in Subpopulation $y$. For example, the entry in the third column of the first row indicates that condition of being a member of Subpopulation 2, there is a 2.6 percent chance that a person who is actually a sales worker will be misclassified as a professional worker. The diagonal elements of the two panels of Table 4 show the probabilities that occupational choices are correctly classified. Averaged across all occupations, the probability that an occupational choice is correctly classified is 0.868 for Subpopulation 2 and 0.840 for Subpopulation 3. One striking feature of the estimated misclassification probabilities is that they provide substantial evidence that misclassification rates vary widely across occupations. For example, in Subpopulation 2 the probability that an occupational choice is correctly classified ranges from a low of 0.56 for sales workers to a high of 0.99 for craftsmen, while in Subpopulation 3 the probability that an occupational choice is correctly classified ranges from a low of 0.60 for sales workers to a high of 0.98 for operatives.

The estimates of the probabilities that a person belongs to Subpopulations 2 and 3 are 42 percent and 19 percent, which leaves an estimated 38 percent of the population belonging to Subpopulation 1, the group whose occupational choices are never misclassified. The fact that a substantial fraction of the population belongs to the subpopulation whose occupational choices are never misclassified highlights the importance of allowing for person-specific heterogeneity in misclassification rates. When averaged over subpopulations, the subpopulation-specific misclassification rates

**Table 4**
*Parameter Estimates- Misclassification Probabilities for Subpopulation 2 ($\alpha_{jk}(2)$)*

Panel A

| Observed/Actual | Professional | Managers | Sales | Clerical | Craftsmen | Operatives | Laborers | Service |
|---|---|---|---|---|---|---|---|---|
| Professional | 0.9570 | 0.0018 | 0.0264 | 0.0017 | 0.0033 | 0.0007 | 0.0380 | 0.0641 |
| | (0.0023) | (0.0096) | (0.0025) | (0.0074) | (0.0002) | (0.0004) | (0.0004) | (0.0017) |
| Managers | 0.0066 | 0.9762 | 0.2128 | 0.0052 | 0.0013 | 0.0021 | 0.0011 | 0.0238 |
| | (0.0041) | (0.0042) | (0.0026) | (0.0082) | (0.0002) | (0.0015) | (0.0022) | (0.0003) |
| Sales | 0.0036 | 0.0148 | 0.5578 | 0.0133 | 0.0000 | 0.0029 | 0.0019 | 0.0774 |
| | (0.0016) | (0.0046) | (0.0001) | (0.0045) | (0.0015) | (0.0017) | (0.0040) | (0.0009) |
| Clerical | 0.0131 | 0.0031 | 0.0131 | 0.9579 | 0.0002 | 0.0021 | 0.0046 | 0.0042 |
| | (0.0002) | (0.0101) | (0.0055) | (0.0046) | (0.0016) | (0.0022) | (0.0067) | (0.0033) |
| Craftsmen | 0.0055 | 0.0022 | 0.1063 | 0.0052 | 0.9897 | 0.0055 | 0.0204 | 0.0024 |
| | (0.0023) | (0.0045) | (0.0098) | (0.0025) | (0.0054) | (0.0030) | (0.0121) | (0.0023) |
| Operatives | 0.0121 | 0.0000 | 0.0456 | 0.0013 | 0.0000 | 0.9849 | 0.0063 | 0.0004 |
| | (0.0025) | (0.0064) | (0.0005) | (0.0082) | (0.0039) | (0.0058) | (0.0009) | (0.0223) |
| Laborers | 0.0000 | 0.0000 | 0.0164 | 0.0136 | 0.0054 | 0.0016 | 0.7029 | 0.0039 |
| | (0.0003) | (0.0131) | (0.0043) | (0.0085) | (0.0021) | (0.0082) | (0.0014) | (0.0079) |
| Service | 0.0018 | 0.0018 | 0.0213 | 0.0014 | 0.0000 | 0.0000 | 0.2243 | 0.8235 |
| | (.0002) | (0.0043) | (0.0008) | (0.0086) | (0.0018) | (0.0022) | (0.0012) | (0.0068) |
| Pr(sub population 2) | 0.4243 (0.0211) | | | | | | | |
| | Ignore misclassification | Correct for misclassification | | | | | | |
| Log-likelihood | −18,695 | −17,821 | | | | | | |

Notes: Element $\alpha(i,j)$ of this table, where $i$ refers to the row and $j$ refers to the column is the probability that occupation $i$ is observed, conditional on $j$ being the true choice: $\alpha(j,k)=Pr(\text{occupation } j \text{ observed} \mid \text{occupation } k \text{ is true choice})$. Standard errors in parentheses. "Subpopulation" refers to the fact that the misclassification model allows for heterogeneity in misclassification rates.

indicate that 91 percent of one-digit occupational choices are correctly classified. This estimate of the overall extent of misclassification in the NLSY data is lower than the misclassification rates reported in validation studies based on other datasets. For example, Mellow and Sider (1983) find an agreement rate of 81 percent at the one-digit level between employee's reported occupations and employer's occupational descriptions in the January 1977 Current Population Survey (CPS). Mathiowetz (1992) finds a 76 percent agreement rate between the occupational descriptions given by workers of a single large manufacturing firm and personnel records.[12]

One possible explanation for the lower misclassification rate found in this study compared to the validation studies is that the NLSY occupation data is of higher quality than both the CPS data and the survey conducted by Mathiowetz (1992). It appears that the procedures used by the CPS and NLSY in constructing occupation codes are quite similar, so it is not clear that one should expect the NLSY data to

12. This study is the first to estimate a parametric model of occupational misclassification, so the validation studies provide the only basis for comparison for the estimated misclassification rates.

**Table 4**
*Misclassification Probabilities for Subpopulation 3 ($\alpha_{ij}(3)$)*

Panel B

| Observed/Actual | Professional | Managers | Sales | Clerical | Craftsmen | Operatives | Laborers | Service |
|---|---|---|---|---|---|---|---|---|
| Professional | 0.9289 | 0.0043 | 0.0394 | 0.0012 | 0.0357 | 0.0007 | 0.0104 | 0.0190 |
| | (0.0022) | (0.0097) | (0.0024) | (0.0079) | (0.0005) | (0.0003) | (0.0003) | (0.0016) |
| Managers | 0.0099 | 0.9641 | 0.0822 | 0.0032 | 0.0046 | 0.0030 | 0.0041 | 0.2548 |
| | (0.0040) | (0.004) | (0.0027) | (0.0081) | (0.0003) | (0.0016) | (0.0021) | (0.0002) |
| Sales | 0.0096 | 0.0248 | 0.6007 | 0.0125 | 0.0002 | 0.0003 | 0.0022 | 0.0026 |
| | (0.0016) | (0.0056) | (0.0001) | (0.0046) | (0.0016) | (0.0018) | (0.0041) | (0.0008) |
| Clerical | 0.0126 | 0.0027 | 0.0052 | 0.9634 | 0.0004 | 0.0012 | 0.0031 | 0.0006 |
| | (0.0001) | (0.0103) | (0.0054) | (0.0045) | (0.0011) | (0.0023) | (0.0061) | (0.0037) |
| Craftsmen | 0.0234 | 0.0025 | 0.0904 | 0.0068 | 0.9475 | 0.0067 | 0.0504 | 0.0041 |
| | (0.0024) | (0.0043) | (0.0096) | (0.0024) | (0.0061) | (0.0031) | (0.0130) | (0.0026) |
| Operatives | 0.0106 | 0.0007 | 0.1335 | 0.0029 | 0.00000 | 0.9833 | 0.0054 | 0.0000 |
| | (0.0026) | (0.0065) | (0.0005) | (0.0081) | (0.0047) | (0.0051) | (0.0008) | (0.0001) |
| Laborers | 0.0000 | 0.0000 | 0.0307 | 0.0082 | 0.0114 | 0.0040 | 0.6215 | 0.0042 |
| | (0.0001) | (0.0141) | (0.0041) | (0.0084) | (0.0020) | (0.0084) | (0.0005) | (0.0069) |
| Service | 0.0049 | 0.0008 | 0.0176 | 0.0016 | 0.0000 | 0.0006 | 0.3028 | 0.7139 |
| | (0.0002) | (0.0043) | (0.0009) | (0.0088) | (0.0017) | (0.0024) | (0.0008) | (0.0048) |
| Pr(subpopulation 3) | 0.1937 | | | | | | | |
| | (0.0235) | | | | | | | |

Notes: Element $\alpha(i,j)$ of this table, where $i$ refers to the row and $j$ refers to the column is the probability that occupation $i$ is observed, conditional on $j$ being the true choice: $\alpha(j,k) = \Pr(\textit{occupation j observed} \mid \textit{occupation k is true choice})$. Standard errors in parentheses. "Subpopulation" refers to the fact that the misclassification model controls for unobserved heterogeneity in misclassification rates by allowing for a discrete number of subpopulations that are each allowed to have different misclassification matrices.

have a lower misclassification rate than the CPS. An alternative explanation is that the employer reports of occupation codes that are assumed to be completely free from classification error in validation studies are in fact measured with error.[13] If this is true, then comparing noisy self-reported data to noisy employer reported data would cause validation studies to overstate the extent of classification error in occupation codes. The idea that this type of validation study may result in an overstatement of classification error in occupation or industry codes is not a new one. For example, Krueger and Summers (1988) assume that the error rate for one-digit industry classifications is half as large as the one reported by Mellow and Sider (1983) as a rough correction for the overstatement of classification error in validation studies.

The wide variation in misclassification rates across occupations along with the patterns in misclassification suggest that certain types of jobs are likely to be misclassified in particular directions. In addition, the misclassification matrix is highly asymmetric. For example, there is only a 1.4 percent chance that a manager will be misclassified as a sales worker, but there is a 21 percent chance that a sales worker

13. It is widely acknowledged that although validation studies are frequently based on the premise that one source of data is completely free from error, in reality no source of data will be completely free from measurement error. See Bound, Brown, and Mathiowetz (2001) for a discussion of this issue.

will be misclassified as a manager. Reading down the laborers column of Panel A of Table 4 shows that laborers are frequently misclassified as service workers (22 percent), but service workers are very unlikely to be misclassified as laborers (.39 percent). Further evidence of asymmetric misclassification is found throughout Table 4.

### *B. Occupational Choice Model Parameter Estimates*

The parameter estimates for the occupational choice model estimated with and without correcting for classification error are presented in Table 5. In addition, this table presents a measure of the difference between each parameter in the baseline ($\beta_b$) and

**Table 5**
*Parameter Estimates—Wage Equation*

Panel A

| | Professional | | | Managers | | |
|---|---|---|---|---|---|---|
| Wage equation | Ignore Classification Error | Correct for Classification error | Normalized Difference | Ignore Classification Error | Correct for Classification Error | Normalized difference |
| Age | 0.0233 | 0.0079 | | 0.0474 | 0.0351 | |
| | (0.0154) | (0.0073) | 2.12 | (0.0184) | (0.0091) | 1.35 |
| Age$^2$/100 | −0.2280 | −0.1434 | | −0.4028 | −0.3543 | |
| | (0.0985) | (0.0426) | −1.98 | (0.1230) | (0.0630) | −0.77 |
| Education | 0.0734 | 0.0626 | | 0.0825 | 0.0837 | |
| | (0.0057) | (0.0041) | 2.66 | (0.0082) | (0.0060) | −0.20 |
| Professional experience | 0.0715 | 0.0687 | | 0.0944 | 0.0896 | |
| | (0.0053) | (0.0034) | 0.81 | (0.0130) | (0.0086) | 0.56 |
| Managerial experience | 0.0375 | 0.0644 | | 0.0656 | 0.0547 | |
| | (0.0158) | (0.0123) | −2.19 | (0.0071) | (0.0055) | 1.99 |
| Sales experience | 0.0493 | 0.0499 | | 0.0888 | 0.0879 | |
| | (0.0147) | (0.0101) | −0.06 | (0.0135) | (0.0097) | 0.09 |
| Clerical experience | 0.0430 | 0.0377 | | 0.0191 | 0.0209 | |
| | (0.0191) | (0.0162) | 0.33 | (0.0096) | (0.0073) | −0.25 |
| Craftsmen experience | 0.0280 | 0.0203 | | 0.0488 | 0.0556 | |
| | (0.0092) | (0.0100) | 0.77 | (0.0074) | (0.0062) | −1.10 |
| Operatives experience | 0.0447 | 0.0259 | | 0.0634 | 0.0705 | |
| | (0.0236) | (0.0210) | 0.90 | (0.0124) | (0.0121) | −0.59 |
| Laborer experience | 0.0146 | −0.0083 | | 0.0416 | 0.0233 | |
| | (0.0291) | (0.0232) | 0.99 | (0.0268) | (0.0179) | 1.02 |
| Service experience | 0.0000 | 0.0718 | | 0.0100 | 0.0069 | |
| | (0.0224) | (0.0234) | −3.07 | (0.0140) | (0.0117) | 0.26 |
| North central | −0.0635 | −0.0139 | | −0.1063 | −0.0667 | |
| | (0.0262) | (0.0189) | −2.63 | (0.0302) | (0.0233) | −1.70 |
| South | −0.0448 | 0.0222 | | −0.0726 | −0.0849 | |
| | (0.0245) | (0.0182) | −3.69 | (0.0345) | (0.0284) | 0.43 |
| West | 0.0412 | 0.1046 | | −0.0919 | −0.0531 | |
| | (0.0294) | (0.0205) | −3.09 | (0.0438) | (0.0311) | −1.25 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

**Table 5**
*Parameter Estimates—Wage Equations*

Panel A

| | Sales | | | Clerical | | |
|---|---|---|---|---|---|---|
| Wage equation | Ignore Classification Error | Correct For Classification Error | Normalized Difference | Ignore Classification Error | Correct For Classification Error | Normalized Difference |
| Age | 0.0662 | 0.1354 | | 0.0480 | 0.0413 | |
| | (0.0368) | (0.0272) | −2.55 | (0.0153) | (0.0157) | 0.43 |
| Age$^2$/100 | −1.0006 | −1.0984 | | −0.4330 | −0.3588 | |
| | (0.2662) | (0.1886) | 0.52 | (0.1057) | (0.1095) | −0.68 |
| Education | 0.1837 | 0.1593 | | 0.0528 | 0.0511 | |
| | (0.0189) | (0.0268) | 0.91 | (0.0087) | (0.0081) | 0.21 |
| Professional experience | 0.0672 | 0.0308 | | 0.0957 | 0.1051 | |
| | (0.0366) | (0.0542) | 0.67 | (0.0146) | (0.0230) | −0.41 |
| Managerial experience | 0.1316 | 0.1089 | | 0.0454 | 0.0418 | |
| | (0.0274) | (0.0322) | 0.71 | (0.0104) | (0.0121) | 0.30 |
| Sales experience | 0.1774 | 0.1571 | | 0.0806 | 0.0888 | |
| | (0.0163) | (0.0195) | 1.04 | (0.0162) | (0.0203) | −0.40 |
| Clerical Experience | 0.1281 | 0.0430 | | 0.0562 | 0.0572 | |
| | (0.0333) | (0.0433) | 1.97 | (0.0085) | (0.0093) | −0.11 |
| Craftsmen experience | −0.0183 | −0.0453 | | 0.0502 | 0.0646 | |
| | (0.0258) | (0.0297) | 0.91 | (0.0083) | (0.0119) | −1.21 |
| Operatives experience | 0.0845 | 0.0845 | | 0.0516 | 0.0500 | |
| | (0.0284) | (0.0297) | 0.00 | (0.0118) | (0.0125) | 0.13 |
| Laborer experience | 0.0507 | 0.0521 | | 0.0420 | 0.0345 | |
| | (0.0431) | (0.0552) | −0.03 | (0.0167) | (0.0153) | 0.49 |
| Service experience | 0.0241 | −0.0657 | | 0.0191 | 0.0215 | |
| | (0.0295) | (0.0826) | 1.09 | (0.0177) | (0.0183) | −0.13 |
| North Central | −0.2505 | −0.3711 | | −0.1688 | −0.1965 | |
| | (0.0754) | (0.1051) | 1.15 | (0.0311) | (0.0369) | 0.75 |
| South | 0.1225 | 0.1249 | | −0.0847 | −0.1030 | |
| | (0.0764) | (0.0915) | −0.03 | (0.0307) | (0.0377) | 0.49 |
| West | 0.0979 | 0.1015 | | −0.0228 | −0.0230 | |
| | (0.0945) | (0.1070) | −0.03 | (0.0342) | (0.0362) | 0.01 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

classification error ($\beta_{ce}$) models, $(\beta_b - \beta_{ce})/se(\beta_{ce})$, where $se(\beta_{ce})$ is the standard error of $\beta_{ce}$. In the remainder of the paper this standard error normalized difference will be referred to as the normalized change in the parameter.

### *1. Wage Equation*

While theoretical results regarding the effects of measurement error in simple linear models have been derived, there are no clear predictions for nonlinear models such as this occupational choice model. Broadly speaking, one would expect the patterns of

**Table 5**
*Parameter Estimates—Wage Equations*

Panel A

| | Craftsmen | | | Operatives | | |
|---|---|---|---|---|---|---|
| Wage equation | Ignore Classification Error | Correct For Classification Error | Normalized Difference | Ignore Classification Error | Correct For Classification Error | Normalized Difference |
| Age | 0.0606 | 0.0489 | | 0.0128 | 0.0123 | |
| | (0.0068) | (0.0053) | 2.21 | (0.0085) | (0.0073) | 0.07 |
| $Age^2/100$ | −0.5257 | −0.4576 | | −0.2230 | −0.2605 | |
| | (0.0481) | (0.0398) | −1.71 | (0.0642) | (0.0604) | 0.62 |
| Education | 0.0290 | 0.0254 | | 0.0209 | 0.0079 | |
| | (0.0048) | (0.0045) | 0.80 | (0.0054) | (0.0048) | 2.74 |
| Professional experience | 0.0290 | 0.0188 | | 0.0670 | 0.0751 | |
| | (0.0120) | (0.0210) | 0.49 | (0.0229) | (0.0344) | −0.24 |
| Managerial experience | 0.0558 | 0.0646 | | 0.0432 | 0.0552 | |
| | (0.0115) | (0.0113) | −0.78 | (0.0157) | (0.0152) | −0.79 |
| Sales experience | 0.0100 | 0.0438 | | 0.0200 | 0.0157 | |
| | (0.0169) | (0.0183) | −1.85 | (0.0149) | (0.0176) | 0.24 |
| Clerical experience | 0.0381 | 0.0366 | | 0.0499 | 0.0370 | |
| | (0.0125) | (0.0210) | 0.07 | (0.0110) | (0.0191) | 0.68 |
| Craftsmen experience | 0.0591 | 0.0605 | | 0.0607 | 0.0764 | |
| | (0.0028) | (0.0027) | −0.52 | (0.0067) | (0.0062) | −2.52 |
| Operatives experience | 0.0386 | 0.0352 | | 0.0549 | 0.0470 | |
| | (0.0052) | (0.0048) | 0.71 | (0.0045) | (0.0041) | 1.92 |
| Laborer experience | 0.0217 | 0.0114 | | 0.0708 | 0.0512 | |
| | (0.0069) | (0.0066) | 1.57 | (0.0090) | (0.0077) | 2.56 |
| Service experience | 0.0254 | 0.0361 | | −0.0023 | 0.0285 | |
| | (0.0094) | (0.0106) | −1.00 | (0.0149) | (0.0147) | −2.10 |
| North Central | −0.1034 | −0.1201 | | −0.0637 | −0.0948 | |
| | (0.0197) | (0.0185) | 0.91 | (0.0266) | (0.0222) | 1.40 |
| South | −0.0786 | −0.0828 | | 0.0234 | 0.0026 | |
| | (0.0209) | (0.0182) | 0.23 | (0.0270) | (0.0222) | 0.94 |
| West | 0.0847 | 0.0868 | | 0.0086 | −0.0043 | |
| | (0.0210) | (0.0208) | −0.10 | (0.0307) | (0.0268) | 0.48 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

misclassification present in the data to be a key determinant of the magnitude and direction of the resulting bias. Due to the large number of wage equation parameters, this discussion focuses on a small subset of parameter estimates with the goal of demonstrating that classification error is something that needs to be accounted for when estimating occupation-specific wage equations. In addition, this discussion will attempt to highlight the type of questions in general that one might receive misleading answers to if one examines occupational choices and ignores misclassification.

The wage equation parameter estimates are presented in Panel A of Table 5. The estimates of the wage equation for the professional occupation show a number of

**Table 5**
*Parameter Estimates—Wage Equations*

Panel A

| | Laborers | | | Service | | |
|---|---|---|---|---|---|---|
| Wage equation | Ignore Classification Error | Correct for Classification Error | Normalized Difference | Ignore Classification Error | Correct for Classification Error | Normalized Difference |
| Age | 0.0268 | 0.0235 | | −0.0083 | −0.0116 | |
| | (0.0119) | (0.0119) | 0.28 | (0.0120) | (0.0093) | 0.35 |
| Age$^2$/100 | −0.3202 | −0.3339 | | 0.0234 | 0.0314 | |
| | (0.0994) | (0.0961) | 0.14 | (0.0889) | (0.0666) | −0.12 |
| Education | 0.0331 | 0.0184 | | 0.0965 | 0.0864 | |
| | (0.0087) | (0.0077) | 1.92 | (0.0071) | (0.0070) | 1.45 |
| Professional experience | 0.0715 | 0.0295 | | 0.0285 | 0.0274 | |
| | (0.0515) | (0.0905) | 0.46 | (0.0359) | (0.0258) | −9.49 |
| Managerial experience | 0.0457 | 0.0597 | | 0.0294 | 0.0419 | |
| | (0.0232) | (0.0478) | −0.29 | (0.0151) | (0.0316) | −0.40 |
| Sales experience | −0.0165 | 0.0364 | | 0.0132 | −0.0121 | |
| | (0.0633) | (0.0378) | −1.40 | (0.0178) | (0.0414) | 0.61 |
| Clerical experience | 0.0445 | 0.0401 | | 0.0240 | 0.0086 | |
| | (0.0234) | (0.0247) | 0.18 | (0.0185) | (0.0391) | 0.39 |
| Craftsmen experience | 0.0559 | 0.0683 | | 0.0681 | 0.0167 | |
| | (0.0082) | (0.0088) | −1.41 | (0.0103) | (0.0362) | 1.42 |
| Operatives experience | 0.0525 | 0.0584 | | 0.0304 | −0.0382 | |
| | (0.0083) | (0.0088) | −0.67 | (0.0179) | (0.0199) | 3.44 |
| Laborer experience | 0.0504 | 0.0556 | | 0.0177 | 0.0674 | |
| | (0.0085) | (0.0083) | −0.63 | (0.0219) | (0.0341) | −1.46 |
| Service experience | 0.0040 | 0.0009 | | 0.0562 | 0.0542 | |
| | (0.0158) | (0.0195) | 0.16 | (0.0066) | (0.0062) | 0.32 |
| North Central | −0.0866 | −0.0675 | | −0.2492 | −0.2297 | |
| | (0.0393) | (0.0363) | −0.53 | (0.0291) | (0.0239) | −0.81 |
| South | −0.1109 | −0.0859 | | −0.1181 | −0.0865 | |
| | (0.0408) | (0.0376) | −0.67 | (0.0304) | (0.0315) | −1.00 |
| West | −0.0043 | 0.0235 | | −0.1278 | −0.1273 | |
| | (0.0492) | (0.0524) | −0.53 | (0.0290) | (0.0307) | −0.02 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

large changes in the estimated effects of occupation-specific work experience on wages between the model that ignores classification error in occupations and the one that accounts for classification error. For example, the effect of a year of managerial experience on wages in the professional occupation is biased downward by 42 percent from 0.064 to 0.037 when misclassification is ignored. The standard error normalized difference for this parameter is -2.19, so the bias appears relatively large relative to the standard error. The bias in this particular parameter is also interesting because the estimated misclassification probabilities show that professionals are rarely misclassified as managers ($\alpha_{21}(2) = 0.0066$, $\alpha_{21}(3) = 0.0099$), and managers

**Table 5**
*Parameter Estimates—Error Standard Deviations*

Panel A

| Occupation | Ignore Classification Error | Correct for Classification Error | Normalized Difference |
|---|---|---|---|
| Professional | 0.3249 | 0.2394 | |
| | (0.0055) | (0.0069) | 3.41 |
| Managers | 0.3701 | 0.2493 | |
| | (0.0080) | (0.0163) | 1.33 |
| Sales | 0.5724 | 0.6850 | |
| | (0.0217) | (0.0248) | −2.56 |
| Clerical | 0.2763 | 0.2636 | |
| | (0.0136) | (0.0211) | 1.70 |
| Craftsmen | 0.3039 | 0.2683 | |
| | (0.0051) | (0.0068) | 6.40 |
| Operatives | 0.3317 | 0.2643 | |
| | (0.0063) | (0.0105) | 3.56 |
| Laborers | 0.3364 | 0.3411 | |
| | (0.0109) | (0.0122) | −0.20 |
| Service | 0.3250 | 0.2802 | |
| | (0.0090) | (0.0154) | 0.57 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

are rarely misclassified as professionals ($\alpha_{12}(2) = 0.0018$, $\alpha_{12}(3) = 0.0043$). The low misclassification rates between these occupations combined with the large bias in the experience coefficient illustrates the point that even a small amount of misclassification can produce large biases in estimates of the transferability of human capital across occupations.

Sales workers are the most frequently misclassified workers in both Subpopulations 2 and 3. Averaged across all three subpopulations, only 72 percent of sales workers are correctly classified. In the most common subpopulation, sales workers are most likely to be misclassified as managers ($\alpha_{23}(2) = 0.21$), so one might expect significant bias in estimates of the parameters of the managerial and sales wage equations. The estimates show that ignoring classification error causes the value of experience as a manager in the managerial occupation to be overstated by 19 percent (normalized change = 1.99). In addition, ignoring classification error leads to the misleading conclusion that one year of clerical experience increases wages by nearly 13 percent in the sales occupation, and this effect is statistically significant at the 5 percent level. However, once classification error is corrected for, the estimated effect of clerical experience on sales wages falls by 2/3, and the effect is not statistically different at conventional levels. Similarly, ignoring classification error leads to an overstatement in the value of professional experience in the sales occupation (0.0672 vs. 0.0308), although the normalized difference for this parameter is only 0.67.

**Table 5**
*Parameter Estimates—Nonpecuniary Utility*

Panel B

| | Professionals | | | Managers | | Normalized difference |
|---|---|---|---|---|---|---|
| | Ignore Classification Error | Correct for Classification Error | Normalized Difference | Ignore Classification Error | Correct for Classification Error | 1.89 |
| Age | 0.1203 | 0.0973 | | −0.0409 | −0.1448 | |
| | (0.0799) | (0.0305) | 0.75 | (0.0809) | (0.0551) | −1.40 |
| Age$^2$/100 | −0.2295 | −0.0345 | | 0.4907 | 1.0633 | |
| | (0.6142) | (0.2001) | 0.97 | (0.5784) | (0.4094) | −2.02 |
| Education | 0.4260 | 0.4715 | | 0.2145 | 0.2631 | |
| | (0.0543) | (0.0370) | −1.23 | (0.0570) | (0.0240) | 0.35 |
| High school diploma | −0.6393 | −0.3529 | | −0.2384 | −0.3125 | |
| | (0.2354) | (0.2103) | −1.36 | (0.2213) | (0.2106) | −0.99 |
| College diploma | 0.0984 | 0.3509 | | 0.2867 | 0.5197 | |
| | (0.1881) | (0.2144) | −1.18 | (0.2013) | (0.2353) | −0.46 |
| Professional experience | 0.4819 | 0.4894 | | 0.3134 | 0.3707 | |
| | (0.1484) | (0.1162) | −0.06 | (0.1468) | (0.1250) | −0.63 |
| Managerial experience | −0.0761 | −0.0229 | | 0.2605 | 0.3433 | |
| | (0.0830) | (0.1472) | −0.36 | (0.0688) | (0.1308) | −0.17 |
| Sales experience | −0.1569 | −0.1803 | | 0.0811 | 0.1052 | |
| | (0.1171) | (0.1604) | 0.15 | (0.1009) | (0.1430) | −0.75 |
| Clerical experience | −0.1028 | −0.0184 | | 0.1471 | 0.2410 | |
| | (0.1126) | (0.1637) | −0.52 | (0.0860) | (0.1249) | −1.08 |
| Craftsmen experience | 0.1531 | 0.2813 | | 0.2197 | 0.3523 | |
| | (0.0657) | (0.1271) | −1.01 | (0.0579) | (0.1224) | −1.06 |
| Operatives experience | −0.1836 | −0.1056 | | 0.0218 | 0.1383 | |
| | (0.0874) | (0.1784) | −0.44 | (0.0608) | (0.1102) | −1.39 |
| Laborer experience | −0.0459 | 0.1008 | | −0.0207 | 0.2366 | |
| | (0.1420) | (0.2052) | −0.71 | (0.1182) | (0.1849) | 0.10 |
| Service experience | −0.4737 | −0.8955 | | −0.2765 | −0.2843 | |
| | (0.0645) | (0.1467) | 2.88 | (0.0574) | (0.0820) | −1.62 |
| Previously a professional | 2.469 | 3.108 | | 1.237 | 2.022 | |
| | (0.339) | (0.368) | −1.74 | (0.379) | (0.484) | −1.47 |
| Previously a manager | 0.792 | 1.181 | | 2.780 | 3.717 | |
| | (0.340) | (0.665) | −0.59 | (0.261) | (0.636) | 0.14 |
| Previously sales | 1.194 | 0.893 | | 1.703 | 1.623 | |
| | (0.459) | (0.594) | 0.51 | (0.432) | (0.591) | −1.20 |
| Previously clerical | 1.628 | 1.546 | | 1.853 | 2.198 | |
| | (0.354) | (0.364) | 0.22 | (0.322) | (0.287) | −1.71 |
| Previously a craftsman | 1.042 | 1.064 | | 1.673 | 2.482 | |
| | (0.298) | (0.485) | −0.05 | (0.294) | (0.472) | −0.18 |
| Previously an operative | 0.752 | 0.537 | | 0.400 | 0.493 | |
| | (0.305) | (0.488) | 0.44 | (0.320) | (0.516) | −0.20 |
| Previously a laborer | 0.634 | 0.341 | | 0.839 | 0.931 | |
| | (0.346) | (0.509) | 0.58 | (0.333) | (0.471) | 1.89 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

**Table 5**
*Parameter Estimates—Nonpecuniary Utility*

Panel B

| | Sales | | | Clerical | | |
|---|---|---|---|---|---|---|
| | Ignore Classification Error | Correct for Classification Error | Normalized Difference | Ignore Classification Error | Correct for Classification Error | Normalized Difference |
| Age | −0.1327 | −0.3511 | | −0.1327 | −0.3511 | |
| | (0.1137) | (0.0484) | 4.52 | (0.1137) | (0.0484) | 1.38 |
| $Age^2/100$ | 1.3350 | 2.8694 | | 1.3350 | 2.8694 | |
| | (0.8122) | (0.4692) | −3.27 | (0.8122) | (0.4692) | −1.19 |
| Education | 0.1403 | 0.1338 | | 0.1403 | 0.1338 | |
| | (0.0764) | (0.0563) | 0.12 | (0.0764) | (0.0563) | 0.28 |
| High school diploma | −0.0762 | −0.3263 | | −0.0762 | −0.3263 | |
| | (0.3236) | (0.2981) | 0.84 | (0.3236) | (0.2981) | 0.15 |
| College diploma | 0.6676 | 0.9465 | | 0.6676 | 0.9465 | |
| | (0.2308) | (0.2761) | −1.01 | (0.2308) | (0.2761) | −1.04 |
| Professional experience | 0.0865 | 0.0444 | | 0.0865 | 0.0444 | |
| | (0.1731) | (0.1797) | 0.23 | (0.1731) | (0.1797) | 0.08 |
| Managerial experience | 0.0223 | 0.0827 | | 0.0223 | 0.0827 | |
| | (0.0903) | (0.1555) | −0.39 | (0.0903) | (0.1555) | −0.40 |
| Sales experience | 0.1072 | 0.0814 | | 0.1072 | 0.0814 | |
| | (0.1016) | (0.1502) | 0.17 | (0.1016) | (0.1502) | 0.34 |
| Clerical experience | −0.0090 | 0.0779 | | −0.0090 | 0.0779 | |
| | (0.1083) | (0.1501) | −0.58 | (0.1083) | (0.1501) | −0.53 |
| Craftsmen experience | 0.1471 | 0.3264 | | 0.1471 | 0.3264 | |
| | (0.0948) | (0.1370) | −1.31 | (0.0948) | (0.1370) | −0.70 |
| Operatives experience | 0.0325 | 0.1358 | | 0.0325 | 0.1358 | |
| | (0.0869) | (0.1214) | −0.85 | (0.0869) | (0.1214) | −1.20 |
| Laborer experience | −0.0951 | 0.0596 | | −0.0951 | 0.0596 | |
| | (0.1618) | (0.1700) | −0.91 | (0.1618) | (0.1700) | −0.87 |
| Service experience | −0.3775 | −0.4288 | | −0.3775 | −0.4288 | |
| | (0.0972) | (0.1928) | 0.27 | (0.0972) | (0.1928) | −0.25 |
| Previously a professional | 1.312 | 1.934 | | 1.312 | 0.1934 | |
| | (0.476) | (0.599) | −1.04 | (0.476) | (0.0599) | −1.11 |
| Previously a manager | 1.837 | 2.194 | | 1.837 | 2.194 | |
| | (0.393) | (0.735) | −0.49 | (0.393) | (0.735) | −0.85 |
| Previously sales | 3.262 | 2.869 | | 3.262 | 2.869 | |
| | (0.411) | (0.544) | 0.72 | (0.411) | (0.544) | 0.48 |
| Previously clerical | 2.005 | 1.864 | | 2.005 | 1.864 | |
| | (0.388) | (0.0427) | 0.33 | (0.388) | (0.427) | −0.73 |
| Previously a craftsman | 1.358 | 1.778 | | 1.358 | 1.778 | |
| | (0.407) | (0.573) | −0.73 | (0.407) | (0.573) | −0.72 |
| Previously an operative | 1.272 | 1.049 | | 1.272 | 1.049 | |
| | (0.361) | (0.457) | 0.49 | (0.361) | (0.457) | 0.37 |
| Previously a laborer | 1.358 | 1.015 | | 1.358 | 1.015 | |
| | (0.457) | (0.545) | 0.63 | (0.457) | (0.545) | 0.65 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

**Table 5**
*Parameter Estimates—Nonpecuniary Utility*

Panel B

| | Craftsmen | | | Operatives | | |
|---|---|---|---|---|---|---|
| | Ignore classification error | Correct for classification error | Normalized difference | Ignore classification error | Correct for classification error | Normalized difference |
| Age | −0.1896 | −0.2998 | | −0.1519 | −0.2535 | |
| | (0.0717) | (0.0799) | 1.82 | (0.0551) | (0.0558) | 1.87 |
| $Age^2/100$ | 1.1693 | 1.8996 | | 1.3557 | 2.0207 | |
| | (0.5598) | (0.6139) | −1.43 | (0.4459) | (0.4634) | −1.51 |
| Education | 0.1443 | 0.1253 | | −0.0703 | −0.0873 | |
| | (0.0638) | (0.0678) | 0.40 | (0.0479) | (0.0422) | 0.36 |
| High school diploma | 0.2760 | 0.2466 | | 0.1959 | 0.1931 | |
| | (0.2437) | (0.1995) | 0.02 | (0.1839) | (0.1680) | −0.09 |
| College diploma | 0.5009 | 0.7951 | | −0.4700 | −0.4137 | |
| | (0.2163) | (0.2838) | −0.16 | (0.2633) | (0.3614) | −0.05 |
| Professional experience | 0.1874 | 0.1779 | | 0.1858 | 0.1962 | |
| | (0.1529) | (0.1211) | −0.08 | (0.1581) | (0.1387) | 0.79 |
| Managerial experience | 0.0188 | 0.0697 | | −0.1568 | −0.0980 | |
| | (0.0762) | (0.1283) | −0.45 | (0.0753) | (0.1321) | 0.00 |
| Sales experience | −0.1264 | −0.1851 | | −0.2418 | −0.3401 | |
| | (0.1093) | (0.1752) | 0.54 | (0.1272) | (0.1816) | −0.14 |
| Clerical experience | 0.3591 | 0.4253 | | −0.1887 | −0.1428 | |
| | (0.0857) | (0.1258) | −0.32 | (0.0887) | (0.1439) | −0.46 |
| Craftsmen Experience | 0.1197 | 0.2104 | | 0.3067 | 0.4074 | |
| | (0.0637) | (0.1293) | −0.86 | (0.0520) | (0.1172) | −0.64 |
| Operatives experience | 0.0786 | 0.2151 | | 0.0571 | 0.1824 | |
| | (0.0707) | (0.1136) | −1.24 | (0.0552) | (0.1010) | −1.36 |
| Laborer experience | 0.0089 | 0.1485 | | 0.0430 | 0.2381 | |
| | (0.1066) | (0.1601) | −1.35 | (0.0848) | (0.1449) | −1.30 |
| Service experience | −0.3782 | −0.3587 | | −0.4665 | −0.5178 | |
| | (0.0623) | (0.0787) | 0.74 | (0.0448) | (0.0697) | 0.41 |
| Previously professional | 1.338 | 1.866 | | 0.1124 | 1.337 | |
| | (0.380) | (0.478) | −0.40 | (0.0394) | (0.526) | −0.84 |
| Previously a manager | 1.477 | 2.034 | | 0.1527 | 2.115 | |
| | (0.325) | (0.653) | −0.90 | (0.0312) | (0.651) | −0.55 |
| Previously sales | 1.710 | 1.415 | | 0.1413 | 1.059 | |
| | (0.457) | (0.618) | 0.66 | (0.0489) | (0.536) | 0.52 |
| Previously clerical | 2.804 | 2.874 | | 0.1198 | 1.166 | |
| | (0.301) | (0.097) | 0.10 | (0.0333) | (0.324) | −0.07 |
| Previously a craftsman | 1.105 | 1.462 | | 0.2903 | 3.368 | |
| | (0.307) | (0.492) | −1.26 | (0.0195) | (0.368) | −1.02 |
| Previously an operative | 0.763 | 0.609 | | 0.1521 | 1.415 | |
| | (0.280) | (0.416) | 0.36 | (0.0195) | (0.294) | 0.42 |
| Previously a laborer | 1.672 | 1.411 | | 0.1636 | 1.312 | |
| | (0.286) | (0.399) | 0.98 | (0.0231) | (0.329) | 1.06 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

**Table 5**
*Parameter Estimates—Nonpecuniary Utility*

Panel B

| | Laborers | | |
|---|---|---|---|
| | Ignore classification error | Correct for classification error | Normalized difference |
| Age | −0.2017 | −0.3403 | |
| | (0.0634) | (0.0650) | 2.13 |
| $Age^2/100$ | 1.8105 | 2.7642 | |
| | (0.5104) | (0.5240) | −1.82 |
| Education | −0.1514 | −0.1099 | |
| | (0.0613) | (0.0545) | −0.76 |
| High school diploma | 0.2912 | 0.1422 | |
| | (0.2274) | (0.2124) | 0.70 |
| College diploma | 0.0821 | 0.3030 | |
| | (0.3341) | (0.3584) | −0.62 |
| Professional experience | −0.4791 | −0.3477 | |
| | (0.2656) | (0.4226) | −0.31 |
| Managerial experience | −0.2364 | −0.3256 | |
| | (0.1162) | (0.1955) | 0.46 |
| Sales experience | −0.2337 | −0.2623 | |
| | (0.1279) | (0.1937) | 0.15 |
| Clerical experience | 0.0255 | 0.0713 | |
| | (0.0883) | (0.1468) | −0.31 |
| Craftsmen experience | 0.0943 | 0.1882 | |
| | (0.0594) | (0.1207) | −0.78 |
| Operatives experience | 0.0370 | 0.1673 | |
| | (0.0575) | (0.1032) | −1.26 |
| Laborer experience | 0.3250 | 0.4753 | |
| | (0.0910) | (0.1501) | −1.00 |
| Service experience | −0.4093 | −0.4625 | |
| | (0.0654) | (0.0884) | 0.60 |
| Previously a professional | 0.943 | 0.175 | |
| | (0.484) | (0.695) | −0.33 |
| Previously a manager | 0.609 | 0.129 | |
| | (0.400) | (0.776) | −0.67 |
| Previously sales | 0.604 | 0.754 | |
| | (0.699) | (0.707) | −0.21 |
| Previously clerical | 1.310 | 1.322 | |
| | (0.354) | (0.351) | −0.04 |
| Previously a craftsman | 1.525 | 1.832 | |
| | (0.240) | (0.435) | −0.70 |
| Previously an operative | 1.139 | 0.976 | |
| | (0.204) | (0.311) | 0.53 |
| Previously a laborer | 1.870 | 1.579 | |
| | (0.213) | (0.331) | 0.88 |

Note: Standard errors in parentheses. Normalized difference = [β(ignore class. error)- β(correct for class. error)]/[Standard error of β(correct for class. error)].

**Table 5**
*Parameter Estimates—Unobserved Heterogeneity: Classification Error Model*

Panel C

| | Type 1 | | Type 2 | | Type 3 | |
|---|---|---|---|---|---|---|
| | Parameter | Standard error | Parameter | Standard error | Parameter | Standard error |
| Nonpecuniary Intercepts | | | | | | |
| Professional | −4.7210 | 0.3270 | −4.1600 | 0.2920 | −2.8250 | 0.3810 |
| Managers | −3.1880 | 0.0930 | −3.0920 | 0.1770 | −2.2050 | 0.2510 |
| Sales | −6.1960 | 0.4940 | −0.9120 | 0.3780 | 0.0160 | 0.3850 |
| Clerical | −1.7920 | 0.3340 | −1.7200 | 0.3460 | −0.5640 | 0.3520 |
| Craftsmen | −0.1250 | 0.2410 | −0.0660 | 0.2260 | 0.5370 | 0.3130 |
| Operatives | 0.0310 | 0.2470 | 0.0570 | 0.2310 | 0.6560 | 0.3100 |
| Laborers | 0.3220 | 0.2590 | 0.4030 | 0.2180 | 1.2000 | 0.3180 |
| Wage intercepts | | | | | | |
| Professional | 1.9360 | 0.0220 | 1.1810 | 0.0250 | 1.6380 | 0.0220 |
| Managers | 1.4510 | 0.0350 | 1.0740 | 0.0260 | 1.5990 | 0.0360 |
| Sales | 2.3700 | 0.2600 | −0.2990 | 0.1770 | 0.2740 | 0.1850 |
| Clerical | 1.4400 | 0.0380 | 1.1220 | 0.0450 | 1.5480 | 0.0300 |
| Craftsmen | 1.6460 | 0.0260 | 1.3670 | 0.0250 | 1.9630 | 0.0300 |
| Operatives | 1.6220 | 0.0240 | 1.3810 | 0.0230 | 1.9710 | 0.0260 |
| Laborers | 1.4130 | 0.0480 | 1.3000 | 0.0470 | 1.7150 | 0.0420 |
| Service | 1.5020 | 0.0310 | 1.0620 | 0.0240 | 0.0010 | 0.1240 |
| Type probabilities | | | | | | |
| Pr(Type 1) | 0.1216 | 0.032 | | | | |
| Pr(Type 2) | 0.3675 | 0.041 | | | | |
| Pr(Type 3) | 0.5109 | 0.042 | | | | |

Further evidence of large changes in estimates of the transferability of human capital across occupations is found in the craftsman occupation. The model that does not correct for classification error implies that a year of professional experience increases a craftsman's wages by 2.9 percent, and this effect is statistically significant at the 5 percent level. Once classification error is accounted for this effect falls to 1.8 percent and it is not statistically different from zero at the 5 percent level. This finding suggests that the type of skills accumulated during employment as a professional have little or no value in craftsman jobs. It appears that the false transitions created by classification error lead to an overstatement of the transferability of human capital between the professional occupation and this seemingly unrelated lower skill occupation.

Another way of comparing the wage equations in the baseline and measurement error model is to determine the number of hypothesis tests where the results of the test change between the baseline and classification error models. For example, one hypothesis that is commonly of interest is the null hypothesis that the effect of each individual explanatory variable on wages equals zero. Comparing the results of these hypothesis tests for the baseline model and the classification error model shows that the rejection or acceptance of the null hypothesis at the 5 percent level

**Table 5**
*Parameter Estimates—Unobserved Heterogeneity: Model that Ignores Classification Error*

Panel C

| | Type 1 | | Type 2 | | Type 3 | |
|---|---|---|---|---|---|---|
| | Parameter | Standard error | Parameter | Standard error | Parameter | Standard error |
| Nonpecuniary intercepts | | | | | | |
| Professional | −3.6890 | 0.3330 | −3.4730 | 0.3160 | −2.1610 | 0.3520 |
| Managers | −2.4600 | 0.3300 | −2.5340 | 0.3060 | −1.5880 | 0.3640 |
| Sales | −7.2570 | 0.7340 | −2.0600 | 0.4340 | −1.0310 | 0.4350 |
| Clerical | −1.8030 | 0.2820 | −2.0260 | 0.2860 | −0.9600 | 0.3590 |
| Craftsmen | −0.1680 | 0.2170 | −0.3450 | 0.2110 | 0.5080 | 0.2910 |
| Operatives | −0.1820 | 0.2210 | −0.1820 | 0.2180 | 0.5370 | 0.2930 |
| Laborers | −0.0110 | 0.2560 | −0.0090 | 0.2420 | 0.6280 | 0.3030 |
| Wage intercepts | | | | | | |
| Professional | 1.7720 | 0.0630 | 1.0550 | 0.0610 | 1.5460 | 0.0600 |
| Managers | 1.3740 | 0.0750 | 0.9420 | 0.0720 | 1.4420 | 0.0720 |
| Sales | 1.8580 | 0.1800 | −0.0220 | 0.1420 | 0.4980 | 0.1390 |
| Clerical | 1.4640 | 0.0470 | 1.1000 | 0.0510 | 1.5630 | 0.0490 |
| Craftsmen | 1.5540 | 0.0320 | 1.2910 | 0.0300 | 1.8530 | 0.0340 |
| Operatives | 1.5590 | 0.0380 | 1.3020 | 0.0360 | 1.7940 | 0.0360 |
| Laborers | 1.4670 | 0.0570 | 1.2880 | 0.0550 | 1.7770 | 0.0600 |
| Service | 1.4630 | 0.0520 | 1.0170 | 0.0480 | 1.3190 | 0.0690 |
| Type probabilities | | | | | | |
| Pr(Type 1) | 0.0456 | 0.033 | | | | |
| Pr(Type 2) | 0.5030 | 0.039 | | | | |
| Pr(Type 3) | 0.4514 | 0.040 | | | | |

changes for 17 variables in the wage equation between the two models. In other words, ignoring classification error would cause one to mistakenly accept or reject the null hypothesis that the effect of an explanatory variable equals zero for 17 wage equation variables.

The final parameters of the wage equation are the standard deviations of the random shock to wages in each occupation, $\sigma_{eq}$, for $q = 1, \ldots, 8$. The estimates of these standard deviations show that random fluctuations in wages are overstated in six out of the eight occupations in the model that ignores classification error. The intuition behind the direction of this bias is that the model must provide an explanation for the large number of short duration occupation switches that occur in the data. When classification error is ignored, the model accomplishes this through relatively large transitory wage shocks.

The determinants of occupational choices have been the subject of considerable research interest, and several recent papers have examined the related question of the role of occupation-specific human capital in determining wages. Although labor economists have typically focused on determining the roles of firm tenure and

general work experience in determining wages, new evidence suggests that in fact occupation-specific skills play an important role in determining wages.[14] Comparing the estimates of the wage equation found in this paper to existing estimates is difficult for a number of reasons. First, there is no existing paper that estimates directly comparable occupation-specific wage equations at the one-digit level. Second, existing papers that estimate wage equations that are similar in some respects do not allow for the type of cross-occupation experience effects found in this study.[15] However, overall the wage equation estimates appear to be broadly consistent with existing research in this area. For example, Both Kambourov and Manovskii (2007) and Sullivan (2007) find that while experience in a workers' current occupation has as an important effect on wages, wages are strongly impacted by total work experience. This finding is consistent with the relatively large cross-occupation experience effects reported in this paper. Keane and Wolpin (1997) also find relatively large cross-occupation experience effects between blue collar and white collar employment, which is again broadly consistent with the wage equation estimates reported in this paper. It is also possible to get a rough sense of how the magnitudes of the estimated effects of occupation-specific work experience on wages in this paper compare to existing research. The estimates in this paper suggest that when classification error is ignored, averaged across all occupations one year of occupation-specific work experience increases wages by approximately 7 percent. Kambourov and Manovskii (2007) do not report a parameter estimate that is directly comparable to this number, but combining the different parameter estimates that they report suggest that wages grow by approximately 5 percent to 8 percent with each year that a worker spends in an occupation.

### *2. Nonpecuniary Utility Flows & Unobserved Heterogeneity*

The occupational choice model presented in this paper allows occupational choices to depend on nonpecuniary utility flows as well as wages. The importance of modeling occupational choices in a utility maximizing framework rather than in an income maximizing framework is demonstrated in work by Keane and Wolpin (1997) and Gould (2002). The parameter estimates for the nonpecuniary utility flow equations for the models estimated with and without accounting for classification error are presented in Panel B of Table 5. These results show that ignoring classification error leads to significant biases in estimates of the effects of variables such as age, education, and work experience on occupational choices.

The nonpecuniary utility flow parameters are all measured in log-wage units relative to the base choice of service employment. For example, the estimate of the effect of working as a professional in the previous time period on the professional utility flow is 2.469 in the model that ignores classification error. This means that

14. See, for example, Kambourov and Manovskii (2007) and Sullivan (2007).

15. Kambourov and Manovskii (2007) and Sullivan (2007) consider the special case of the wage equation estimated in this paper where all of the cross-occupation experience effects are equal. However, these studies also consider firm tenure and industry specific work experience. Keane and Wolpin (1997) allow for cross-occupation experience effects, but their work uses occupation codes aggregated to the level of blue- and white-collar jobs.

a person who previously worked as a professional receives utility that is 2.469 log wage units higher than a person who was previously employed as a service worker but is currently employed as a professional. The effect of previous professional employment on the professional utility flow is biased downwards by 21 percent when classification error is ignored (normalized difference = -1.74). It appears that the false transitions between occupations created by classification error lead to an understatement of the importance of state dependence in professional employment. Overall, the estimates of the effects of lagged occupational choices on current occupation-specific utility flows are fairly sensitive to classification error.

As is the case with the wage equation, another way of examining the consequences of not correcting for misclassification is to determine the number of hypothesis tests where the results of the test at the 5 percent level change between the baseline and classification error models. Comparing the results of these hypothesis tests for the baseline model and the classification error model show that the rejection or acceptance of the null hypothesis that the effect of each variable equals zero changes for 22 variables in the nonpecuniary utility flow equation between the two models. In other words, ignoring classification error would cause one to mistakenly accept or reject the null hypothesis that the effect of an explanatory variable on nonpecuniary utility equals zero for 22 variables.

The estimates of the wage intercepts ($\mu$'s) and nonpecuniary intercepts ($\phi$'s) for the three types of people in the model are presented in Panel C of Table 5. These parameter estimates reveal the extent of unobserved heterogeneity in skills and preferences for employment in each occupation. The final section of Panel C of Table 5 shows the averages of the wage and nonpecuniary intercepts across the three types of people for the models that correct for and ignore classification error in occupation codes. The largest bias among these parameters occurs in parameters that measure preferences for employment in each occupation ($\phi$'s). For example, the average preference for working as a craftsman changes from 0.048 in the model that ignores classification error to 0.23 in the model that corrects for classification error. Biases of similar magnitudes are found in the average preferences for employment as operatives and laborers. The relatively large biases in estimates of preference parameters caused by ignoring classification error occurs because unobserved heterogeneity in preferences helps explain occupational transitions that are not well explained by the other parts of the model. When classification error is ignored and all occupational transitions are treated as true occupation switches, the model attempts to explain transitions that are not well explained by wages or the deterministic portion of nonpecuniary utility flows in part through preference heterogeneity.

## V. Simulating Data that is Free from Misclassification

One application of the model presented in this paper is that the estimated model can be used to simulate occupational choice data that is free from classification error. The simulated data is used to examine which workers tend to be identified as misclassified by the model, the predicted patterns in misclassification over workers' careers, and the predicted relationship between wages and misclassification.

### *A. Simulated Occupational Choices*

#### *1. Which Workers are Misclassified?*

One explanatory variable that is of central importance when investigating occupational choices is education, since there is strong sorting across occupations based on completed education. Given this fact, it is useful to see how completed education levels vary between choices that are identified as misclassified choices in the simulated data compared to choices that are identified as correctly classified choices.

Table 6 shows the distribution of completed education for correctly classified and misclassified occupational choices, disaggregated by occupation. For example, the table shows that the model predicts that 10.8 percent of those workers who are

**Table 6**
*Completed Education by Observed Occupation for Correctly Classified and Misclassified Occupational Choices*

| Observed Occupation in NLSY Data | | Percent No College Completed | Percent College Graduate |
|---|---|---|---|
| Professional | Correctly classified | 10.8% | 71.8% |
| | Misclassified | 48.6% | 30.2% |
| Managers | Correctly classified | 39.8% | 36.8% |
| | Misclassified | 47.2% | 28.7% |
| Sales | Correctly classified | 25.2% | 54.2% |
| | Misclassified | 44.8% | 24.7% |
| Clerical | Correctly classified | 54.9% | 23.7% |
| | Misclassified | 36.0% | 49.0% |
| Craftsmen | Correctly classified | 77.9% | 2.1% |
| | Misclassified | 53.3% | 18.7% |
| Operatives | Correctly classified | 85.2% | 2.5% |
| | Misclassified | 61.3% | 21.5% |
| Laborers | Correctly classified | 83.7% | 3.2% |
| | Misclassified | 73.0% | 11.7% |
| Service | Correctly classified | 60.2% | 13.7% |
| | Misclassified | 74.2% | 8.1% |

Notes: Generated using the simulated data that identifies occupational choices as correctly or incorrectly classified. The ''correctly classified'' row refers to observations where the occupation in the leftmost column matches the true occupation code generated by the model. The ''misclassified'' row refers to observations where a person is observed in the occupation in the leftmost column and the simulated true occupation differs from the observed occupation. So, 71.8 percent of correctly classified professionals graduated from college, while only 30.2 percent of those incorrectly classified as professionals graduated from college.

correctly classified as professionals have not completed any years of college, while 48.6 percent of workers who are misclassified as professionals have not completed any years of college. A correctly classified professional has a 71.8 percent change of being a college graduate, while a worker misclassified as a professional has only a 30.2 percent chance of being a college graduate. Clearly, education serves as a strong predictor of which observations are likely to be true professionals as opposed to observations that are falsely classified as professionals. These results are consistent with the fact that the jobs located in the professional occupation are overwhelmingly ones that require a college degree, or at least some amount of completed higher education.

Across the other occupations, similarly strong and sensible relationships exist between education and misclassification. For example, in blue collar occupations, one would expect to see the opposite relationship between misclassification and education from the one found in the professional occupations. This is in fact what the results in Table 6 show. For example, the percentage of correctly classified workers who have graduated from college is 2.1 percent for craftsmen, 2.5 percent for operatives, and 3.2 percent for laborers. In contrast, for workers who are falsely classified in these occupations the percentage of workers who are college graduates is 18.7 percent for craftsmen, 21.5 percent for operatives, and 11.7 percent for laborers. In general, the model tends to flag workers as misclassified who have reported education levels that appear to be inconsistent with their reported occupation.

### *2. The Frequency of Misclassification over an Individual's Career*

Given the panel nature of the data, the simulated occupational choice data can be used to examine how often occupational choices are misclassified over a typical individual's career. Table 7 presents the distribution of the total number of times that occupational choices are misclassified over the course of a person's career. The majority of workers never experience misclassification (57.2 percent), 17.6 percent of workers are misclassified once over their career, and very few workers are misclassified more than five times over their career (4.3 percent). Table 7 also provides information about the distribution of the lengths of misclassification spells. For example, the first entry in the final column of Table 7 shows that conditional on an occupational choice being misclassified, there is a 72.9 percent chance that the person will be *correctly* classified in the next survey. Conditional on being misclassified, there is an 18.3 percent chance that a person will be misclassified in two consecutive periods, and there is only a 5.2 percent chance that a person will be misclassified in three consecutive periods.[16]

### *3. True Occupational Choices, Observed Choices, and Wages*

Table 8 shows the average true occupational choice probabilities conditional on observed choices and observed wages that are predicted by the empirical model. This

16. One implication of the relatively short durations of misclassification spells is that the model does not tend to repeatedly flag individuals as misclassified who have consistently high (or low) wages for their reported occupation over the course of their entire career.

**Table 7**
*Distribution of Total Number of Times a Person's Occupational Choices are Misclassified Over the Career and Length of Misclassification Spells*

| Total Number of Times Misclassified | Percentage | Number of Consecutive Times Misclassified | Percentage |
|---|---|---|---|
| 0 | 57.2% | 1 | 72.9% |
| 1 | 17.6% | 2 | 18.3% |
| 2 | 11.5% | 3 | 5.2% |
| 3 | 6.3% | 4 | 1.8% |
| 4 | 3.0% | 5 | 0.7% |
| 5 | 2.2% | >5 | 0.93% |
| 6-9 | 1.9% | | |
| >9 | 0.20% | | |

Entries are the distributions of the number of times that a person's occupational choices are misclassified over the course of the career and lengths misclassification spells based on the simulated data.

analysis shows how the classification error rates generated by the model vary with observed wages and provides a more detailed analysis of the type of occupational choice and wage combinations that are likely to be affected by classification error.

Observed occupational choices are listed in the far left column of Table 8, while actual occupational choices are listed in the top row. Conditional on the observed choice and wage (and all of the other explanatory variables), the model is used to calculate the conditional probability that the actual choice is each of the eight occupations for each occupational choice observed in the data. The average of each probability for each occupation is presented in Table 8. Probabilities are disaggregated by the percentile of the observed wage in the wage distribution of the observed occupation to show how misclassification rates vary with observed wages. For example, the top left cell of Table 8 shows that a worker observed in the data as a professional worker with a wage in the top 10 percent of the professional wage distribution has a 90.9 percent chance of being correctly classified as a professional worker. However, a worker observed as a professional with a wage in the bottom 10 percent of the professional distribution has only a 75.7 percent chance of actually being a professional worker. People observed in the data as low wage professional workers are primarily service workers (9.5 percent).

Similar patterns of misclassification are found in the sales and clerical occupations, where workers in certain areas of the wage distribution are more likely to be misclassified than those in other areas of the wage distribution. For example, 91.6 percent of clerical workers in the top 10 percent of the clerical wage distribution are correctly classified, but 3.9 percent of those observed as high wage clerical workers are actually professionals. However, the unconditional probability that a professional is misclassified as a clerical worker is much lower ($\alpha_{41}(2) = 0.013$, $\alpha_{41}(3) = 0.013$).

**Table 8**
*Average True Choice Probabilities by Observed Choice and Wage Percentile*

| Observed/Actual | | Professional | Managers | Sales | Clerical | Craftsmen | Operatives | Laborers | Service |
|---|---|---|---|---|---|---|---|---|---|
| Professional | Top 10% | 0.909 | 0.000 | 0.074 | 0.000 | 0.000 | 0.000 | 0.004 | 0.011 |
| | Middle 10% | 0.953 | 0.001 | 0.006 | 0.000 | 0.013 | 0.000 | 0.012 | 0.014 |
| | Bottom 10% | 0.757 | 0.001 | 0.053 | 0.001 | 0.020 | 0.001 | 0.070 | 0.095 |
| Managers | Top 10% | 0.052 | 0.565 | 0.374 | 0.001 | 0.002 | 0.000 | 0.000 | 0.005 |
| | Middle 10% | 0.020 | 0.858 | 0.067 | 0.002 | 0.003 | 0.002 | 0.001 | 0.046 |
| | Bottom 10% | 0.010 | 0.544 | 0.272 | 0.004 | 0.004 | 0.005 | 0.012 | 0.148 |
| Sales | Top 10% | 0.039 | 0.033 | 0.916 | 0.017 | 0.000 | 0.000 | 0.000 | 0.038 |
| | Middle 10% | 0.033 | 0.018 | 0.911 | 0.000 | 0.000 | 0.002 | 0.002 | 0.026 |
| | Bottom 10% | 0.004 | 0.005 | 0.834 | 0.016 | 0.000 | 0.007 | 0.007 | 0.127 |
| Clerical | Top 10% | 0.039 | 0.005 | 0.916 | 0.001 | 0.000 | 0.000 | 0.000 | 0.038 |
| | Middle 10% | 0.033 | 0.017 | 0.911 | 0.008 | 0.000 | 0.002 | 0.002 | 0.026 |
| | Bottom 10% | 0.004 | 0.005 | 0.834 | 0.016 | 0.000 | 0.007 | 0.007 | 0.127 |
| Craftsmen | Top 10% | 0.031 | 0.001 | 0.091 | 0.000 | 0.872 | 0.000 | 0.003 | 0.000 |
| | Middle 10% | 0.008 | 0.000 | 0.015 | 0.000 | 0.965 | 0.002 | 0.007 | 0.000 |
| | Bottom 10% | 0.005 | 0.000 | 0.124 | 0.003 | 0.818 | 0.005 | 0.041 | 0.002 |
| Operatives | Top 10% | 0.084 | 0.000 | 0.110 | 0.000 | 0.000 | 0.801 | 0.003 | 0.000 |
| | Middle 10% | 0.009 | 0.000 | 0.008 | 0.000 | 0.000 | 0.979 | 0.003 | 0.000 |
| | Bottom 10% | 0.003 | 0.000 | 0.119 | 0.000 | 0.000 | 0.869 | 0.006 | 0.000 |
| Laborers | Top 10% | 0.000 | 0.000 | 0.065 | 0.012 | 0.032 | 0.002 | 0.885 | 0.003 |
| | Middle 10% | 0.000 | 0.000 | 0.003 | 0.007 | 0.008 | 0.003 | 0.976 | 0.001 |
| | Bottom 10% | 0.000 | 0.000 | 0.071 | 0.004 | 0.003 | 0.002 | 0.915 | 0.004 |
| Service | Top 10% | 0.054 | 0.004 | 0.072 | 0.000 | 0.000 | 0.000 | 0.150 | 0.719 |
| | Middle 10% | 0.005 | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 | 0.251 | 0.732 |
| | Bottom 10% | 0.000 | 0.000 | 0.100 | 0.000 | 0.000 | 0.000 | 0.174 | 0.725 |

Note: Entries are the average true choice probabilities found in the simulated data conditional on the observed choice and wage. Top, middle, and bottom 10 percent refer to the location of the observed wage in the wage distribution of the observed occupation.

### *B. Sensitivity Analysis: Measurement Error in Wages*

One important question regarding the model presented in this paper is the sensitivity of the results to the existence of measurement error in wages. One way of addressing this question is to simulate noisy wage data, reestimate the model using the noisy wage data (leaving the rest of the NLSY data unchanged), and see how the estimates of misclassification parameters change when the noisy wage data is used in place of the actual wages found in the NLSY data. The noisy wages ($w_{it}^{me}$) are generated using the following equation,

$$(18) \quad w_{it}^{me} = w_{it}^{obs} + \nu_{it}, \text{ where } \nu_{it} \sim N(0, \sigma_{\nu}^2).$$

Recall that $w_{it}^{obs}$ is a log wage, so the extent of measurement error in the noisy log wage data is captured by $\sigma_{\nu}^2$. A number of validation studies have quantified the extent of measurement error in wages, see Bound, Brown, and Mathiowetz (2001) for a thorough survey of this literature. Actual estimates of $\sigma_{\nu}^2$ do not exist for the NLSY, so in simulating the noisy data the measurement error term is set towards the upper end of the reported estimates found in the literature based on other data sources. The exact value used is $\sigma_{\nu}^2 = .10$. This value of $\sigma_{\nu}^2$ creates a substantial amount of measurement error in the noisy wage data, since in the noisy data, measurement error accounts for approximately one-third of the total variation in log wages.

Rather than presenting a complete set of parameter estimates for the misclassification model estimated using the noisy data, it is sufficient to summarize the overall effect that the noisy wage data has on the parameter estimates. When the noisy wage data are used in place of the NLSY wage data, the average parameter in the model changes by approximately 2 percent, so it appears that the overall bias introduced by measurement error is relatively small. The primary concern about measurement error in wages is that it may impact the estimates of the extent of measurement error in occupation codes. The overall extent of misclassification is summarized by the diagonal elements of the misclassification rate matrices for Subpopulations 2 and 3, $\alpha_{jj}(y)$, for $j = 1, \ldots, Q$, and $y = 2, 3$. Across both subpopulations, the use of noisy wage data results in the average probability of correct classification decreasing by only $-0.006$ from $-0.8546$–$0.8486$. Adding measurement error slightly increases the overall estimated rate of misclassification, but the magnitude of the increase is quite small. The corresponding average absolute change in the probability of correct classification is only 0.008, and the average change in the off-diagonal elements is only 0.0015, so it appears that estimates of the overall extent of misclassification in the NLSY occupation data are quite robust to measurement error in wages.

There are a number of reasons why the estimates of the misclassification parameters are robust to a relatively large amount of measurement error in wages. The first reason is that, as discussed earlier in the paper, wages are not the only source of information that the model uses to infer that an occupational choice is misclassified. Another key point is that many of the occupational choices that are flagged in the simulations as misclassifications are associated with extremely large differences between the reported wage and the average wage in the reported occupation.

Differences of this magnitude are unlikely to be generated in large numbers by a reasonable amount of measurement error in wages. For example, the median wage for workers who are identified in the simulations as falsely classified professionals is $5.59, while the median wage for workers who are correctly classified as professionals is $10.32.

## VI. Conclusion

Although occupational choices have been a topic of considerable research interest, existing research has not studied occupational choices in a framework that addresses the biases created by classification error in self-reported occupation data. This paper develops an approach to estimating a panel data occupational choice model that corrects for classification error in occupations by incorporating a model of misclassification within an occupational choice model. Estimating this model provides a solution to the problems created by measurement error in the discrete dependant variable of an occupational choice model. Methodologically, this approach contributes to the literature on misclassification in discrete dependant variables by demonstrating how simulation methods can be used to address the problems created in a panel data setting where measurement error in a discrete dependant variable creates measurement error in explanatory variables. The simulation technique is applicable to any discrete choice panel data model where misclassification in a current period dependent variable creates measurement error in future explanatory variables. This paper also contributes to the literature on misclassification by using observed wages within the framework of an occupational choice model to obtain information about misclassified occupational choices.

The main findings of this paper are that a substantial number of occupational choices in the NLSY are affected by misclassification, with an overall misclassification rate of 9 percent. The results also suggest that person-specific heterogeneity in misclassification rates is an important feature of the data. An estimated 38 percent of the population never experiences a misclassified occupational choice, and the remaining two subpopulations have substantially different propensities to have their occupational choices misclassified in particular directions. The parameter estimates also indicate that misclassification rates vary widely across occupations, and that the probability of a worker being misclassified into each occupation is strongly influenced by the worker's actual occupation. Most importantly, this paper demonstrates the large bias in parameter estimates that results from estimating a model of occupational choices that ignores the fact that occupations are frequently misclassified. Consistent with existing research in the area of misclassified dependant variables, the results show that even relatively small amounts of misclassification creates substantial bias in parameter estimates. Especially large biases are found in parameters that measure the transferability of occupation-specific work experience across occupations, since these parameters are quite sensitive to the false occupational transitions created by classification error.

Overall, the results indicate that one should use caution when interpreting the parameter estimates from occupational choice models that are estimated without correcting for classification error in self-reported occupations. In addition, these

results suggest that similar bias may arise when occupation dummy variables are used as explanatory variables, as is commonly done in a wide range of studies. A possible avenue for future research would be to investigate the effects of classification error in occupation codes on parameter estimates in this wider class of models, such as simple wage regressions that make use of self-reported occupation data.

## APPENDIX 1

## Data

The goal of this paper is to follow workers from the time they make a permanent transition to the labor market and start their career. There is no clear best way to identify this transition to the labor market, so this analysis follows people from the month they reach age 18 or stop attending school, whichever occurs later. Individuals are followed until they reach age 35, or exit from the sample due to missing data. There are 6,111 men in the nationally representative cross-sectional sample of the NLSY. Of these workers, there are 2,439 white males who are candidates to be used in this analysis. As noted in the text, respondents are between the ages of 14 and 21 in the first year of the NLSY. One issue raised by the fact that respondents enter the sample at different ages is that there is an initial conditions problem for the older workers because the data does not contain any information on their work history before they enter the sample. For example, if someone enters the sample at age 21 they may have never worked before, or they could have accumulated a number of years of experience in a particular occupation.

Based on these considerations, individuals who enter the NLSY sample at an age older than 18 are dropped from the sample, so that each individual enters the sample at the start of his career. Since individuals are nearly evenly distributed across initial ages, approximately 43 percent of the sample is dropped as a result of the age restriction. In addition, individuals are dropped if they serve in the military at some point during their career, if they ever work as farmers, if they report being self-employed, if it is not possible to determine when they stopped attending school and started their career, or if information on completed schooling is missing. Finally, observations are truncated if geographic data, wage data, or occupation data are missing, or if the individual exits from the sample. The final sample contains 954 individuals and 10,573 observations.

Since the estimation sample is considerable smaller that the entire NLSY sample due to the age restriction, one might wonder how this sample compares to a broader sample from the NLSY. Table 2 shows descriptive statistics for the estimation sample and a broader sample of white men from the NLSY cross-sectional sample that includes roughly twice as many individuals because it includes individuals who enter the NLSY sample at ages older than 18. Overall, the samples appear to be quite similar in terms of observable characteristics. As an additional check, Equation 19 shows the estimates of a simple regression of wages on age, education, and occupation dummy variables for the estimation subsample, and equation number 20 shows the estimates based on the broader sample (standard errors in parentheses),

$$\begin{aligned}(19)\quad \ln(w) = {} & 1.16 + 0.035age + 0.054educ - 0.062manag - 0.129sales) \\ & (0.034)(0.001)(0.003)(0.018)(0.023)) \\ & - 0.173cleric - 0.046craft - 0.175oper - 0.222labor - 0.331serv) \\ & (0.021)(0.017)(0.018)(0.021)(0.020))\end{aligned}$$

$$\begin{aligned}(20)\quad \ln(w) = {} & 0.829 + 0.026age + 0.062educ - 0.039manag - 0.097sales) \\ & (0.035)(0.001)(0.002)(0.014)(0.016)) \\ & - 0.190cleric - 0.037craft - 0.142oper - 0.220labor - 0.360serv.) \\ & (0.016)(0.013)(0.013)(0.015)(0.015))\end{aligned}$$

## APPENDIX 2

## Simulating the Likelihood Function

### *The Likelihood Function*

Let $Exp_{it}^{A}$ represent person $i$'s actual, or true experience in occupation $q$ in time period $t$. Define $Exp^{A}$ as a $Q \times 1$ vector of experience in each occupation. Let $Lastocc_{it}^{A}$ represent a $Q \times 1$ vector of dummy variables where the $q$th element is equal to one if person $i$'s true occupational choice was $q$ in time period $t-1$. Let $F_{it}(Exp_{it}^{A}, Lastocc_{it}^{A})$ represent the distribution of true occupation-specific experience and lagged occupational choices for person $i$ in time period $t$. This distribution is a function of each person's observed characteristics, and observed choices and wages, but these conditioning variables are suppressed for brevity of notation. The likelihood function can be evaluated by integrating over the distribution of the unobserved state variables,

$$(21)\quad L(\theta) = \prod_{i=1}^{N} \int \prod_{t=1}^{T(i)} L_{it}(\theta | Exp_{it}^{A}, Lastocc_{it}^{A}) dF_{it}(Exp_{it}^{A}, Lastocc_{it}^{A})$$

However, in practice this is very difficult to do because the distribution $F_{it}(Exp_{it}^{A}, Lastocc_{it}^{A})$ is intractably complex. This problem can be overcome by simulated the likelihood function using a recursive simulation algorithm that is similar to the Geweke (1991), Hajivassiliou (1990), and Keane (1994) (GHK) algorithm. The GHK algorithm breaks a choice probability up into a sequence of transition probabilities, and then recursively simulates the sequence. Simulation methods have not been used extensively in this manner to solve problems created by measurement error, although it is a natural application of these techniques.

### *The Simulation Algorithm*

This section provides the details of the simulation algorithm used to evaluate the likelihood function. For simplicity, the algorithm is outlined for the case where the number of unobserved heterogeneity types *(M)* equals one because the extension to multiple types is straightforward. The object that must be simulated is

$$(21)\quad L(\theta) = \prod_{i=1}^{N} \int \prod_{t=1}^{T(i)} \sum_{q=1}^{Q} 1\{O_{it} = q\} P_{it}(q, w_{it}^{obs} | \theta, Z_{it}, X_{it}, Exp_{it}^{A}, Lastocc_{it}^{A}) \, dF_{it}(Exp_{it}^{A}, Lastocc_{it}^{A})$$
$$= \prod_{i=1}^{N} \int \prod_{t=1}^{T(i)} L_{it}(O_{it}, w_{it}^{obs} | \theta, Z_{it}, X_{it}, Exp_{it}^{A}, Lastocc_{it}^{A}) \, dF_{it}(Exp_{it}^{A}, Lastocc_{it}^{A})$$

Let variables with a * superscript represent simulated variables, and let $r = 1, \ldots, R$ index simulation draws. Using this notation, $O_{it}^{*}(r|\theta, O_{it}, w_{it}^{obs}, Z_{it,}X_{it}, Exp_{it}^{*}, Lastocc_{it}^{*})$ is a simulated occupational choice, $Exp_{it+1}^{*}(r|\theta, O_{it}, w_{it}^{obs}, Z_{it,}X_{it}, Exp_{it}^{*}, Lastocc_{it}^{*})$ is a $Q \times 1$ vector of simulated occupation-specific experience, $Lastocc_{it+1}^{*}(r|\theta, O_{it}, w_{it}^{obs}, Z_{it,}X_{it}, Exp_{it}^{*}, Lastocc_{it}^{*})$ is a vector of dummy variables representing the simulated occupational choice in the previous period, and $L_{it}^{*}(r, O_{it}, w_{it}^{obs}|\theta, Z_{it,}X_{it}, Exp_{it}^{*}, Lastocc_{it}^{*})$ is a simulated likelihood contribution. For brevity of notation, define the set of conditioning variables for the simulated choices as $\rho = \{\theta, O_{it}, w_{it}^{obs}, Z_{it,}X_{it}, Exp_{it}^{*}, Lastocc_{it}^{*}\}$. The simulation algorithm for Person $i$ is:

(1) Start in Time Period $t = 1$, simulation draw $r = 1$. All experience variables equal zero at the start of the career by definition, so initialize the simulated experience vector to zero for Time Periods $t = 1, \ldots, T$.

(2) Evaluate and store the simulated likelihood contribution for year $t$, simulation draw $r$, $L_{it}^{*}(r, O_{it}, w_{it}^{obs}|\theta, Z_{it,}X_{it}, Exp_{it}^{*}(r), Lastocc_{it}^{*}(r))$.

(3) Compute and store the probability that person $i$'s true choice in time period $t$ ($\hat{O}_{it}$) is each of the $Q$ possible occupations, conditional on the parameter vector ($\theta$), observed choice ($O_{it}$), observed wage ($w_{it}^{obs}$), explanatory variables ($Z_{it,}X_{it}$), and simulated previous occupational choice ($Lastocc_{it}^{*}(r)$) and experience variables ($Exp_{it}^{*}(r)$). Let $\Omega_{it}(r, q|\rho)$ for $q = 1, \ldots, Q$ represent the conditional probability for simulation draw $r$ that the true occupational choice is $q$ for person $i$ in Time Period $t$. These probabilities can be written using Bayes' rule as

$$(22)\quad \Omega_{it}(r, q|\rho) = \Pr(\hat{O}_{it} = q|\rho) = \frac{\alpha_{O(it),1} \hat{P}_{it}(q, w_{it}^{obs})}{\sum_{k=1}^{Q} \alpha_{O(it),k} \hat{P}_{it}(k, w_{it}^{obs})}.$$

Recall that $\hat{P}_{it}(\bullet)$ is a function of all of the variables that $\Omega_{it}(\bullet)$ is conditioned on, but these arguments are suppressed here. This implies that the conditional true choice probabilities ($\Omega_{it}(\bullet)$) are a function of the observed wage and all of the explanatory variables in the model.

(4) Use the $Q$ computed conditional true choice probabilities, $\Omega_{it}(r, q|\rho)$, to define the discrete distribution of true occupational choices $\{\Pr(O_{it}^{*}(r) = q) = \Omega_{it}(r, q|\rho)\}$, $\{q = 1, \ldots, Q\}$. Next, randomly draw a simulated true occupational choice $O_{it}^{*}(r|\rho)$

for person $i$ at time period $t$ from the discrete distribution of the $Q$ possible true occupational choices.

(5) Use the simulated choice $O^*_{it}(r|\rho)$ to update the vectors of simulated experience and lagged occupational choice vectors, $Exp^*_{it+1}(r)$ and $Lastocc^*_{it+1}(r)$. The updating rules are to increase the element of the experience vector by one in the simulated occupation, and leave all other elements of the vector unchanged. For the previous occupation dummy, set the element of the $Lastocc^*_{it+1}(r)$ vector corresponding to the simulated occupation in time $t$ equal to one and set all other elements of the vector to zero.

(6) If $t = T(i)$ (the final time period for Person $i$), go to step 7. Otherwise, Set $t = t+1$ and go back to Step 2.

(7) Compute the likelihood function for simulated path $r$,

$$(23)\quad L^r_i(\theta) = \prod_{t=1}^{T(i)} L^*_{it}(r, O_{it}, w^{obs}_{it}|\theta, Z_{it,}X_{it}, Exp^*_{it}(r), Lastocc^*_{it}(r)).$$

(8) Repeat this algorithm $R$ times, and the simulated likelihood function is the average of the $R$ path probabilities over the $R$ draws,

$$(24)\quad L^*_i(\theta) = \frac{1}{R}\sum_{r=1}^{R} L^r_i(\theta).$$

During estimation, antithetic acceleration is used to reduce the variance of the simulated integrals. The number of simulation draws is set at $R = 60$. Increasing the number of simulation draws to $R = 600$ leads to only a 0.01 percent change in the value of the likelihood function at the simulated maximum likelihood parameter estimates.[17]

## References

Abowd, John, and Arnold Zellner. 1985. "Estimating Gross Labor Force Flows." *Journal of Business and Economic Statistics* 3(3):254-83.

Bollinger, Christopher. 1996. "Bounding Mean Regressions when a Binary Regressor is Mismeasured." *Journal of Econometrics* 73(2):387-99.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, ed. Edward Learner and James Heckman, 3705-3843. New York: North Holland Publishing.

Chen, Xiaohong, Han Hong, and Elie Tamer. 2005. "Measurement Error Models with Auxiliary Data." *The Review of Economic Studies* 72(2):343-66.

17. As a further check on the robustness of the parameter estimates to the choice of $R$, the model was reestimated using $R = 300$. The program converged to essentially the same parameter vector as it did when $R = 60$ was used.

Chua, Tin Chiu, and Wayne Fuller. 1987. "A Model for Multinomial Response Error Applied to Labor Flows." *Journal of the American Statistical Association* 82(397):46-51.

Dustmann, Christian, and Arthur van Soest. 2001. "Language Fluency and Earnings: Estimation with Misclassified Language Indicators." *The Review of Economics and Statistics* 83(4):663-74.

Geweke, John. 1991. "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints." In *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, ed. E. M. Keramidas, 571-78. Fairfax: Interface Foundation of North America, Inc.

Gould, Eric. 2002. "Rising Wage Inequality, Comparative Advantage, and the Growing Importance of General Skills in the United States." *Journal of Labor Economics* 20(1):105-47.

Hajivassiliou, Vassilis. 1990. "Smooth Simulation Estimation of Panel Data LDV Models." New Haven: Yale University. Unpublished.

Hausman, Jerry, Jason Abrevaya, and Fiona Scott-Morton. 1998. "Misclassification of the Dependant Variable in a Discrete-Response Setting." *Journal of Econometrics* 87(2):239-69.

Heckman, James, and Guilherme Sedlacek. 1985. "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self Selection in the Labor Market." *Journal of Political Economy* 93(6):1077-1125.

_____. 1990. "Self-Selection and the Distribution of Hourly Wages." *Journal of Labor Economics* 8(1): S329-S363.

Heckman, James, and Burton Singer. 1984. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica* 52(2):271-320.

Kambourov, Gueorgui, and Iorii Manovskii. 2007. "Occupational Specificity of Human Capital." *International Economic Review.* Forthcoming.

Keane, Michael. 1994. "A Computationally Practical Simulation Estimator for Panel Data." *Econometrica* 62(1):95-116.

Keane, Michael, and Robert Sauer. 2006. "Classification Error in Dynamic Discrete Choice Models: Implications for Female Labor Supply Behavior." Discussion Paper 2332, Bonn: IZA.

Keane, Michael, and Kenneth Wolpin. 1997. "The Career Decisions of Young Men." *Journal of Political Economy* 105(3):474-521.

_____. 2001. "The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment." *International Economic Review* 42(4):1051-1103.

Kreider, Brent, and John Pepper. 2007A. "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors." *Journal of the American Statistical Association* 102(478):432-41.

_____. 2007B. "Inferring Disability Status from Corrupt Data." *Journal of Applied Econometrics.* Forthcoming.

Krueger, Alan and Lawrence Summers. 1988. "Efficiency Wages and Inter-Industry Wage Structure." *Econometrica* 56(2):259-93.

Lavy, Victor, Michael Palumbo, and Steven Stern. 1998. "Simulation of Multinomial Probit Probabilities and Imputation of Missing Data." In *Advances in Econometrics*, Volume 13. eds. Thomas Fomby and R. Carter Hill. Oxford: Elsevier.

Li, Tong, Pravin Trivedi, and Jiequn Guo. 2003. "Modeling Response Bias in Count: A Structural Approach with an Application to the National Crime Victimization Survey Data." *Sociological Methods and Research* 31(4):514-44.

Magnac, Thierry, and Michael Visser. 1999. "Transition Models with Measurement Errors." *Review of Economics and Statistics* 81(3):466-74.

Mathiowetz, Nancy. 1992. ‘‘Errors in Reports of Occupation.’’ *The Public Opinion Quarterly* 56(3):352-55.
McCall, Brian. 1990. ‘‘Occupational Matching: A Test of Sorts.’’ *Journal of Political Economy* 98(1):45-69.
Mellow, Wesley, and Hal Sider. 1983. ‘‘Accuracy of Response in Labor Market Surveys: Evidence and Implications.’’ *Journal of Labor Economics* 1(4):331-44.
Mroz, Thomas. 1999. ‘‘Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome.’’ *Journal of Econometrics* 92(2):233-74.
Neal, Derek. 1999. ‘‘The Complexity of Job Mobility Among Young Men.’’ *Journal of Labor Economics* 17(2):237-61.
Poterba, James, and Lawrence Summers. 1995. ‘‘Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification.’’ *The Review of Economics and Statistics* 77(2):207-16.
Ramalho, Esmeralda. 2002. ‘‘Regression Models for Choice-based Samples with Misclassification in the Response Variable.’’ *Journal of Econometrics* 106(1):171-201.
Roy, Andrew. 1951. ‘‘Some Thoughts on the Distribution of Earnings.’’ *Oxford Economic Papers* 3(2):135-46.
Stinebrickner, Todd. 1999. ‘‘Estimation of a Duration Model in the Presence of Missing Data.’’ *The Review of Economics and Statistics* 81(3):529-42.
Stinebrickner, Ralph, and Todd Stinebrickner. 2004. ‘‘Time Use and College Outcomes.’’ *Journal of Econometrics* 121(1):243-69.
Sullivan, Paul. 2007. ‘‘Empirical Evidence on Occupation and Industry Specific Human Capital.’’ Working Paper: http://www-personal.umich.edu/~paulsull/research.html.