Estimation of an Occupational Choice Model when Occupations are Misclassified

Sullivan, Paul

Bureau of Labor Statistics

November 2006

Estimation of an Occupational Choice Model when Occupations are Misclassified

Paul Sullivan*
Bureau of Labor Statistics

November 2006

Abstract

This paper examines occupational choices using a discrete choice model that accounts for the fact that self-reported occupation data is measured with error. Despite evidence from validation studies which suggests that there is a substantial amount of measurement error in self-reported occupations, existing research has not corrected for classification error when estimating models of occupational choice. This paper develops a panel data model of occupational choices that corrects for misclassification in occupational choices and measurement error in occupation-specific work experience variables. The model is used to estimate the extent of measurement error in self-reported 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% of occupational choices in the 1979 cohort of the National Longitudinal Survey of Youth are misclassified. Ignoring misclassification biases the median parameter in the occupational choice model by 25%.

JEL codes: J24, C25, C15

Keywords: Occupational choice, Misclassification, Discrete choice, Simulation methods

*I would like to thank John Bound and Steven Stern for many helpful comments and discussions during the course of this project. I would also like to thank Timothy Erickson, Loren Smith, and seminar participants at the University of Virginia, the Society of Labor Economists Annual Meeting, and the Bureau of Labor Statistics for helpful comments. Finally, I gratefully acknowledge funding for this research provided by a National Institute on Aging Post Doctoral Fellowship. Email: Sullivan.Paul.Joseph@bls.gov. Phone: (202) 691-6593. Mail: Bureau of Labor Statistics, Postal Square Building Room 3105 MC 204, 2 Massachusetts Ave. N.E., Washington, D.C. 20212-0001.

1 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% of one-digit occupational choices are misclassified, it is surprising that existing research has not corrected for 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.[1]

This paper develops a panel data model of occupational choices that corrects for the measurement error in the dependant variable created by misclassification of occupations, estimates the extent of misclassification in the data, and demonstrates the substantial bias in parameter estimates caused by ignoring classification error when estimating an occupational choice model. The estimation method developed in this paper also employs simulation methods to correct for measurement error in the occupation specific work experience variables used as explanatory variables in the model.

The classification error literature consists of two broadly defined approaches to estimating parametric models in the presence of classification error.[2] One approach uses assumptions about

[1] See Bound, Brown, and Mathiowetz (2001) for a discussion of the effects of measurement error in dependant and independent variables for both linear and nonlinear models.

[2] 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

the measurement error process along with auxiliary information on error rates, which typically takes the form of validation or re-interview 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. Examples of this approach are found in Hausman, Abrevaya, and Scott-Morton (1998) who develop a maximum likelihood estimator that corrects for misclassification in the dependant variable of a binary choice model, and Li, Trivedi, and Guo (2003) who estimate a count model with misclassification. A related methodology is employed by Dustmann and van Soest (2001), who estimate a model of the relationship between language fluency and earnings that corrects for misclassification in self reported language fluency.

The occupational choice model developed in this paper combines features of the two existing approaches to misclassification. Instead of relying solely on 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.

The model of occupational choices presented in this paper builds on the models of self selec-

independent variables in a linear regression, and Kreider and Pepper's (2004A, 2004B) work on misclassification in disability status.

tion in sectoral and occupational choices used by Heckman and Sedlacek (1985,1990) and Gould (2002). Workers in the model self select into one-digit occupations based on their skills and preferences which influence the wages and non-pecuniary utility received while employed in each of the eight occupations in the economy. The model expands on previous occupational choice models by explicitly allowing for misclassification in observed occupational choices by incorporating misclassification probabilities that indicate the probability of observing a worker in each occupation conditional on the worker's actual occupational choice. The misclassification probabilities are estimated along with the other parameters of the model, and these estimates provide direct evidence on the extent of misclassification in the data as well as information about the patterns of misclassification between occupations. In addition, the model allows misclassification rates to be heterogenous across people. It is necessary to control for this person-specific heterogeneity because in panel data, some individuals may persistently provide poor descriptions of their occupations that are likely to be misclassified when these verbatim descriptions are translated into occupation codes.[3]

One key 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. This situation arises in a panel data occupational choice model because when a current period occupational choice is misclassified it creates measurement error in future occupation specific work experience variables.[4] This problem has not been addressed in existing models of misclassification or in the occupational choice literature. It is addressed in this work by using the model of misclassification to derive the dis-

[3]See Dustmann and van Soest (2001) for a model of misclassification applied to panel data that allows for person-specific heterogeneity in propensity to falsely report language fluency.

[4]This is the case because the amount of occupation specific work experience that a worker has accumulated as of year t in occupation q is simply the total number of times that the individual reported working in occupation q in the previous years.

tribution of true occupation specific work experience conditional on the observed occupational choices, wages, and other explanatory variables in the model. The distribution of the true occupation specific experience variables for a given person is used to integrate out the effects of measurement error on each individual's likelihood contribution. This approach creates serious computational problems because treating occupation specific work experience as an unobserved state variable creates a likelihood function composed of high dimensions integrals that are extremely difficult to evaluate. This research addresses this problem by employing recent advances in integral simulation techniques to approximate the otherwise intractable integrals over the distribution of true occupation specific experience that appear in the likelihood function. This application of simulation methods adds to a growing literature that uses simulation methods to solve problems created by missing data and measurement error.[5] The simulation algorithm developed in this paper is applicable in a wide range of settings beyond occupational choice models. For example, a natural application of these techniques would be to studies of labor force participation or unemployment, where current labor force status is measured with error and accumulated work experience impacts the probability of employment.

The parameter estimates show that a substantial fraction of occupational choices (9%) are misclassified in the NLSY data. The extent of misclassification varies widely across occupations, with 96% of craftsmen classified in the correct occupation, while only 77% of service workers are correctly classified. The estimates also indicate 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 91% of professionals with reported wages in the top 10% of the professional wage distribution are correctly classified as professionals, but

[5]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.

only 75% of those observed in the bottom 10% of the professional wage distribution are correctly classified as professionals. There is a similarly strong and intuitively plausible relationship between education and misclassification in occupation codes. For example, 71.8% of workers who are correctly classified as professionals graduated from college, while only 30.2% of workers who are incorrectly classified as professionals graduated from college.

The bias caused by ignoring classification error when estimating a one-digit occupational choice model is substantial. The average parameter is biased by 60% when classification error is ignored, while the median parameter is biased by 25%. The largest biases are found in parameters that measure the transferability of occupation specific human capital across occupations. For example, ignoring misclassification in occupation codes overstates the effect of experience as a craftsman or operative on wages in the professional occupation by 38% and 73%, respectively. Classification error in occupations creates serious bias in estimates of the parameters of an occupational choice model, so researchers should be careful to examine the robustness of their results to misclassification when studying occupational choices and the returns to occupation specific human capital.[6]

An additional application of the model developed in this paper is that it can be used to simulate occupational choice data that is free from misclassification, because estimating the model recovers the distribution of true occupational choices conditional on observed occupational choices and wages. This simulated data can be used in place of the noisy occupational choice data in a wide range of applications, ranging from simple descriptive analyses of the patterns in occupational mobility to estimation of dynamic structural models of occupational choices.

The remainder of this paper is organized in the following manner. Section 2 discusses the

[6]While estimating the returns to firm specific and general work experience has long been a major research topic for economists, in recent years attention has turned to the importance of occupation and industry specific work experience. See, for example, Neal (1995), Parent (2000), and Kambourov and Manovskii (2006).

data. Section 3 presents the occupational choice model with misclassification, and Section 4 presents the parameter estimates. Section 5 discusses how the model can be used to simulate occupational choice data that is free from misclassification and examines the simulated data. Section 6 concludes examines the sensitivity of the results to the existence of measurement error in wages, and Section 7 concludes.

2 Data

The National Longitudinal Survey of Youth (NLSY) is a panel dataset 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 21 when first interviewed in 1979. The employment data contains 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. Individuals who ever report serving in the military, working as farmers, or being self-employed are excluded from the sample. The NLSY work history files are used to construct a monthly history of each individual's primary employment using the weekly employment records. This analysis considers only full time employment, which is defined as a job where the weekly hours worked are at least 20. The intent of this analysis 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.

The weekly labor force record is aggregated into a monthly employment record based on the number of weeks worked in each full time job in each month. An individual's primary job for

each month is defined as the one in which the most weeks were spent during that month. The monthly employment record is used to create a running tally of accumulated work experience in each occupation for each worker.

Descriptions of the one-digit occupation classifications along with average wages are presented in Table 1a. 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 1b. There are 954 individuals in the sample who contribute a total of 10,573 observations to the data. On average, each individual contributes approximately 11 observations to the data.

2.1 Measurement Error in Occupation Codes & Descriptive Statistics

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 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% at the three digit level and 81% at the one digit level. As one would expect, there appears to be less measurement error in the fairly broadly defined one digit classifications compared to the more narrowly defined three digit groupings. Additional evidence on measurement error in occupation codes is presented by Mathiowetz (1992). Mathiowetz (1992) independently creates one and three-digit occupation codes based on occupational descriptions from employees of a large man-

ufacturing firm and job descriptions found in these worker's personnel files. The agreement rate between these independently coded one-digit occupation codes is 76%, while the agreement rate for three-digit codes is only 52%. 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% at the three-digit level.

Table 1a lists the one digit occupation classifications used throughout this study along with the mean wage in each occupation. Average wages vary widely across occupations, with managers earning the highest average hourly wage of $12.89, and service workers earning the lowest wage of $6.34. Table 2 provides information about occupational mobility in the form of a transition matrix. The top entry in each cell represents the percentage of employment spells in the NLSY data that start in the left column occupation and end in the top row occupation. Table 2 shows that persistence in occupational choices varies widely across occupations. For example, 74.7% of professionals remain in the professional occupation from one employment spell to the next, while only 36.2% of laborers remain in the laborer occupation from one spell to the next. Mobility occurs frequently between the closely related blue collar occupations of operatives, craftsmen, and laborers. Mobility is also quite common from the sales to managerial occupation, although mobility in the opposite direction is roughly half as common.

3 Occupational Choice Model with Misclassification

3.1 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 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 non-pecuniary utility,

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

where w_{iqt} is the log wage of person i in occupation q at time t, H_{iqt} is the non-pecuniary utility that person i receives from working in occupation q at time t, and ε_{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

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

where μ_{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.[7] Note that the commonly estimated log wage equation which assumes that only total work experience influences wages,

[7] See Keane and Wolpin (1997) for an example of a paper that allows for cross-occupation experience effects. Their occupational choice model allows blue collar and white collar experience to enter into the wage equations in both the blue and white collar occupations.

rather than occupation specific work experience, is nested within this specification. Equation (2) reduces to this "standard" wage equation when all the δ's in the model are equal, ($\delta_{11} = \delta_{qk}, q = 1, ..., Q, \;\; k = 1, ..., Q$). The final term, $e_{iqt,}$ represents a random wage shock. The non-pecuniary utility flow equation is specified as

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

where X_{it} is a vector of explanatory variables, Exp_{ikt} is person i's experience at time t in occupation k, $Lastocc_{ikt}$ is a dummy variable equal to 1 if person i worked in occupation k at time $t-1$. This variable allows switching occupations to have a direct impact on non-pecuniary utility, as it would if workers incur non-pecuniary costs when switching occupations. The final term, $\phi_{iq,}$ represents person i's innate preference for working in occupation q.

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 1 to Q. A person's true occupational choice may differ from the one observed in the data if classification error exists. Let $\widehat{O}_{it}$ represent the true occupational choice, which is simply the occupation that yields the highest utility,

$$\widehat{O}_{it} = q \;\; \text{if } V^*_{iqt} = \max\{V^*_{i1t}, V^*_{i2t}, ..., V^*_{iQt}\}. \tag{4}$$

The model of misclassification used in this paper builds on the model of misclassification in a binary dependant variable developed by Hausman, Abrevaya, and Scott-Morton (1998) and the multinomial logit model with misclassification developed by Poterba and Summers (1995).[8] In this framework the probability of misclassification depends on the value of the latent variable V^*_{iqt}. The misclassification probabilities are denoted as

$$\alpha_{jk} = \Pr(O_{it} = j \mid \widehat{O}_{it} = k), \text{ for } j = 1, ..., Q, \;\; k = 1, ..., Q. \tag{5}$$

[8] An important distinction between these two papers is that Hausman et al. (1998) estimate misclassification probabilities jointly with the other parameters of their binary choice model, while Poterba and Summers (1995) consider the case where misclassification rates are known.

That is, α_{jk} represents the probability that the occupation observed in the data is j, conditional on the actual occupational choice being k. The α_{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,

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

Throughout this paper the term "misclassification probabilities" will be used when referring to all of the α_{jk}'s, but of course only the terms where $j \neq k$ truly represent misclassification probabilities, since the terms with $j = k$ are actually "correct classification probabilities." Note that this occupational choice model nests a standard occupational choice model which assumes that occupations are always correctly classified. When $\alpha_{jj} = 1$ for $(j = 1, ..., Q)$, and $\alpha_{jk} = 0$ for $j \neq k$ and $(j, k = 1, ..., Q)$, occupations are never misclassified. This model builds on existing models of misclassification such as Douglas, Smith Conway, and Ferrier (1995), Hausman, Abrevaya, and Scott-Morton (1998), and Dustmann and van Soest (2001). Following studies of this type, the model assumes that the misclassification probabilities $\{\alpha_{jk} : k = 1, .., Q, j = 1, ..., Q\}$ depend only on j and k, and not on the other explanatory variables in the model. This is a standard assumption in this type of 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. For example, it is possible that some workers consistently provide poor descriptions of their occupations over the course of their career which results in frequent misclassifications. On the other hand, other workers may provide very detailed descriptions that are much less likely to result in misclassification. Section 3.4 of this paper

presents an extension of the model that allows for this type of within-person correlation in misclassification rates.

This model of misclassification implies that the occupation specific experience variables, Exp_{iqt}, will be measured with error, since measurement error in a current occupational choice creates measurement error in future experience variables because the experience variables are calculated using a worker's sequence of observed occupations. This measurement error is non-classical because it is correlated with observed choices. A method for dealing with this problem is presented in the next section.

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 ϕ_i represent a $Q \times 1$ vector of person i's preferences for working in each occupation, and let μ_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 θ represent the vector of parameters in the model, $\theta = \{\beta_k, \gamma_{kj}, \chi_{kj}, \alpha_{kj}, \pi_k, \delta_{jk}, \sigma_{ek}, F(\mu, \phi) : k = 1, ..., Q,\ j = 1, ..., Q\}$. Define $\widehat{P}_{it}(q, w^{obs}_{it})$ 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}. 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 $f(e_{-q})$ as the joint density of the wage error terms excluding the error term for occupation q.

The outcome probability is

$$\begin{aligned}
\widehat{P}_{it}(\ q, w_{it}^{obs} |\ \mu, \phi) &= \Pr(V_{iqt}^* = \max\{V_{i1t}^*, V_{i2t}^*, ..., V_{iQt}^*\} \cap w_{iqt} = w_{it}^{obs}) \qquad (7) \\
&= \Pr(V_{iqt}^* = \max\{V_{i1t}^*, V_{i2t}^*, ..., V_{iQt}^*\} \mid w_{iqt} = w_{it}^{obs}) \times \Pr(w_{iqt} = w_{it}^{obs}) \\
&= \int \cdots \int \frac{\exp(w_{it}^{obs} + H_{iqt})}{\exp(w_{it}^{obs} + H_{iqt}) + \sum_{j \neq q} \exp(w_{ijt} + H_{ijt})} f(e_{-q}) de_{-q} \times \\
&\quad \frac{1}{\sigma_{eq}} \Phi\left(\frac{w_{it}^{obs} - \mu_{iq} - Z_{it}\pi_q - \sum_{k=1}^{Q} \delta_{qk} Exp_{ikt}}{\sigma_{eq}}\right),
\end{aligned}$$

where Φ represents the standard normal pdf. During the evaluation of the likelihood function the integral over the distribution of $f(e_{-q})de_{-q}$ is simulated by taking random draws from the distribution and computing the average of $\widehat{P}_{it}(\ q, w_{it}^{obs} |\ \mu, \phi)$ over the draws.[9] The likelihood function for the observed data is constructed using the misclassification probabilities and the true choice probabilities. Define $P_{it}(q, w_{it}^{obs})$ as the probability that person i is *observed* working in occupation q at time period t with a wage of w_{it}^{obs}. This probability is the sum of the true occupational choice probabilities weighted by the misclassification probabilities,

$$P_{it}(q, w_{it}^{obs} \mid \mu, \phi) = \sum_{k=1}^{Q} \alpha_{qk} \widehat{P}_{it}(k, w_{it}^{obs} \mid \mu, \phi). \qquad (8)$$

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 worker intentionally misrepresents his occupation and simultaneously provides a false wage consistent with the false occupation. This assumption implies that observed wages provide information about true occupational choices. 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

[9] During estimation, 60 draws 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 re-started 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.

they are in the sample,

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

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

where $1\{\bullet\}$ denotes the indicator function which is equal to 1 if its argument is true and 0 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.[10] Suppose that there are M types of people, each with a $Q \times 1$ vector of wage intercepts μ^m and $Q \times 1$ vector of preferences ϕ^m. Let ω_m represent the proportion of the mth type in the population. The unconditional likelihood function is simply a weighted average of the type specific likelihoods,

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

3.2 Evaluating the Likelihood Function

The parameters of the model can be estimated by maximizing the likelihood function shown in equation number (11). The major complication arises from the fact that classification error in occupation codes creates measurement error in the occupation specific work experience variables and previous occupational choice dummy variables that are used as explanatory variables. This section describes the relationship between measurement error in occupation codes and measure-

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

ment error in occupation specific work experience variables and explains how simulation methods can be used to correct for this measurement error during the evaluation of the likelihood function.

The intuition behind this approach is that the model of misclassification of occupational choices presented in the previous section defines a relationship between the occupational choices observed in the data and the true occupational choices predicted by the model. This relationship implies that conditional on the occupational choices, wages, and other explanatory variables observed in the data, the model implies there is a distribution of true values of occupation specific experience and true lagged occupational choices. The distribution of the true data conditional on the observed data can be used to integrate out the effects of measurement error. Unfortunately, the distribution of true lagged occupational choices and experience variables is intractably complex. This work overcomes this limitation by using simulation methods to evaluate the otherwise intractable integrals that arise when misclassified occupational choices creates measurement error in explanatory variables.

Let $\widehat{Exp}_{iqt}$ represent person i's true experience in occupation q in time period t. Define $\widehat{Exp}_{it}$ as a $Q \times 1$ vector of experience in each occupation. These experience variables are not observed in the data, because the data only contains information about reported occupation specific experience, Exp_{iqt}, which is measured with error. Let $\widehat{Lastocc}_{it}$ represent a $Q \times 1$ vector of dummy variables where the qth element is equal to 1 if person i's true occupational choice was q in time period t. Let $F(\widehat{Exp}, \widehat{Lastocc})$ represent the distribution of true occupation specific experience and lagged occupational choices. This distribution is a function of each person's observed characteristics, and observed choices and wages. The likelihood function must be integrated over this distribution when it is evaluated during estimation,

$$L(\theta) = \prod_{i=1}^{N} \int L_i(\theta | \widehat{Exp}, \widehat{Lastocc}) dF(\widehat{Exp}, \widehat{Lastocc}). \tag{12}$$

The likelihood function is difficult to evaluate because the distribution of actual occupation

specific experience and lagged choices is intractably complex, but recent advances in integral simulation methods provide a way to evaluate the likelihood function. The likelihood function can be simulated using a modified Geweke (1991), Hajivassiliou (1990), and Keane (1994) (GHK) algorithm to simulate the likelihood contribution. 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.

3.2.1 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. In the case of multiple types, the algorithm is simply repeated for each type to obtain the type-specific likelihood contributions found in the likelihood function, because the likelihood function is simply a weighted average of the type-specific likelihood contributions. The object that must be simulated is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} \int \prod_{t=1}^{T(i)} \sum_{q=1}^{Q} 1\{O_{it}=q\} P_{it}(q, w_{it}^{obs} \mid \theta, Z_{it,} X_{it}, \widehat{Exp}_{ikt}, \widehat{Lastocc}_{ikt}) dF(\widehat{Exp}, \widehat{Lastocc}) \\ &= \prod_{i=1}^{N} \int \prod_{t=1}^{T(i)} L_{it}(O_{it}, w_{it}^{obs} \mid \theta, Z_{it,} X_{it}, \widehat{Exp}_{ikt}, \widehat{Lastocc}_{ikt}) dF(\widehat{Exp}, \widehat{Lastocc}) \end{aligned} \tag{13}$$

Let variables with a $*$ superscript represent simulated variables, and let $r = 1, ..., R$ 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, and $Lastocc_{it+1}^{*}(r|\theta, O_{it}, w_{it}^{obs}, Z_{it,} X_{it}, Exp_{it}^{*}, Lastocc_{it}^{*})$ 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 since the career begins at the first job, so initialize the simulated experience vector to zero for time periods $t = 1, ..., T$: $Exp_{i1}^{*}(r) = 0$,and $Lastocc_{i1}^{*}(r) = 0$.

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

3. Compute and store the probability that person i's true choice in time period t ($\widehat{O}_{it}$) is each of the Q possible occupations, conditional on the parameter vector (θ), observed choice (O_{it}), observed wage (w_{it}^{obs}), explanatory variables ($Z_{it,}X_{it}$), and simulated previous occupational choice ($Lastocc_{it}^{*}$) and experience variables (Exp_{it}^{*}). Let $\Omega_{it}(r, q|\rho)$ for $q = 1, ..., Q$ 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 a function of the previously defined outcome probabilities ($\widehat{P}_{it}(\bullet)$) and misclassification probabilities (α's),

$$\Omega_{it}(r, q|\rho) = \Pr(\widehat{O}_{it} = q \mid \rho) \tag{14}$$

$$= \frac{\alpha_{O_{it},1}\widehat{P}_{it}(q, w_{it}^{obs})}{\sum_{k=1}^{Q}\alpha_{Oit,k}\widehat{P}_{it}(k, w_{it}^{obs})} \, . \tag{15}$$

   Recall that $\widehat{P}_{it}(\bullet)$ is a function of all of the variables that $\Omega_{it}(\bullet)$ is conditioned on, but they are suppressed here as they were in equation (7). This implies that the observed wage and all the explanatory variables provide information about the conditional true choice probabilities ($\Omega_{it}(\bullet)$).

4. Use the Q computed conditional true choice probabilities, $\Omega_{it}(r, q|\rho)$, to define the discrete

distribution of true occupational choices $\{\Pr(O^*_{it} = q) = \Omega_{it}(r, q|\rho)\}$, $q = 1, ..., Q\}$. 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, $Exp^*_{it+1}(r|\rho)$ and $Lastocc^*_{it+1}(r|\rho)$. 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 $Lastocc^*_{it+1}$ vector corresponding to the simulated occupation in time t equal to one and set all other elements of the vector to zero. More precisely, increment the jth element of the vector $Exp^*_{it+1}(r)$ by one if $O^*_{it}(r) = j$, and leave all other elements of $Exp^*_{it+1}(r)$ unchanged from their values in time period t. Set the jth element in the vector $Lastocc^*_{it+1}(r)$ equal to one, and set all other elements of $Lastocc^*_{it+1}(r)$ equal 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,

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

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

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

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 .01% change in the value of the likelihood function at the simulated maximum likelihood parameter estimates.[11]

3.3 Identification

This section presents the identification conditions for the occupational choice model with misclassification and discusses the intuition behind how the misclassification model identifies certain occupational choices as likely to be misclassified.

3.3.1 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,

$$\sum_{k \neq j} \alpha_{jk} < \alpha_{jj}, \ j = 1, ..., Q. \tag{16}$$

This condition implies that on average, the occupational choices observed in the data are correct. The intuition behind this identification condition is 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. A key implication of this identification condition is that when the likelihood

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

function is being maximized during simulated maximum likelihood (SML) estimation, the SML parameter vector is confined to an area of the parameter space where the true occupational choices generated by the model correspond to those observed in the data to a certain minimum extent. This rules out extreme situations where misclassification accounts for the majority of the observed occupational choices in the data. For example, this assumption rules out the extreme case where the model evaluated at the SML parameter vector assigns extremely low true choice probabilities to every occupational choice observed in the data and instead accounts for all observed occupational choices through misclassification.[12]

3.3.2 An Illustrative Example

This section presents an actual occupational choice sequence drawn from the NLSY and discusses how the misclassification model uses the predicted true choice probabilities and wage data to infer the probability that an occupational choice is misclassified. Consider the following sequence of occupational choices found in the data for a particular person in the NLSY.

Age	24	25	26	27
Observed occupation	Professional	Craftsman	Professional	Professional
Observed wage	\$10.75	\$11.85	\$13.83	\$13.90
Years of College	4	4	4	4

At the time that this person is observed switching from the professional occupation to the craftsmen occupation he has worked as a professional for two years and has never worked as a craftsman. In addition, this worker is a college graduate, and college graduates do not typically

[12] During estimation, the identification constraint ($\sum_{k \neq j} \alpha_{jk} < \alpha_{jj},\ j = 1, ..., Q$) was never directly imposed on the parameter vector. One could do this by parameterizing the α's in such a way that the condition is required to hold, or by adding a penalty function to the likelihood function, but neither of these approaches was used during estimation. Although it was possible for the identification to be violated during the course of estimation, in practice this never occured during runs of the estimation program.

work as craftsmen. Also, the mean wage for a professional is \$11.19, while the mean wage for a craftsman is only \$8.53. Given this information, it seems plausible that this transition from professional to craftsmen employment is a false one created by classification error. From an intuitive standpoint, the consistent pattern of this worker choosing professional employment over his career, the cycling between professional and craftsmen employment, the patterns in the observed wages, and the relationship between college graduation and occupational choices all combine to make this a suspicious occupational transition. As the following discussion will demonstrate, the misclassification model incorporates all of these considerations when it determines whether or not an occupational choice is likely to be misclassified.

First, consider the information provided by the panel nature of the data. The identification condition shown in equation 16 implies that on average the occupational choices observed in the data are correct, so it would require an extremely unlikely sequence of misclassifications to account for a person being falsely observed as a professional over the course of their entire career. Consistently observing a worker as a professional and observing the associated wages provides information about a person's ability and preference for professional employment relative to other occupations.

At this point it is useful to first examine a simple occupational choice model with misclassification that does not incorporate wage data. This model is the one found in Section 3 under the restriction that $w_{iqt} = 0$. For simplicity, suppose that there are only two occupations, where the professional occupation is defined as occupation 1 and the craftsman occupation is defined as occupation 2. The probability that this person is observed as a craftsman is

$$P_{it}(2) = \alpha_{22}\widehat{P}_{it}(2) + \alpha_{21}\widehat{P}_{it}(1), \tag{17}$$

where α_{22} is the probability that this person is correctly classified as a craftsman and $\widehat{P}_{it}(2)$ is the probability that working as a craftsman is the optimal choice. Recall that the true occupa-

tional choice probabilities ($\widehat{P}_{it}$'s) are functions of all of the explanatory variables in the model, such as education. Estimating the parameters of the model with maximum likelihood involves maximizing a likelihood function composed of observed choice probabilities of this form. The derivatives of the observed choice probability with respect to the α's are

$$\frac{\partial P_{it}(2)}{\partial \alpha_{22}} = \widehat{P}_{it}(2) \quad \text{and} \quad \frac{\partial P_{it}(2)}{\partial \alpha_{21}} = \widehat{P}_{it}(1). \tag{18}$$

This example shows that, roughly speaking, when it is very likely that this person actually chooses to work in occupation 2 ($\widehat{P}_{it}(2)$ is large) this particular person's contribution to the likelihood function will be maximized by making α_{22} large. In the context of this example, if college educated workers are on average very unlikely to work as craftsmen, but very likely to work as professionals, then $\widehat{P}_{it}(2)$ will be small relative to $\widehat{P}_{it}(1)$. This example illustrates that the estimates of the misclassification probabilities will be determined by the extent to which the choices observed in the data are likely to be generated as optimal occupational choices by the model.

One clear shortcoming of the preceding model is that potentially useful information found in wages is excluded. The outcome probabilities in the model that incorporates wages include the joint density of observed choices and wages,

$$P_{it}(2, w_{it}^{obs} = \$11.12) = \alpha_{22}\widehat{P}_{it}(2, w_2 = \$11.85) + \alpha_{21}\widehat{P}_{it}(1, w_1 = \$11.85), \tag{19}$$

where $\widehat{P}_{it}(2, w_2 = \$11.85)$ is the joint probability that occupation 2 is the optimal choice and a wage of \$11.85 is observed in occupation 2. The derivatives of this outcome probability with respect to the a's are

$$\frac{\partial P_{it}(2)}{\partial \alpha_{22}} = \widehat{P}_{it}(2, w_2 = \$11.85) \quad \text{and} \quad \frac{\partial P_{it}(2)}{\partial \alpha_{21}} = \widehat{P}_{it}(1, w_1 = \$11.85). \tag{20}$$

Once wages are incorporated into the model, the derivatives of the outcome probabilities with respect to the α's depend on the probability of observing a wage of \$11.85 in each occupation in

addition to the choice probabilities predicted by the model, which capture the effect of variables such as education on true occupational choices. The model takes into account how consistent the observed wage is with the wage distribution in each occupation while the α's are being estimated. If, for example, a wage of \$11.85 is unlikely to be observed in the craftsman occupation (occupation 2) but is much more likely to be observed in the professional occupation (occupation 1), then α_{22} should be relatively small, while the misclassification probability α_{21} should be relatively large. In addition, the fact that wages vary strongly with occupation specific work experience provides further variation in the wage distribution across occupations that helps to identify wages that don't appear to fit well in the reported occupation. For example, a craftsman with 15 years of experience may be fairly likely to earn a wage of \$11.85, but it is probably very unlikely for a person working as a craftsman for the first time to earn a wage of \$11.85 when the mean wage in the craftsmen occupation is only \$8.53. In general, choice-wage combinations where the reported wage is unlikely to be observed in the reported occupation and where the observed occupational choice is unlikely to be generated as an optimal choice in the model are the ones that support the existence of misclassification. Estimates of the misclassification probabilities will be determined by the likelihood of the choices observed in the data being generated by the model and by the extent to which observed wages are consistent with reported occupational choices.

3.4 An Extended Model: Heterogeneity in Misclassification Rates

The model of misclassification presented in Section 3.1 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. This type of time-persistent misclassification has been examined by Dustmann and van Soest (2001) in their model of misclassification of self reported language fluency. Dustmann and van Soest (2001) allow for several subpopulations who have different propensities to over or under report their language fluency, and they estimate the subpopulation-specific misclassification rates along with the proportions of each subpopulation in the overall population. 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).

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

$$\alpha_{jk}(y) = \Pr(O_{it} = j|\widehat{O}_{it} = k),\ j = 1, ..., Q,\ k = 1, ..., Q, \tag{21}$$

$$\sum_{j=1}^{Q} \alpha_{jk}(y) = 1, \quad k = 1, ..., Q,\ y = 1, 2, 3. \tag{22}$$

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. For example, if $\alpha_{11}(2) > \alpha_{11}(3)$, then conditional on the true occupational choice being occupation 1, a person from subpopulation 2 is always more likely to be correctly classified in occupation 1 over his entire career than a person from subpopulation 3. Note that there are $3\times[(Q\times Q) - Q]$ misclassification probabilities

that must be estimated when there are three subpopulations in the economy. 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 1 are always correctly classified. Under this restriction the misclassification probabilities for subpopulation one are: $\alpha_{jk}(1) = 1$ if $j = k$, $\alpha_{jk}(1) = 0$ if $j \neq k$, for $j = 1, ..., Q$, and $k = 1, ..., Q$.[13] The misclassification parameters for the second and third subpopulations are not restricted during estimation. Note that the subpopulation probabilities are estimated, so although this specification restricts members of subpopulation 1 to always be correctly classified, this is not a restrictive assumption because as $\xi(1)$ approaches zero the proportion of people who are always correctly classified approaches zero.

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 level of misclassification and the particular patterns in misclassification are allowed to vary between subpopulations. Estimating the parameters of the model reveals the extent of misclassification in occupations and the importance of person-specific heterogeneity in misclassification rates.

The likelihood function presented in section 3.1 can be modified to account for person-specific heterogeneity in misclassification. The observed choice probabilities presented are easily modified

[13] 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.

so that they are allowed to vary by subpopulation,

$$P_{it}(q, w_{it}^{obs} \mid \mu, \phi, y) = \sum_{k=1}^{Q} \alpha_{qk}(y) \widehat{P}_{it}(k, w_{it}^{obs} \mid \mu, \phi), \tag{23}$$

where $y = 1, 2, 3$ indexes subpopulations. Conditional on subpopulations, the likelihood function is

$$\begin{aligned} L(\theta|y) &= \prod_{i=1}^{N} \int \prod_{t=1}^{T(i)} \sum_{q=1}^{Q} 1\{O_{it} = q\} P_{it}(q, w_{it}^{obs} \mid \mu, \phi, y) dF(\mu, \phi) & (24) \\ &= \prod_{i=1}^{N} \int L_i(\theta|\mu, \phi, y) dF(\mu, \phi), & (25) \end{aligned}$$

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

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} \sum_{y=1}^{3} \sum_{m=1}^{M} \xi(y) \omega_m L_i(\theta \mid y, \ \mu_i = \mu^m, \phi_i = \phi^m) & (26) \\ &= \prod_{i=1}^{N} L_i(\theta) \ . \end{aligned}$$

4 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.

4.1 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 3. 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 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% 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 3 show the probabilities that occupational choices are correctly classified. Averaged across all occupations, the probability that an occupational choice is correctly classified is .868 for subpopulation 2 and .840 for subpopulation 3. One striking feature of the estimated misclassification probabilities is the large variation in misclassification rates across occupations. In subpopulation 2 the probability that an occupational choice is correctly classified ranges from a low of .56 for sales workers to a high of .99 for craftsmen, while in subpopulation 3 the probability that an occupational choice is correctly classified ranges from a low of .60 for sales workers to a high of .98 for operatives.

The estimates of the probabilities that a person belongs to subpopulations 2 and 3 are 42% and 19%, which leaves an estimated 38% 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. The subpopulation-specific misclassification rates discussed in the previous paragraph must be averaged over the three subpopulations to produce an estimate of the overall extent of misclassification in the NLSY data. The estimates indicate that $(.42 \times .86 + .19 \times .84 + .38 \times 1 = 91\%)$ 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% 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% agreement rate between the occupational descriptions given by workers of a single large manufacturing firm and personnel records.

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 would be possible to test this hypothesis by re-estimating the occupational choice model developed in this paper using the CPS data. This extension is left for future research, since it appears that the procedures used by the CPS and NLSY in constructing occupation codes are quite similar. Given these similarities, it is not clear that one should expect the NLSY data to 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.[14] 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

[14]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.

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 show that certain types of jobs are likely to be misclassified in particular directions. Simpler models of misclassification that restrict the probability of misclassification to be constant across occupations or impose symmetry in the misclassification rate matrix are clearly inadequate. The estimates of the misclassification probabilities for subpopulation 2 show that the sales occupation is the occupation that is most frequently misclassified. Conditional on a person being employed as a sales worker, there is a 21% chance that in the data they will be misclassified as a manager. The classification error matrix is highly asymmetric. Note that there is only a 1.4% chance that a manager will be misclassified as a sales worker, but there is a 21% chance that a sales worker will be misclassified as a manager. The high misclassification rate for sales workers may be due in part to the existence of a three-digit occupation of sales managers, which falls under the one-digit classification of managers.

The estimated misclassification probabilities for the blue collar occupations shown in panel A of Table 3 show that these workers are most commonly misclassified into closely related low skill occupations, although there are several exceptions. Reading down the laborers column of panel A of Table 3 shows that laborers are frequently misclassified as service workers (22%), but service workers are very unlikely to be misclassified as laborers (.39%). Service workers are most frequently misclassified as professionals (6.4%) and sales workers (7.7%). The misclassification rates between service and professional employment provide another example of asymmetry in

the misclassification rate matrix, since professional workers are unlikely to be misclassified as service workers (.18%), but service workers are frequently misclassified as professionals (6.4%). The relatively large number of service workers misclassified as professionals may be caused by health service workers such as health aides and nursing aides who are incorrectly coded as health professionals. Ignoring this type of misclassification may result in serious biases in studies of wage differences between occupations, since professionals are one of the highest paid occupations, while service workers are the lowest paid one-digit occupation.

The overall rate of misclassification rate for subpopulation 3 is approximately 3 percentage points higher than the misclassification rate found in subpopulation 2, but the similarity of the overall misclassification rates masks several large differences between subpopulations in the patterns of misclassification between occupations. For example, only 71% of service workers are correctly classified in subpopulation 3, and these workers are largely misclassified as managers (25%). In contrast, 82% of service workers are correctly classified in subpopulation two, and these workers are relatively unlikely to be misclassified as managers (2%). Similarly, sales workers in subpopulation two are frequently misclassified as managers (21%), but in subpopulation three the corresponding misclassification rate is only 8%. Overall, these results show that there is considerable heterogeneity in the patterns of misclassification across people in the NLSY. In addition, the occupations of a sizeable fraction of the population (38%) are never misclassified. The variation in misclassification rates across subpopulations suggests that a sizeable component of the measurement error in occupation codes is due to errors or ambiguities in the occupational descriptions provided by survey respondents, rather that due to errors introduced by coders as they translate the job descriptions into occupation codes. If all misclassification in occupation codes arises from mistakes made by coders, then one would not expect to find evidence of person-

specific heterogeneity in misclassification rates.[15]

4.2 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 6. In addition, this table presents the percent change in each parameter from the model that corrects for classification error compared to the model that ignores classification error. Let β_E represent the estimated parameter in the misclassification model, and let β_B represent the same parameter in the baseline model that does not incorporate classification error. The percent bias in absolute value resulting from ignoring classification error and examining occupational choices using the model that does not correct for classification error is

$$\% \text{ abs dev} = \frac{|\beta_B - \beta_E|}{|\beta_E|}.$$

Before examining the effects of classification error in occupations on individual parameters it is useful to summarize the overall effects of ignoring classification error on the parameter estimates of the occupational choice model. The preceding section demonstrates that misclassification rates are substantial, but the most important question to be addressed when examining measurement error in occupation codes is the bias resulting from estimating models that do not take into account misclassification. The average and median of the percent absolute deviations between the baseline and misclassification models are presented in Table 4. The average parameter in the occupational choice model is biased by 59.9% when the occupational choice model is estimated without accounting for misclassification in occupation codes. The large average bias is driven upwards by a number of large outliers, but the median bias is still substantial at 24.7%.

[15] In the NLSY, the occupation coders translated the occupational descriptions into census occupation codes after each yearly survey was conducted, so there is little chance of coder-specific measurement error. In addition, when coding the occupations for a given year each occupation coder did not have access to the occupational descriptions provided by the respondent in previous interviews.

These results indicate that ignoring classification error creates significant bias in estimates of the parameters of an occupational choice model. These findings are consistent with the results of Hausman et al. (1998), who find that even small amounts of misclassification in the dependant variable of a binary discrete choice model creates substantial bias in parameter estimates.

4.2.1 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.[16] One obvious problem created by the misclassification of occupations is that some wage observations used to estimate occupation specific wage function are classified into the wrong occupation. The patterns of misclassification present in the data will be a key determinant of the magnitude and direction of the resulting bias. If workers are generally misclassified into occupations with wage distributions similar to their actual occupation, one would expect the bias to be less than if workers are frequently misclassified from low to high paying occupations. The second problem created by misclassification is measurement error in occupation specific experience variables that arises when reported occupations are used to create experience variables. Again, it seems likely that the patterns in misclassification will be a key factor in determining the severity of the bias resulting from measurement error in occupation specific experience variables.

Table 5 shows how the bias in wage equation parameters in each occupation varies with misclassification rates by listing the probability of a worker being misclassified “out of" or “into" each occupation along with the average and median percent deviations of the wage equation parameters in the baseline model from those in the classification error model. For example, the first row of Table 5 shows that the probability of a worker being falsely classified as a

[16] See Bound, Brown, and Mathiowetz (2001) for a discussion of the effects of measurement error in both linear and nonlinear models.

professional worker is .07, while the probability of a professional worker being misclassified into another occupation is .06. The average bias caused by ignoring classification error for a parameter in the professional wage equation is 110%, and the median bias is 59%. There is considerable variation in the misclassification rates into (.04 to .22) and out of (.02 to .42) occupations as well as in the median bias created by ignoring misclassification (14% to 59%). There is no obvious relationship between the misclassification rates and the bias created by misclassification. This result highlights the fact that the level of misclassification is not the sole determinant of how much bias is created by misclassification, it is the level weighted by the importance of the misclassified choices and observed wages. For example, the largest median bias is found in the professional wage equation even though this occupation has among the lowest rates of misclassification.

The wage equation parameter estimates are presented in Panel A of Table 6. There are a large number of wage equation parameters because there is a separate wage equation for each occupation, so this section focuses on the effects of ignoring classification error in occupation codes on selected parameter estimates. The estimates of the wage equation in the professional occupation show 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% from .064 to .037 when misclassification is ignored. The model that ignores classification error overstates the effect of experience as a craftsman or operative on professional wages by 38% (.0280 to .0203) and 73% (.0447 to .0259). Accounting for classification error removes false transitions in the data where craftsmen and operatives are observed switching to professional employment, and so the effects of craftsman and operative experience on professional wages are greatly reduced. The substantial bias in estimates of the value of experience in other occupations on wages in the

professional occupation is relevant for studies of wage growth over the career as well as studies of occupational mobility because wage changes accompanying occupational switches partly reflect the transferability of skills across occupations. The bias in these particular parameters is also interesting because the estimated misclassification probabilities show that professionals are rarely misclassified as managers ($\alpha_{21}(2) = .0066$, $\alpha_{21}(3) = .0099$), and managers are rarely misclassified as professionals ($\alpha_{12}(2) = .0018$, $\alpha_{12}(3) = .0043$). Similarly, the misclassification rates averaged across subpopulations between the professional and operatives and craftsmen occupations are below 1%. The low misclassification rates between these occupations combined with the large bias in the experience coefficients shows 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% of sales workers are correctly classified. In the most common subpopulation, sales workers are most likely to be misclassified as managers ($\alpha_{23}(2) = .21$), 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 results in a relatively small overestimate of the effect of a year of sales experience on wages in the managerial occupation (.0888 vs. .0879), while the value of experience as a manager in the managerial occupation is overstated by 19%. Correcting for classification error also causes large changes in estimates of the effects of experience on wages in the sales occupation. The model that ignores classification error indicates that a year of sales experience increases a sales worker's wages by 17%, but this estimate falls by 13% to 15% once classification error is accounted for. The effect of clerical experience on a sales worker's wages is overstated by 197% when classification error is ignored, even though misclassification rates between the sales and clerical occupations are low relative to the misclassification rate between sales and managerial employment. Ignoring

classification error leads to the misleading conclusion that clerical experience has a large and statistically significant effect on wage in the sales occupation, but correcting for classification error shows that clerical experience does not have a statistically significant effect on sales wages at any conventional significance level. Similarly, ignoring classification error leads to a 118% overstatement in the value of professional experience in the sales occupation (.0672 vs. .0308). The bias in these parameters is another example of large biases in estimates of the effects of human capital on wages resulting from ignoring classification error in occupations.

Further evidence of large changes in estimates of the transferability of human capital across occupations are found in the remaining occupations. For example, there are several large changes in the wage equation for craftsmen and operatives between the models with and without classification error. The model that does not correct for classification error implies that a year of professional experience increases a craftsman's wages by 2.9%, and this effect is statistically significant at the 5% level. Once classification error is accounted for this effect falls to 1.8% and it is not statistically different from zero at the 5% 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. Even though misclassification leads to relatively few of these false transitions, the bias is substantial.

The estimates of the service occupation wage function show that ignoring classification error leads to a 178% overstatement of the value of clerical experience in the service occupation, and a 307% overstatement of the value of operative experience in the service occupation. On the other hand, the transferability of skills between the laborer and service occupations is vastly understated when classification error is ignored (.0177 vs. .0674).

The estimates of the wage equations show that misclassification creates substantial bias in wage equation parameter estimates. The average wage equation parameter is biased by 59%, while the median parameter is biased by 21%. The effects of misclassification are quite complicated, and parameters may be biased upwards or downward by measurement error. One of the key insights derived from these estimates is that substantial bias is created by classification error even in occupations where approximately 98% of choices are correctly classified. An important implication of these results is that an analysis of human capital wage functions that does not take into account classification error in occupations will lead in some cases to misleading conclusions about the effects of occupation specific human capital on wages. Given that misclassification results in many false transitions between occupations, it seems reasonable that some of the most seriously biased parameters are those that measure the transferability of human capital across occupations.

The final parameters of the wage equation are the standard deviations of the random shock to wages in each occupation, σ_{eq}, for $q = 1, ..., 8$. 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. Ignoring classification error biases the estimate of the standard deviation of the wage shock upwards by 36% for professionals, 48% for managers, 13% for craftsmen, 25% for operatives, and 16% for service workers. The intuition behind the direction of this bias is that when classification error is ignored the model must provide an explanation for the large number of short duration occupation switches that occur in the data. One way the model can explain these transitions is through large wage shocks that create short duration occupation switches. The model that allows for classification error provides an alternative explanation which is that some occupation switches are created by classification error. Once this alternative explanation is available, the variance of the wage shocks is reduced because

classification error provides an explanation for some of the patterns in observed occupational mobility and observed wages that is more consistent with the data than large wage shocks.

4.2.2 Non-pecuniary Utility Flows & Unobserved Heterogeneity

The occupational choice model presented in this paper allows occupational choices to depend on non-pecuniary utility flows as well as wages. The importance of modelling 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 non-pecuniary utility flow equations for the models estimated with and without accounting for classification error are presented in Panel B of Table 6. These results show that the average parameter in the non-pecuniary utility flow equations is biased by 59% when classification error is ignored, and the median parameter is biased by 30%. 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 non-pecuniary 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 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% when classification error is ignored. 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 extremely sensitive to classification error. This result seems sensible since one would expect estimates of the effects of lagged choices to be quite sensitive to the false transitions between occupations created by classification error.

Estimates of the effects of occupation specific work experience on non-pecuniary utility are also quite sensitive to classification error in occupation codes. For example, the effect of experience as a manager on the non-pecuniary utility flow from being employed as a manager is biased downward by 24% when classification error is ignored. Estimates of the effects of experience in other occupations on the operative utility flow are biased by even larger amounts. The effect of craftsman experience on operative utility is biased downward by 51%, and the effect of laborer experience on operative utility is biased downward by 82%. Ignoring classification error leads to serious bias in estimates of the effects of occupation-specific work experience on non-pecuniary utility.

The estimates of the wage intercepts (μ's) and non-pecuniary intercepts (ϕ's) for the three types of people in the model are presented in Panel C of Table 6. These parameter estimates reveal the extent of unobserved heterogeneity in skills and preferences for employment in each occupation. The estimates of the wage and non-pecuniary intercepts for the model that corrects for classification error show that preferences for employment in each occupation vary widely across types. For example, the professional non-pecuniary intercept ranges from -4.72 for a type 1 person to -2.82 for a type 3 person, and the clerical non-pecuniary intercept ranges from -1.79 to $-.56$ across types. These intercepts are measured relative to the value of employment in the service occupation.

The final section of Panel C of Table 6 shows the averages of the wage and non-pecuniary 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 occupations (ϕ's). The average preference for working as a craftsman changes from .048 in the model that ignores classification error to .23 in the model that corrects for classification error, a change of 79%. The average preference for employment as operatives and laborers are each biased by approximately 60% when classification error is ignored, while the average preference for employment as a sales worker is biased by 69%. The 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 non-pecuniary utility flows in part through preference heterogeneity.

The bias in estimates of the average occupation-specific ability parameters (μ's) is much lower than the bias in the preference parameters. The bias in the mean wage intercepts is lower than 14% across all occupations. The bias is extremely low in the clerical (3.8%) and laborer (.5%) occupations. It appears that ignoring classification error causes the model to explain observed patters in occupational mobility largely though unobserved heterogeneity in preferences rather than heterogeneity in ability, which results in larger bias for parameters that measure preferences compared to those that measure ability.

5 Simulating Data that is Free from Misclassification

One useful 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. Estimating the parameters of the model amounts to estimating the distribution of true occupational choices conditional on the choices and wages observed in the data, so it is fairly straightforward to simulate

occupational choices by drawing from this distribution. The simulation algorithm outlined in section 3.2.1 explains how the model can be used to simulate true occupational choices for each person in the sample conditional on the observed choices, wages, and other explanatory variables found in the data. The only minor complication is that each person must be randomly assigned both a type (vectors of μ's and ϕ's) and a misclassification subpopulation before their occupational choices are simulated. Define w_i as a vector of person i's observed wages over his entire career: $w_i = \{w_{it},\ t = 1, ..., T(i)\}$. Define X_i, Z_i, Exp_i, and $Lastocc_i$ as the analogous vectors of these explanatory variables over person i's career, and let $\Theta_i = \{X_i, Z_i, Exp_i, Lastocc_i\}$. The conditional probability that a particular person is of type k and subpopulation (pop) j is

$$\begin{aligned}\Pr(type = k, pop = j | O_i, w_i, \Theta_i) &= \frac{\Pr(O_i, w_i | type = k, pop = j, \Theta_i) \Pr(type = k, pop = j)}{\Pr(O_i, w_i | \Theta_i)} \\ &= \frac{L_i(O_i, w_i | type = k, pop = j, \Theta_i)\omega_k \xi_j}{L_i(O_i, w_i | \Theta_i)}.\end{aligned}$$

Occupational choices are simulated by first computing $\Pr(type = k, pop = j | O_i, w_i, \Theta_i)$ for each person in the sample, and then randomly assigning a type and subpopulation to each person using these probabilities. Then, occupational choices are simulated conditional on the simulated type and subpopulation using the algorithm outlined in section 3.2.1.

5.1 Simulated Occupational Choices

Table 2 presents occupational transition matrices for the actual data (top entry) and simulated data (bottom entry) together to facilitate a comparison of the changes in the patterns of occupational mobility that result from correcting for classification error in occupations. The simulated data is based on $2,000$ simulated careers. The diagonal elements of the matrix are larger in the simulated data compared to the actual data. This indicates that the net effect of misclassification is to create false transitions between occupations that lead to an overstatement of occupational

mobility. Eliminating the false transitions created by classification error leads to the largest increase in the persistence of occupational choices for professional workers (74.7% to 78.5%) and service workers (59.5% to 63.7%).

The fact that occupational choices become more persistent in the simulated data provides information about the types of occupational choices that are likely to be flagged as misclassified by the misclassification model. One possible concern is that wage outliers may cause the model to incorrectly flag occupational choices as misclassified because workers have (accurately measured) wage outliers at certain points over their career. If this concern is valid, one would expect occupational transitions to increase in the simulated data. However, the fact that the simulated data shows more persistence in occupational choices than the noisy data provides evidence against this concern, because overall the misclassification model is removing occupational transitions, not creating new transitions.[17]

The increase in the persistence of occupational choices is of course accompanied by a corresponding decreases in occupational mobility. Some of the noteworthy decreases in mobility occur between the sales and managerial occupations and the service and managerial occupations. Classification error causes the data to overstate the mobility of sales workers into managerial employment by 62%, overstate the mobility of sales workers into clerical employment by 22%, and overstate the mobility of clerical workers into professional employment by 18%. The simulated data indicates that across all occupations, 9% of all occupational choices are misclassified.

5.1.1 Which Workers are Misclassified?

One explanatory variable that is of central importance when investigating occupational choices is education. There is strong sorting across occupations based on completed education. Given this

[17] Section 5.1.2 presents evidence that the misclassification model also does not repeatedly flag individuals as misclassified who have unusually high or low wages in their reported occupation.

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. This type of analysis provides information about the extent to which the model uses variation in occupational choice probabilities with education levels to identify misclassified occupations.

Table 7 shows the distribution of completed education for correctly classified and misclassified occupational choices, disaggregated by occupation. For example, the table shows that 10.8% of those workers who are correctly classified as professionals have not completed any years of college, while 48.6% of workers who are misclassified as professionals have not completed any years of college. A correctly classified professional has a 71.8% change of being a college graduate, while a worker misclassified as a professional has only a 30.2% 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. When the model is used to generate simulated occupational choices it tends to remove workers who have not completed any college from the professional occupation. 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 level of completed higher education. It is reassuring that the model tends to flag workers as misclassified who appear to have reported education levels that are inconsistent with their reported occupation.

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, since college graduates are unlikely to work in low skill occupations. This is in fact what the results in Table 7 show. For example, the percentage of correctly classified

workers who have graduated from college is 2.1% for craftsmen, 2.5% for operatives, and 3.2% for laborers. In contrast, for workers who are falsely classified in these occupations the percentage of workers who are college graduates is 18.7% for craftsmen, 21.5% for operatives, and 11.7% for laborers. In general, the workers who are misclassified into these blue collar occupations are much more likely to be college graduates compared to workers who are correctly classified in these occupations.

5.1.2 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 8 presents the distribution of the total number of times that occupational choices are misclassified over the course of a person's career. The final column of Table 8 shows that across all three subpopulations, 57.2% of people never have any of their occupational choices misclassified at any point during their career. The relatively large number of people who are never misclassified is made up of two groups. First, an estimated 39% of the population belongs to subpopulation 1 and therefore by definition never experience misclassification. Second, some members of the other two subpopulations never experience misclassification because of the random nature of misclassification. Reading down the final column of Table 8 shows that the majority of workers never experience misclassification (57.2%), 17.6% of workers are misclassified once over their career, and 11.5% of workers are misclassified twice over their career. To provide some context for these results, recall that the average worker contributes approximately 11 observations to the data set.

One important feature of Table 8 is that it shows that it is very unlikely that a worker's occupational choices will be consistently misclassified over the course of his career. For example,

only 4.3% of the sample is misclassified more than five times over the course of the career. Another notable feature of Table 8 is that the number of times that a person is misclassified is extremely similar for subpopulations 2 and 3. This result is driven by the fact that the misclassification rates averaged across occupations are quite close for the two subpopulations (.87 for subpop. 1 and .84 for subpop. 2). While subpopulations 2 and 3 experience substantial differences in the patterns of misclassification between occupations, the overall error rates are quite similar. More detail about misclassification over the course of the career is presented in Table 8, which provides information about the lengths of misclassification spells. Table 8 shows the distribution of the number of times a person is consecutively misclassified, conditional on being misclassified. For example, the first entry in the final column of Table 9 shows that conditional on an occupational choice being misclassified, there is a 72.9% chance that the person will be *correctly* classified in the next survey. Conditional on being misclassified, there is an 18.3% chance that a person will be misclassified in two consecutive periods, and there is only a 5.2% chance that a person will be misclassified in three consecutive periods.[18]

5.1.3 True Occupational Choices, Observed Choices, and Wages

The comparison of the occupational transition matrix observed in the data with the transition matrix generated by the model highlights the overall changes in occupational mobility when classification error in occupations is corrected for. Table 10 extends this analysis by showing the average true occupational choice probabilities conditional on observed choices and observed wages. This analysis shows how classification error rates vary with observed wages across occupations and provides a more detailed analysis of the type of occupational choice and wage

[18] 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.

combinations that are likely to be affected by classification error.

Observed occupational choices are listed in the far left column of Table 10, 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 10. 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 10 shows that a worker observed in the data as a professional worker with a wage in the top 10% of the professional wage distribution has a 90.9% chance of being correctly classified as a professional worker. However, a worker observed as a professional with a wage in the bottom 10% of the professional distribution has only a 75.7% chance of actually being a professional worker. People observed in the data as low wage professional workers are primarily service workers (9.5%).

Among workers observed in the data classified as managers, 85.8% of those in the middle 10% of the managerial wage distribution are actually managers, 56.5% of the highest paid workers are actually managers, while only 54.4% of those in the bottom 10% of the managerial wage distribution are correctly classified. The vast majority of workers misclassified as managers are actually sales workers. The wide variation in misclassification rates with observed wages lends support to the use of wages to identify false occupational choices. Apparently, the variation in the wage distribution across occupations provides enough information for the model to assign true choice probabilities that vary substantially with observed wages. Combinations of choices and wages that do not fit the wage distribution are assigned high misclassification probabilities by the model.

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% of clerical workers in the top 10% of the clerical wage distribution are correctly classified, but 3.9% of those observed as high wage clerical workers are actually professionals. The simulated choices reveal that conditional on being observed at the top of the clerical wage distribution there is a significant chance that the worker is actually a misclassified professional worker (3.9%). However, the unconditional probability that a professional is misclassified as a clerical worker is much lower ($\alpha_{41}(2) = .013, \alpha_{41}(3) = .013$), so the simulations demonstrate that wages provide a substantial amount of information about which occupational choices are misclassified. The difference between the unconditional misclassification probabilities and the misclassification probabilities that condition on wages highlights the value of using wages to identify misclassified occupational choices.

There are some patterns present for workers observed as operatives that are worth discussing. Conditional on being in the bottom 10% of the operative wage distribution there is an 86.9% chance that the worker is correctly classified as an operative. The model suggests that the majority of these low wages workers misclassified as operatives are actually sales workers (11.9%). A similar pattern of misclassification is found for craftsmen, where the simulations indicate that 12% of those observed as low wage craftsmen are actually sales workers. These patterns of misclassification are somewhat surprising because the operatives and sales occupations are composed of jobs that appear to be quite different overall. However, a closer look a the three-digit occupations that make up the sales category shows that this occupation includes three-digit occupations such as “manufacturing industry sales" and “construction sales". These jobs may be the ones where respondents’ descriptions of their sales jobs could be mistaken for descriptions of craftsman or operative employment.

The fraction of service workers that are correctly classified is fairly stable across wage deciles, ranging from $.719 - .732$. The majority of workers misclassified as service workers are actually employed in the closely related low skill occupation of laborers, but a number of those observed as high wage service workers are actually professionals (5.4%). In addition, a sizeable fraction of those observed as low wage professionals are actually service workers (9.5%). These patterns of misclassification are somewhat surprising, since the unconditional probability that a professional worker is misclassified as a service worker is quite low ($\alpha_{81}(2) = .0011, \alpha_{81}(3) = .0049$), but conditional on observed wages the probabilities are fairly large. One explanation for the relationship between observed wages and predicted misclassification is that there are certain high paying professional jobs that are misclassified as closely related, but much lower paying service occupations. For example, low wage health aides and nursing aids being falsely classified as high wage health professionals.

5.2 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. It is widely known that wages are measured with error, so it is important to examine whether or not the existence of measurement error in wages affects the estimates of the extent of measurement error in occupation codes. One way of addressing this question is to simulate noisy wage data, re-estimate 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,

$$w_{it}^{me} = w_{it}^{obs} + \nu_{it}, \text{ where } \nu_{it} \sim N(0, \sigma_{\nu}^2). \tag{27}$$

Recall that w_{it}^{obs} is a log wage, so the extent of measurement error in the noisy log wage data is captured by σ_ν^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 σ_ν^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 σ_ν^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 that 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. The parameter estimates found in Table 3 and the columns in Table 6 labelled "correct for classification error" serve as the baseline, since these parameter estimates were obtained using the NLSY wage data. When the noisy wage data is used in place of the NLSY wage data the average parameter in the model changes by approximately 2%, 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 two and three, $\alpha_{jj}(y)$, $j = 1, ..., Q$, $y = 2, 3$. Recall that these parameters reflect the probabilities that occupational choices are correctly classified. Across both subpopulations, the use of noisy wage data results in the average probability of correct classification ($\frac{1}{2Q}\sum_{y=1,2}\sum_{j=1,Q}\alpha_{jj}(y)$) decreasing by only $-.006$ from .8546 to .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 .008, so

it appears that estimates of the overall extent of misclassification in the NLSY occupation data are quite robust to measurement error in wages.

Of course, it is possible for the overall level of misclassification to remain approximately constant while the patterns in misclassification between occupations change substantially, so it is necessary to check if the off diagonal elements of the misclassification matrices ($\alpha_{jk}(y)$, $j \neq k$) are affected by measurement error in wages. The average absolute change in these misclassification rates is only .0015, so these off diagonal elements are not greatly impacted by measurement error in wages. The results of this simulation exercise show that even with a substantial amount of measurement error in wages that accounts for 30% of total variation in wages, the estimates of the misclassification parameters are extremely robust.

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. For example, the panel nature of the data as well as the strong relationship between observable variables (such as education) and occupational choices provide a large amount of information about misclassification. Another key point is that many of the occupational choices that are flagged in the simulations as misclassifications are associated with extremely large wage changes. Wage changes of this magnitude are unlikely to be generated in large numbers by measurement error in wages that is of the magnitude found in validation studies. 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.

6 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%. The results also suggest that person-specific heterogeneity in misclassification rates is an important feature of the data. An estimated 38% 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 oc-

cupation. 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 create substantial bias in parameter estimates. The median parameter is biased by 25% when classification error in occupations is ignored, and the magnitude and direction of the bias varies widely across parameters. 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. Ignoring classification error also creates significant biases in estimates of the importance of unobserved heterogeneity in preferences and random wage shocks in determining career choices.

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. Another interesting avenue for future research would be to examine classification error in three-digit occupation codes as opposed to the more broadly defined one-digit codes considered in this paper. However, the extremely detailed nature of the three-digit occupation codes makes this a challenging problem for future research.

Table 1a: Description of Occupations

One-Digit Occupation	Mean Wage	Example Three-Digit Occupations
Professional, technical & kindred workers	$11.19	Accountants, chemical engineers, physicians, social scientists
Managers & 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 & unskilled workers	$7.48	Bank tellers, cashiers, receptionists, secretaries
Craftsmen & 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 in 1979 dollars.

Table 1b: Descriptive Statistics

Variable	Mean	Standard Deviation
Age	27.09	2.24
Education	14.26	.91
North central	.34	.47
South	.29	.45
West	.17	.37
Number of observations	10,573	
Number of individuals	954	

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

Table 2: Occupational Transition Matrix – NLSY Data (top entry) and Simulated Data (bottom entry)

	Professional	Managers	Sales	Clerical	Craftsmen	Operatives	Laborers	Service	Total
Professional	74.7 78.5	6.9 5.6	2.3 4.2	4.5 3.7	5.0 3.2	3.0 2.2	1.3 1.4	2.2 1.2	100
Managers	6.4 6.6	57.4 58.5	7.2 9.4	7.3 7.4	10.7 10.3	3.5 2.9	2.5 2.6	5.0 2.3	100
Sales	8.0 7.6	14.9 9.2	53.5 55.2	7.7 6.3	5.4 6.8	5.2 5.9	2.2 5.2	3.2 3.6	100
Clerical	10.3 8.7	12.4 11.4	5.9 7.2	44.8 45.8	6.8 6.3	7.0 6.8	8.3 9.8	4.6 4.0	100
Craftsmen	2.9 2.0	5.3 4.7	1.0 2.3	2.2 2.0	66.6 67.4	11.1 10.8	8.1 9.6	2.6 1.2	100
Operatives	2.4 1.9	2.2 1.3	2.1 3.3	3.1 2.9	18.4 18.3	56.8 56.3	10.1 11.6	4.9 4.4	100
Laborers	2.7 2.5	3.3 2.7	1.8 4.0	7.9 7.3	23.2 21.6	18.6 16.5	36.2 39.1	6.1 6.2	100
Service	3.9 3.7	7.8 4.2	1.5 2.8	3.5 3.1	8.4 6.8	6.8 6.2	8.6 9.5	59.5 63.7	100
Total	14.0 13.9	11.5 9.5	5.3 7.9	7.6 7.3	25.8 25.2	16.9 16.2	9.6 11.5	9.4 8.4	100

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

Table 3, Panel A: Parameter Estimates- Misclassification Probabilities for Subpopulation 2 ($\alpha_{jk}(2)$)

Observed/Actual	Professional	Managers	Sales	Clerical	Craftsmen	Operatives	Laborers	Service
Professional	.9570 (.0023)	.0018 (.0096)	.0264 (.0025)	.0017 (.0074)	.0033 (.0002)	.0007 (.0004)	.0380 (.0004)	.0641 (.0017)
Managers	.0066 (.0041)	.9762 (.0042)	.2128 (.0026)	.0052 (.0082)	.0013 (.0002)	.0021 (.0015)	.0011 (.0022)	.0238 (.0003)
Sales	.0036 (.0016)	.0148 (.0046)	.5578 (.0001)	.0133 (.0045)	.0000 (.0015)	.0029 (.0017)	.0019 (.0040)	.0774 (.0009)
Clerical	.0131 (.0002)	.0031 (.0101)	.0131 (.0055)	.9579 (.0046)	.0002 (.0016)	.0021 (.0022)	.0046 (.0067)	.0042 (.0033)
Craftsmen	.0055 (.0023)	.0022 (.0045)	.1063 (.0098)	.0052 (.0025)	.9897 (.0054)	.0055 (.0030)	.0204 (.0121)	.0024 (.0023)
Operatives	.0121 (.0025)	.0000 (.0064)	.0456 (.0005)	.0013 (.0082)	.0000 (.0039)	.9849 (.0058)	.0063 (.0009)	.0004 (.0223)
Laborers	.0000 (.0003)	.0000 (.0131)	.0164 (.0043)	.0136 (.0085)	.0054 (.0021)	.0016 (.0082)	.7029 (.0014)	.0039 (.0079)
Service	.0018 (.0002)	.0018 (.0043)	.0213 (.0008)	.0014 (.0086)	.0000 (.0018)	.0000 (.0022)	.2243 (.0012)	.8235 (.0068)
Pr(subpopulation 2)	.4243 (.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.

Table 3, Panel B : Misclassification Probabilities for Subpopulation 3 ($\alpha_{ij}(3)$)

Observed/Actual	Professional	Managers	Sales	Clerical	Craftsmen	Operatives	Laborers	Service
Professional	.9289 (.0022)	.0043 (.0097)	.0394 (.0024)	.0012 (.0079)	.0357 (.0005)	.0007 (.0003)	.0104 (.0003)	.0190 (.0016)
Managers	.0099 (.0040)	.9641 (.004)	.0822 (.0027)	.0032 (.0081)	.0046 (.0003)	.0030 (.0016)	.0041 (.0021)	.2548 (.0002)
Sales	.0096 (.0016)	.0248 (.0056)	.6007 (.0001)	.0125 (.0046)	.0002 (.0016)	.0003 (.0018)	.0022 (.0041)	.0026 (.0008)
Clerical	.0126 (.0001)	.0027 (.0103)	.0052 (.0054)	.9634 (.0045)	.0004 (.0011)	.0012 (.0023)	.0031 (.0061)	.0006 (.0037)
Craftsmen	.0234 (.0024)	.0025 (.0043)	.0904 (.0096)	.0068 (.0024)	.9475 (.0061)	.0067 (.0031)	.0504 (.0130)	.0041 (.0026)
Operatives	.0106 (.0026)	.0007 (.0065)	.1335 (.0005)	.0029 (.0081)	.00000 (.0047)	.9833 (.0051)	.0054 (.0008)	.0000 (.0001)
Laborers	.0000 (.0001)	.0000 (.0141)	.0307 (.0041)	.0082 (.0084)	.0114 (.0020)	.0040 (.0084)	.6215 (.0005)	.0042 (.0069)
Service	.0049 (.0002)	.0008 (.0043)	.0176 (.0009)	.0016 (.0088)	.0000 (.0017)	.0006 (.0024)	.3028 (.0008)	.7139 (.0048)
Pr(subpopulation 3	.1937 (.0235)							

Table 4: Summary of Bias Caused by Ignoring Classification Error

	Average % deviation	Median % deviation
Wage equation	59.4	20.6%
Non-pecuniary utility flow equation	59.2	29.6%
All parameters	59.9	24.7%

Notes: Entries are the percent absolute deviations of parameters in the baseline model from the model that corrects for misclassification.

Table 5: Misclassification Rates by Occupation and Bias in Wage Equations from Ignoring Misclassification

Occupation	Misclassified into	Misclassified out of	Average % deviation	Median % deviation
Professional	.07	.06	110%	59%
Managers	.20	.02	27%	14%
Sales	.14	.42	27%	14%
Clerical	.06	.02	47%	18%
Craftsmen	.04	.02	11%	10%
Operatives	.05	.01	25%	14%
Laborers	.04	.20	113%	30%
Service	.22	.14	70%	26%

Notes: "Misclassified into" refers to Pr(observed in listed occupation $\cap$ actually work in another occupation). Misclassified out of refers to Pr(actually work in listed occupation $\cap$ observed in another occupation). Entries in the rightmost two columns are the percent absolute deviations of parameters in the baseline model from the model that corrects for misclassification.

Table 6 Panel A: Parameter Estimates – Wage Equation

Wage equation	*Professional*		% absolute deviation	*Managers*		% absolute deviation
	Ignore classification error	Correct for classification error		Ignore classification error	Correct for classification error	
Age	0.0233 (0.0154)	0.0079 (0.0073)	1.9644	0.0474 (0.0184)	0.0351 (0.0091)	0.3504
$Age^2/100$	-0.2280 (0.0985)	-0.1434 (0.0426)	0.5904	-0.4028 (0.1230)	-0.3543 (0.0630)	0.1368
Education	0.0734 (0.0057)	0.0626 (0.0041)	0.1725	0.0825 (0.0082)	0.0837 (0.0060)	0.0143
Professional experience	0.0715 (0.0053)	0.0687 (0.0034)	0.0408	0.0944 (0.0130)	0.0896 (0.0086)	0.0536
Managerial experience	0.0375 (0.0158)	0.0644 (0.0123)	0.4177	0.0656 (0.0071)	0.0547 (0.0055)	0.1993
Sales experience	0.0493 (0.0147)	0.0499 (0.0101)	0.0120	0.0888 (0.0135)	0.0879 (0.0097)	0.0102
Clerical experience	0.0430 (0.0191)	0.0377 (0.0162)	0.1406	0.0191 (0.0096)	0.0209 (0.0073)	0.0861
Craftsmen experience	0.0280 (0.0092)	0.0203 (0.0100)	0.3793	0.0488 (0.0074)	0.0556 (0.0062)	0.1223
Operatives experience	0.0447 (0.0236)	0.0259 (0.0210)	0.7259	0.0634 (0.0124)	0.0705 (0.0121)	0.1007
Laborer experience	0.0146 (0.0291)	-0.0083 (0.0232)	2.7676	0.0416 (0.0268)	0.0233 (0.0179)	0.7854
Service experience	0.0000 (0.0224)	0.0718 (0.0234)	1.0005	0.0100 (0.0140)	0.0069 (0.0117)	0.4504
North central	-0.0635 (0.0262)	-0.0139 (0.0189)	3.5683	-0.1063 (0.0302)	-0.0667 (0.0233)	0.5935
South	-0.0448 (0.0245)	0.0222 (0.0182)	3.0180	-0.0726 (0.0345)	-0.0849 (0.0284)	0.1449
West	0.0412 (0.0294)	0.1046 (0.0205)	0.6061	-0.0919 (0.0438)	-0.0531 (0.0311)	0.7307

Note: Standard errors in parentheses.

Table 6 Panel A: Parameter Estimates – Wage Equations

Wage equation	*Sales*		% absolute deviation	*Clerical*		% absolute deviation
	Ignore classification error	Correct for classification error		Ignore classification error	Correct for classification error	
Age	0.0662 (0.0368)	0.1354 (0.0272)	0.5110	0.0480 (0.0153)	0.0413 (0.0157)	0.1622
$Age^2/100$	-1.0006 (0.2662)	-1.0984 (0.1886)	0.0890	-0.4330 (0.1057)	-0.3588 (0.1095)	0.2069
Education	0.1837 (0.0189)	0.1593 (0.0268)	0.1528	0.0528 (0.0087)	0.0511 (0.0081)	0.0333
Professional experience	0.0672 (0.0366)	0.0308 (0.0542)	1.1818	0.0957 (0.0146)	0.1051 (0.0230)	0.0899
Managerial experience	0.1316 (0.0274)	0.1089 (0.0322)	0.2086	0.0454 (0.0104)	0.0418 (0.0121)	0.0861
Sales experience	0.1774 (0.0163)	0.1571 (0.0195)	0.1296	0.0806 (0.0162)	0.0888 (0.0203)	0.0923
Clerical experience	0.1281 (0.0333)	0.0430 (0.0433)	1.9799	0.0562 (0.0085)	0.0572 (0.0093)	0.0175
Craftsmen experience	-0.0183 (0.0258)	-0.0453 (0.0297)	0.5960	0.0502 (0.0083)	0.0646 (0.0119)	0.2229
Operatives experience	0.0845 (0.0284)	0.0845 (0.0297)	0.0000	0.0516 (0.0118)	0.0500 (0.0125)	0.0320
Laborer experience	0.0507 (0.0431)	0.0521 (0.0552)	0.0269	0.0420 (0.0167)	0.0345 (0.0153)	0.2174
Service experience	0.0241 (0.0295)	-0.0657 (0.0826)	1.3668	0.0191 (0.0177)	0.0215 (0.0183)	0.1116
North central	-0.2505 (0.0754)	-0.3711 (0.1051)	0.3250	-0.1688 (0.0311)	-0.1965 (0.0369)	0.1409
South	0.1225 (0.0764)	0.1249 (0.0915)	0.0195	-0.0847 (0.0307)	-0.1030 (0.0377)	0.1774
West	0.0979 (0.0945)	0.1015 (0.1070)	0.0358	-0.0228 (0.0342)	-0.0230 (0.0362)	0.0087

Note: Standard errors in parentheses.

Table 6 Panel A: Parameter Estimates – Wage Equations

Wage equation	*Craftsmen*		% absolute deviation	*Operatives*		% absolute deviation
	Ignore classification error	Correct for classification error		Ignore classification error	Correct for classification error	
Age	0.0606 (0.0068)	0.0489 (0.0053)	0.2393	0.0128 (0.0085)	0.0123 (0.0073)	0.0407
$Age^2/100$	-0.5257 (0.0481)	-0.4576 (0.0398)	0.1490	-0.2230 (0.0642)	-0.2605 (0.0604)	0.1436
Education	0.0290 (0.0048)	0.0254 (0.0045)	0.1417	0.0209 (0.0054)	0.0079 (0.0048)	1.6523
Professional experience	0.0290 (0.0120)	0.0188 (0.0210)	0.5426	0.0670 (0.0229)	0.0751 (0.0344)	0.1079
Managerial experience	0.0558 (0.0115)	0.0646 (0.0113)	0.1362	0.0432 (0.0157)	0.0552 (0.0152)	0.2174
Sales experience	0.0100 (0.0169)	0.0438 (0.0183)	0.7717	0.0200 (0.0149)	0.0157 (0.0176)	0.2739
Clerical experience	0.0381 (0.0125)	0.0366 (0.0210)	0.0410	0.0499 (0.0110)	0.0370 (0.0191)	0.3486
Craftsmen experience	0.0591 (0.0028)	0.0605 (0.0027)	0.0231	0.0607 (0.0067)	0.0764 (0.0062)	0.2055
Operatives experience	0.0386 (0.0052)	0.0352 (0.0048)	0.0966	0.0549 (0.0045)	0.0470 (0.0041)	0.1681
Laborer experience	0.0217 (0.0069)	0.0114 (0.0066)	0.9035	0.0708 (0.0090)	0.0512 (0.0077)	0.3828
Service experience	0.0254 (0.0094)	0.0361 (0.0106)	0.2964	-0.0023 (0.0149)	0.0285 (0.0147)	1.0811
North central	-0.1034 (0.0197)	-0.1201 (0.0185)	0.1392	-0.0637 (0.0266)	-0.0948 (0.0222)	0.3281
South	-0.0786 (0.0209)	-0.0828 (0.0182)	0.0507	0.0234 (0.0270)	0.0026 (0.0222)	7.8973
West	0.0847 (0.0210)	0.0868 (0.0208)	0.0242	0.0086 (0.0307)	-0.0043 (0.0268)	2.9908

Note: Standard errors in parentheses.

Table 6 Panel A: Parameter Estimates – Wage Equations

Wage equation	*Laborers*		% absolute deviation	*Service*		% absolute deviation
	Ignore classification error	Correct for classification error		Ignore classification error	Correct for classification error	
Age	0.0268 (0.0119)	0.0235 (0.0119)	0.1404	-0.0083 (0.0120)	-0.0116 (0.0093)	0.2819
$Age^2/100$	-0.3202 (0.0994)	-0.3339 (0.0961)	0.0410	0.0234 (0.0889)	0.0314 (0.0666)	0.2548
Education	0.0331 (0.0087)	0.0184 (0.0077)	0.7989	0.0965 (0.0071)	0.0864 (0.0070)	0.1169
Professional experience	0.0715 (0.0515)	0.0295 (0.0905)	1.4237	0.0285 (0.0359)	0.02738 (0.0258)	0.0409
Managerial experience	0.0457 (0.0232)	0.0597 (0.0478)	0.2345	0.0294 (0.0151)	0.0419 (0.0316)	0.2983
Sales experience	-0.0165 (0.0633)	0.0364 (0.0378)	1.4533	0.0132 (0.0178)	-0.0121 (0.0414)	2.0909
Clerical experience	0.0445 (0.0234)	0.0401 (0.0247)	0.1097	0.0240 (0.0185)	0.0086 (0.0391)	1.7810
Craftsmen experience	0.0559 (0.0082)	0.0683 (0.0088)	0.1816	0.0681 (0.0103)	0.0167 (0.0362)	3.0778
Operatives experience	0.0525 (0.0083)	0.0584 (0.0088)	0.1010	0.0304 (0.0179)	-0.0382 (0.0199)	1.7958
Laborer experience	0.0504 (0.0085)	0.0556 (0.0083)	0.0935	0.0177 (0.0219)	0.0674 (0.0341)	0.7374
Service experience	0.0040 (0.0158)	0.0009 (0.0195)	3.4222	0.0562 (0.0066)	0.0542 (0.0062)	0.0369
North central	-0.0866 (0.0393)	-0.0675 (0.0363)	0.2830	-0.2492 (0.0291)	-0.2297 (0.0239)	0.0847
South	-0.1109 (0.0408)	-0.0859 (0.0376)	0.2915	-0.1181 (0.0304)	-0.0865 (0.0315)	0.3649
West	-0.0043 (0.0492)	0.0235 (0.0524)	1.1843	-0.1278 (0.0290)	-0.1273 (0.0307)	0.0037

Note: Standard errors in parentheses.

Table 6 Panel A: Parameter Estimates – Error Standard Deviations

Occupation			
	Ignore classification error	Correct for classification error	% absolute deviation
Professional	0.3249 (0.0055)	0.2394 (0.0069)	0.3575
Managers	0.3701 (0.0080)	0.2493 (0.0163)	0.4847
Sales	0.5724 (0.0217)	0.6850 (0.0248)	0.1643
Clerical	0.2763 (0.0136)	0.2636 (0.0211)	0.0485
Craftsmen	0.3039 (0.0051)	0.2683 (0.0068)	0.1325
Operatives	0.3317 (0.0063)	0.2643 (0.0105)	0.2547
Laborers	0.3364 (0.0109)	0.3411 (0.0122)	0.0137
Service	0.3250 (0.0090)	0.2802 (0.0154)	0.1597

Note: Standard errors in parentheses.

Table 6 Panel B: Parameter Estimates – Non-pecuniary Utility

	Professionals		% absolute deviation	*Managers*		% absolute deviation
	Ignore classification error	Correct for classification error		Ignore classification error	Correct for classification error	
Age	0.1203 (0.0799)	0.0973 (0.0305)	0.2368	-0.0409 (0.0809)	-0.1448 (0.0551)	0.7176
$Age^2/100$	-0.2295 (0.6142)	-0.0345 (0.0000)	5.6531	0.4907 (0.5784)	1.0633 (0.4094)	0.5385
Education	0.4260 (0.0543)	0.4715 (0.0370)	0.0967	0.2145 (0.0570)	0.2631 (0.0240)	0.1848
High school diploma	-0.6393 (0.2354)	-0.3529 (0.2103)	0.8118	-0.2384 (0.2213)	-0.3125 (0.2106)	0.2371
College diploma	0.0984 (0.1881)	0.3509 (0.2144)	0.7196	0.2867 (0.2013)	0.5197 (0.2353)	0.4484
Professional experience	0.4819 (0.1484)	0.4894 (0.1162)	0.0152	0.3134 (0.1468)	0.3707 (0.1250)	0.1545
Managerial experience	-0.0761 (0.0830)	-0.0229 (0.1472)	2.3231	0.2605 (0.0688)	0.3433 (0.1308)	0.2412
Sales experience	-0.1569 (0.1171)	-0.1803 (0.1604)	0.1296	0.0811 (0.1009)	0.1052 (0.1430)	0.2293
Clerical experience	-0.1028 (0.1126)	-0.0184 (0.1637)	4.5877	0.1471 (0.0860)	0.2410 (0.1249)	0.3897
Craftsmen experience	0.1531 (0.0657)	0.2813 (0.1271)	0.4556	0.2197 (0.0579)	0.3523 (0.1224)	0.3763
Operatives experience	-0.1836 (0.0874)	-0.1056 (0.1784)	0.7378	0.0218 (0.0608)	0.1383 (0.1102)	0.8424
Laborer experience	-0.0459 (0.1420)	0.1008 (0.2052)	1.4554	-0.0207 (0.1182)	0.2366 (0.1849)	1.0875
Service experience	-0.4737 (0.0645)	-0.8955 (0.1467)	0.4710	-0.2765 (0.0574)	-0.2843 (0.0820)	0.0275
Previously a professional	2.469 (0.339)	3.108 (0.368)	0.2056	1.237 (0.379)	2.022 (0.484)	0.3880
Previously a manager	0.792 (0.340)	1.181 (0.665)	0.3295	2.780 (0.261)	3.717 (0.636)	0.2522
Previously sales	1.194 (0.459)	0.893 (0.594)	0.3376	1.703 (0.432)	1.623 (0.591)	0.0492
Previously clerical	1.628 (0.354)	1.546 (0.364)	0.0529	1.853 (0.322)	2.198 (0.287)	0.1569
Previously a craftsman	1.042 (0.298)	1.064 (0.485)	0.0215	1.673 (0.294)	2.482 (0.472)	0.3260
Previously an operative	0.752 (0.305)	0.537 (0.488)	0.4004	0.400 (0.320)	0.493 (0.516)	0.1886
Previously a laborer	0.634 (0.346)	0.341 (0.509)	0.8592	0.839 (0.333)	0.931 (0.471)	0.0988

Note: Standard errors in parentheses.

Table 6 Panel B: Parameter Estimates – Non-pecuniary Utility

	Sales		% absolute deviation	*Clerical*		% absolute deviation
	Ignore classification error	Correct for classification error		Ignore classification error	Correct for classification error	
Age	-0.1327 (0.1137)	-0.3511 (0.0484)	0.6221	-0.1327 (0.1137)	-0.3511 (0.0484)	0.3674
$Age^2/100$	1.3350 (0.8122)	2.8694 (0.4692)	0.5347	1.3350 (0.8122)	2.8694 (0.4692)	0.3844
Education	0.1403 (0.0764)	0.1338 (0.0563)	0.0486	0.1403 (0.0764)	0.1338 (0.0563)	0.1520
High school diploma	-0.0762 (0.3236)	-0.3263 (0.2981)	0.7665	-0.0762 (0.3236)	-0.3263 (0.2981)	0.1192
College diploma	0.6676 (0.2308)	0.9465 (0.2761)	0.2946	0.6676 (0.2308)	0.9465 (0.2761)	0.3700
Professional experience	0.0865 (0.1731)	0.0444 (0.1797)	0.9482	0.0865 (0.1731)	0.0444 (0.1797)	0.0536
Managerial experience	0.0223 (0.0903)	0.0827 (0.1555)	0.7304	0.0223 (0.0903)	0.0827 (0.1555)	0.7303
Sales experience	0.1072 (0.1016)	0.0814 (0.1502)	0.3171	0.1072 (0.1016)	0.0814 (0.1502)	0.3171
Clerical experience	-0.0090 (0.1083)	0.0779 (0.1501)	1.1155	-0.0090 (0.1083)	0.0779 (0.1501)	0.1556
Craftsmen experience	0.1471 (0.0948)	0.3264 (0.1370)	0.5495	0.1471 (0.0948)	0.3264 (0.1370)	0.4309
Operatives experience	0.0325 (0.0869)	0.1358 (0.1214)	0.7607	0.0325 (0.0869)	0.1358 (0.1214)	0.6347
Laborer experience	-0.0951 (0.1618)	0.0596 (0.1700)	2.5956	-0.0951 (0.1618)	0.0596 (0.1700)	0.9400
Service experience	-0.3775 (0.0972)	-0.4288 (0.1928)	0.1198	-0.3775 (0.0972)	-0.4288 (0.1928)	0.0543
Previously a professional	1.312 (0.476)	1.934 (0.599)	0.3216	1.312 (0.476)	0.1.934 (0.0599)	0.2832
Previously a manager	1.837 (0.393)	2.194 (0.735)	0.1629	1.837 (0.393)	2.194 (0.735)	0.2739
Previously sales	3.262 (0.411)	2.869 (0.544)	0.1372	3.262 (0.411)	2.869 (0.544)	0.2087
Previously clerical	2.005 (0.388)	1.864 (0.0427)	0.0755	2.005 (0.388)	1.864 (0.427)	0.0245
Previously a craftsman	1.358 (0.407)	1.778 (0.573)	0.2361	1.358 (0.407)	1.778 (0.573)	0.2439
Previously an operative	1.272 (0.361)	1.049 (0.457)	0.2122	1.272 (0.361)	1.049 (0.457)	0.2529
Previously a laborer	1.358 (0.457)	1.015 (0.545)	0.3369	1.358 (0.457)	1.015 (0.545)	0.1852

Note: Standard errors in parentheses.

Table 6 Panel B: Parameter Estimates – Non-pecuniary Utility

	Craftsmen		% absolute deviation	*Operatives*		% absolute deviation
	Ignore classification error	Correct for classification error		Ignore classification error	Correct for classification error	
Age	-0.1896 (0.0717)	-0.2998 (0.0799)	0.4007	-0.1519 (0.0551)	-0.2535 (0.0558)	0.3115
$Age^2/100$	1.1693 (0.5598)	1.8996 (0.6139)	0.3291	1.3557 (0.4459)	2.0207 (0.4634)	0.2851
Education	0.1443 (0.0638)	0.1253 (0.0678)	0.1947	-0.0703 (0.0479)	-0.0873 (0.0422)	0.1682
High school diploma	0.2760 (0.2437)	0.2466 (0.1995)	0.0148	0.1959 (0.1839)	0.1931 (0.1680)	0.0537
College diploma	0.5009 (0.2163)	0.7951 (0.2838)	0.1359	-0.4700 (0.2633)	-0.4137 (0.3614)	20.7143
Professional experience	0.1874 (0.1529)	0.1779 (0.1211)	0.0530	0.1858 (0.1581)	0.1962 (0.1387)	1.4746
Managerial experience	0.0188 (0.0762)	0.0697 (0.1283)	0.6004	-0.1568 (0.0753)	-0.0980 (0.1321)	0.0020
Sales experience	-0.1264 (0.1093)	-0.1851 (0.1752)	0.2889	-0.2418 (0.1272)	-0.3401 (0.1816)	0.1798
Clerical experience	0.3591 (0.0857)	0.4253 (0.1258)	0.3213	-0.1887 (0.0887)	-0.1428 (0.1439)	0.7116
Craftsmen experience	0.1197 (0.0637)	0.2104 (0.1293)	0.2471	0.3067 (0.0520)	0.4074 (0.1172)	0.5117
Operatives experience	0.0786 (0.0707)	0.2151 (0.1136)	0.6869	0.0571 (0.0552)	0.1824 (0.1010)	0.3475
Laborer experience	0.0089 (0.1066)	0.1485 (0.1601)	0.8194	0.0430 (0.0848)	0.2381 (0.1449)	0.8217
Service experience	-0.3782 (0.0623)	-0.3587 (0.0787)	0.0990	-0.4665 (0.0448)	-0.5178 (0.0697)	0.0722
Previously a professional	1.338 (0.380)	1.866 (0.478)	0.1591	0.1124 (0.0394)	1.337 (0.526)	0.2965
Previously a manager	1.477 (0.325)	2.034 (0.653)	0.2778	0.1527 (0.0312)	2.115 (0.651)	0.3681
Previously sales	1.710 (0.457)	1.415 (0.618)	0.3348	0.1413 (0.0489)	1.059 (0.536)	0.3292
Previously clerical	2.804 (0.301)	2.874 (0.097)	0.0275	0.1198 (0.0333)	1.166 (0.324)	0.0207
Previously a craftsman	1.105 (0.307)	1.462 (0.492)	0.1381	0.2903 (0.0195)	3.368 (0.368)	0.1756
Previously an operative	0.763 (0.280)	0.609 (0.416)	0.0752	0.1521 (0.0195)	1.415 (0.294)	0.0477
Previously a laborer	1.672 (0.286)	1.411 (0.399)	0.2470	0.1636 (0.0231)	1.312 (0.329)	0.2746

Note: Standard errors in parentheses.

Table 6 Panel B: Parameter Estimates – Non-pecuniary Utility

	Laborers		% absolute deviation
	Ignore classification error	Correct for classification error	
Age	-0.2017 (0.0634)	-0.3403 (0.0650)	0.4072
$Age^2/100$	1.8105 (0.5104)	2.7642 (0.5240)	0.3450
Education	-0.1514 (0.0613)	-0.1099 (0.0545)	0.3774
High school diploma	0.2912 (0.2274)	0.1422 (0.2124)	1.0480
College diploma	0.0821 (0.3341)	0.3030 (0.3584)	0.7291
Professional experience	-0.4791 (0.2656)	-0.3477 (0.4226)	0.3778
Managerial experience	-0.2364 (0.1162)	-0.3256 (0.1955)	0.2739
Sales experience	-0.2337 (0.1279)	-0.2623 (0.1937)	0.1093
Clerical experience	0.0255 (0.0883)	0.0713 (0.1468)	0.6424
Craftsmen experience	0.0943 (0.0594)	0.1882 (0.1207)	0.4990
Operatives experience	0.0370 (0.0575)	0.1673 (0.1032)	0.7788
Laborer experience	0.3250 (0.0910)	0.4753 (0.1501)	0.3163
Service experience	-0.4093 (0.0654)	-0.4625 (0.0884)	0.1150
Previously a professional	0.943 (0.484)	0.175 (0.695)	0.1975
Previously a manager	0.609 (0.400)	0.129 (0.776)	0.4604
Previously sales	0.604 (0.699)	0.754 (0.707)	0.1989
Previously clerical	1.310 (0.354)	1.322 (0.351)	0.0096
Previously a craftsman	1.525 (0.240)	1.832 (0.435)	0.1673
Previously an operative	1.139 (0.204)	0.976 (0.311)	0.1675
Previously a laborer	1.870 (0.213)	1.579 (0.331)	0.1846

Note: Standard errors in parentheses.

Table 6 Panel C: Parameter Estimates – Unobserved Heterogeneity: Classification Error Model

	Type 1		Type 2		Type 3	
Non-pecuniary intercepts	Parameter	Std. error	Parameter	Std. error	Parameter	Std. error
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	.032				
Pr(Type 2)	0.3675	.041				
Pr(Type 3)	.5109	.042				

Table 6 Panel C: Parameter Estimates – Unobserved Heterogeneity: Model that Ignores Classification Error

	Type 1		Type 2		Type 3	
Non-pecuniary intercepts	Parameter	Std. error	Parameter	Std. error	Parameter	Std. error
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	.033				
Pr(Type 2)	0.5030	.039				
Pr(Type 3)	.4514	.040				

Table 6 Panel C: Average Wage & Non-pecuniary Intercepts Across Types

Average non-pecuniary intercepts (ϕ's)	Ignore classification error	Correct for classification error	% absolute deviation
Professional	-2.890	-3.546	0.184
Managers	-2.103	-2.650	0.206
Sales	-1.832	-1.080	0.696
Clerical	-1.534	-1.138	0.347
Craftsmen	0.048	0.234	0.795
Operatives	0.142	0.359	0.603
Laborers	0.278	0.800	0.652
Average wage intercepts (μ's)			
Professional	1.309	1.506	0.130
Managers	1.187	1.388	0.144
Sales	0.298	0.318	0.062
Clerical	1.325	1.378	0.038
Craftsmen	1.556	1.705	0.087
Operatives	1.535	1.711	0.102
Laborers	1.516	1.525	0.005
Service	1.173	0.5734	1.046

Note: Averages computed across types.

Table 7: Completed Education by Observed Occupation for Correctly Classified and Misclassified Occupational Choices

Observed Occupation in NLSY Data		**% No College Completed**	**% 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% of correctly classified professionals graduated from college, while only 30.2% of those incorrectly classified as professionals graduated from college.

Table 8: Distribution of Total Number of Times a Person's Occupational Choices are Misclassified Over the Career

# of times misclassified	Subpopulation 2	Subpopulation 3	All
0	32.8%	30.4%	57.2%
1	28.4%	27.0%	17.6%
2	18.8%	17.4%	11.5%
3	9.9%	10.1%	6.3%
4	4.2%	5.9%	3.0%
5	3.1%	4.4%	2.2%
6-9	2.5%	4.4%	1.9%
>9	.37%	.26%	.20%

Entries are the frequencies of the number of times that a person's occupational choices are misclassified over the course of the career based on the simulated data. For example, in Subpopulation 2, 9.9% of individuals in the simulated data have their occupational choices misclassified three times over the course of their career.

Table 9: Distribution of Lengths of Misclassification Spells

# of consecutive times misclassified	Subpopulation 2	Subpopulation 3	All
1	73.2%	72.2%	72.9%
2	18.2%	18.5%	18.3%
3	4.9%	5.8%	5.2%
4	2.2%	.9%	1.8%
5	.47%	1.3%	.7%
>5	.83%	1.1%	.93%

Entries are the frequencies of the number of consecutive times that a person's occupational choices are misclassified. For example, in subpopulation 2, conditional on having an occupational choice misclassified, 18.2% of these choices are misclassified for two consecutive survey observations.

Table 10: Average True Choice Probabilities by Observed Choice and Wage Percentile

Observed/Actual		Professional	Managers	Sales	Clerical	Craftsmen	Operatives	Laborers	Service
	Top 10%	.909	.000	.074	.000	.000	.000	.004	.011
Professional	Middle 10%	.953	.001	.006	.000	.013	.000	.012	.014
	Bottom 10%	.757	.001	.053	.001	.020	.001	.070	.095
	Top 10%	.052	.565	.374	.001	.002	.000	.000	.005
Managers	Middle 10%	.020	.858	.067	.002	.003	.002	.001	.046
	Bottom 10%	.010	.544	.272	.004	.004	.005	.012	.148
	Top 10%	.039	.033	.916	.017	.000	.000	.000	.038
Sales	Middle 10%	.033	.018	.911	.000	.000	.002	.002	.026
	Bottom 10%	.004	.005	.834	.016	.000	.007	.007	.127
	Top 10%	.039	.005	.916	.001	.000	.000	.000	.038
Clerical	Middle 10%	.033	.017	.911	.008	.000	.002	.002	.026
	Bottom 10%	.004	.005	.834	.016	.000	.007	.007	.127
	Top 10%	.031	.001	.091	.000	.872	.000	.003	.000
Craftsmen	Middle 10%	.008	.000	.015	.000	.965	.002	.007	.000
	Bottom 10%	.005	.000	.124	.003	.818	.005	.041	.002
	Top 10%	.084	.000	.110	.000	.000	.801	.003	.000
Operatives	Middle 10%	.009	.000	.008	.000	.000	.979	.003	.000
	Bottom 10%	.003	.000	.119	.000	.000	.869	.006	.000
	Top 10%	.000	.000	.065	.012	.032	.002	.885	.003
Laborers	Middle 10%	.000	.000	.003	.007	.008	.003	.976	.001
	Bottom 10%	.000	.000	.071	.004	.003	.002	.915	.004
	Top 10%	.054	.004	.072	.000	.000	.000	.150	.719
Service	Middle 10%	.005	.001	.001	.001	.000	.000	.251	.732
	Bottom 10%	.000	.000	.100	.000	.000	.000	.174	.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% refer to the location of the observed wage in the wage distribution of the observed occupation.

References

[1] Abowd, John and Arnold Zellner (1985). "Estimating Gross Labor Force Flows." *Journal of Business and Economic Statistics*, v. 3: 254-283.

[2] Abrevaya, Jason and Jerry Hausman (1999). "Semiparametric Estimation with Mismeasured Dependant Variables: An Application to Duration Models for Unemployment Spells." *Annales D'Economie Et De Statistique*, v. 55: 243-275.

[3] Bollinger, Chris (1996). "Bounding Mean Regressions when a Binary Regressor is Mismeasured." *Journal of Econometrics*, v. 73: 387-399.

[4] Bound, John, Charles Brown, and Nancy Mathiowetz (2001). "Measurement Error in Survey Data." In Handbook of Econometrics, edited by E. Learner and J. Heckman. Pp. 3705-3843. New York: North Holland Publishing.

[5] Chen, Xiaohong, Han Hong, and Elie Tamer (2005). "Measurement Error Models with Auxiliary Data." *The Review of Economic Studies,* v. 72: 343-366.

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

[7] Douglas, Stratford, Karen Smith Conway, and Gary Ferrier (1995). "A Switching Frontier Model for Imperfect Sample Separation Information: with an Application to Constrained Labor Supply." *International Economic Review*, v. 36 no. 2, pp. 503-526.

[8] Dustmann, Christian and Arthur van Soest (2001). "Language Fluency and Earnings: Estimation with Misclassified Language Indicators." *The Review of Economics and Statistics,* v. 83, no. 4: 663-674.

[9] Geweke, John (1991). "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints." *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, pp. 571-578.

[10] Gould, Eric (2002). "Rising Wage Inequality, Comparative Advantage, and the Growing Importance of General Skills In the United States." *Journal of Labor Economics*, v. 20, no. 1: pp. 105-147.

[11] Hajivassiliou, Vassilis (1990). "Smooth Simulation Estimation of Panel Data LDV Models." Manuscript, Yale University.

[12] Hausman, Jerry, Jason Abrevaya, and Fiona Scott-Morton (1998). "Misclassification of the Dependant Variable in a Discrete-Response Setting." *Journal of Econometrics*, v. 87.

[13] 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,* v. 93: 1077-1125.

[14] Heckman, James, and Guilherme Sedlacek (1990). "Self-Selection and the Distribution of Hourly Wages." *Journal of Labor Economics*, v. 8, no. 1: S329-S363.

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

[16] Kambourov, Gueorgui and Iorii Manovskii (2006). "Occupational Specificity of Human Capital." Working Paper.

[17] Keane, Michael (1994). "A Computationally Practical Simulation Estimator for Panel Data." *Econometrica*, v. 62, no. 1: 95-116.

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

[19] Kreider, Brent and John Pepper (2004A). "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors." Working paper.

[20] Kreider, Brent and John Pepper (2004B). "Inferring Disability Status from Corrupt Data." Working paper.

[21] Krueger, Alan and Lawrence Summers (1988). "Efficiency Wages and Inter-Industry Wage Structure." *Econometrica* v. 56, no. 2: 259-293.

[22] Lavy, Victor, Michael Palumbo and Steven Stern (1998). "Simulation of Multinomial Probit Probabilities and Imputation of Missing Data." In Advances in Econometrics, edited by T. Fomby and C. Hill. JAI Press.

[23] 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*, v. 31, no. 4: 514-544.

[24] Magnac, Thierry and Michael Visser (1999). "Transition Models with Measurement Errors." *The Review of Economics and Statistics*, v. 81, no. 3: 466-474.

[25] Mathiowetz, Nancy (1992). "Errors in Reports of Occupation." *The Public Opinion Quarterly*, v. 56, no. 3: 352-355.

[26] McCall, Brian (1990). "Occupational Matching: A Test of Sorts." *Journal of Political Economy*, v. 98: 45-69.

[27] Mellow, Wesley and Hal Sider (1983). "Accuracy of Response in Labor Market Surveys: Evidence and Implications." *Journal of Labor Economics*, v. 1, no. 4: 331-344.

[28] 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*, v. 92, no. 2: 233-274.

[29] Neal, Derek (1995). "Industry Specific Human Capital: Evidence from Displaced Workers." *Journal of Labor Economics*, v. 13, no. 4.

[30] Neal, Derek (1999). "The Complexity of Job Mobility Among Young Men." *Journal of Labor Economics*, v. 17, No. 2.

[31] Parent, Daniel (2000). "Industry Specific Capital and the Wage Profile: Evidence from the National Longitudinal Survey of Youth and the Panel Study of Income Dynamics." *Journal of Labor Economics*, v. 18, no. 2.

[32] 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*, v. 77, no. 2: 207-216.

[33] Ramalho, Esmeralda (2002). "Regression models for choice-based samples with misclassification in the response variable." *Journal of Econometrics*, v. 106: 171-201.

[34] Roy, Andrew (1951). "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers*, v. 3: 135-146.

[35] Stinebrickner, Todd (1999). "Estimation of a Duration Model in the Presence of Missing Data." *The Review of Economics and Statistics*, v. 81, no. 3: 529-542.

[36] Stinebrickner, Ralph and Todd Stinebrickner (2004). "Time Use and College Outcomes." *Journal of Econometrics* v. 121: 243-269.