

Measurement Error in Survey Data: An Application to a Model of Job Mobility in Ireland

By Adele Bergin¹

*(National University of Ireland, Maynooth and Economic and Social Research
Institute, Dublin)*

Paper Prepared for the 12th IZA European Summer School

This Draft: April 19th, 2009

Abstract

Many studies of labour market dynamics use survey data so it is valuable to know about the quality of the data collected. This paper investigates job mobility in Ireland over the period 1995 to 2001, using the Living in Ireland Survey, the Irish component of the European Community Household Panel. As is common with many Surveys, there is no direct question about changing jobs; instead mobility is inferred from responses of individuals to a question about tenure on two adjacent interview dates. The paper finds that there is substantial inconsistencies or measurement error in the responses used to determine job changes so there is a risk of misclassifying cases as being job changes when truly no change took place and vice versa. The extent of measurement error is similar to what has been found in other studies (e.g. Brown and Light, 1992). Any technique used to estimate the determinants of job change should control for misclassification error, otherwise it can lead to estimates that are biased and inconsistent. A procedure proposed by Hausman, Abrevaya and Scott-Morton (1998) is used to control for misclassification. The paper finds that ignoring misclassification may substantially underestimate the true number of job changes. In addition, ignoring misclassification leads to diminished covariate effects in models of job change.

JEL classification: J62

Keywords: Job Mobility; Binary choice model; Response Error

¹ Email: Adele.Bergin@esri.ie

1. Introduction

Many studies of labour market dynamics use survey data. Therefore it is valuable to know about the quality of the data collected. There may be ambiguity in a survey question, respondents may misunderstand the question, they may have poor recall or responses may be coded incorrectly. This paper investigates job mobility or employment-to-employment transitions in Ireland over the period 1995 to 2001 using the Living in Ireland Survey (LIS), the Irish component of the European Community Household Panel (ECHP). There is no direct question in the LIS about job mobility; instead it is inferred from responses of individuals on two adjacent interview dates. The paper finds that there is substantial measurement error in the data which may lead to misclassifying people who have not changed jobs as having changed jobs and vice versa. Any technique used to estimate the determinants of job mobility should control for misclassification error, otherwise it can lead to estimates that are biased and inconsistent. A procedure proposed by Hausman, Abrevaya and Scott-Morton (1998) is used to control for misclassification in the dependent variable.

This paper is organised as follows: Section 2 examines the reasons for and prevalence of reporting errors in labour market survey data and, in particular, focuses on studies relevant to job mobility. Section 3 explores the extent of measurement error in the LIS data. Section 4 outlines the estimation technique used to control for misclassification. Section 5 presents a simple extension of the estimation technique that allows for covariate-dependent measurement error. Section 6 provides estimation results and Section 7 concludes.

2. Labour Market Survey Data

Most studies of job mobility use survey data and usually surveys do not contain a direct question asking if the respondent has changed jobs in the past year. Instead job changes are inferred from the length of time an employee reports to have been with their current employer. Therefore questions about tenure play a crucial role in most empirical studies of job mobility. There are several reasons to suspect that responses to questions about tenure are measured with error. Respondents may find it difficult to remember when they started working in their current job. Bound *et al.* (2001) describe studies that categorise the question and answer process in a survey as a four-stage procedure. These stages include understanding the question, recovering the information from memory, considering whether the information matches what was requested and communicating the response. Much of the measurement error literature focuses on the stage where respondents retrieve the information from memory. A general principle from this literature is that the longer the length of the recall period the greater the expected bias due to respondent retrieval error. Therefore we might expect respondents with longer tenure to be most likely to misreport tenure. In one sense, this does not pose a serious problem for calculating job changes as job changes are associated with people who have short tenures; provided those with longer tenures who misreport do not significantly underestimate their tenure. Farber (1999) and Ureta (1992) find a heaping of tenure responses at round counts of years or round calendar years and this rounding indicates that individuals do not provide precise responses about tenure.

There may also be ambiguity in the wording of the question about tenure or there may be changes to the wording of the question in other waves. Farber (1999) points out how the mobility supplements to the Current Population Survey in the US from 1951

to 1981 asked workers what year they “...started working at their present job or business” while in later years the supplement asked workers how many years they have “...been working continuously for the present employer”. The earlier question refers to time on the present job rather than time with the present employer. Workers may experience other types of internal labour mobility (e.g. promotion, reassignment) which means that their tenure on the job will be shorter than tenure with the employer. The interviewer notes for the LIS provide clarity in distinguishing between employer changes and other types of internal labour mobility as they state that the question refers to when they started working with their present employer even if there have been position changes with that employer. In addition, there were no changes to the wording of the question about tenure in the LIS. The interviewer notes in the LIS do not provide guidance on how to handle interrupted employment spells (in particular when someone returns to a previous employer). Farber (1999) mentions that if no reference is made to the continuity of employment that the natural inclination of workers will be to ignore interruptions of “reasonable” length.

Brown and Light (1992) examine the extent of measurement error in tenure responses in the Panel Study of Income Dynamics (PSID). They find that tenure responses are frequently inconsistent with calendar time.² In addition, they perform a validation exercise to gauge the accuracy of their measure of job changes. They adopt various definitions of job mobility (based on tenure responses) and use them to partition the data into distinct jobs. They assess the accuracy of the various definitions by comparing the number of jobs and the number of times each job is observed with those identified by the National Longitudinal Survey (NLS). The NLS contains unique employer codes which can be compared across interviews and so provides a more accurate count of the ‘true’ number of jobs.

Brown and Light (1992) investigate various measures of job mobility and examine which one performs best when there is measurement error in tenure data. One definition of job mobility they employ is to assume that a job change has taken place whenever reported tenure is less than the time elapsed since the previous interview. If tenure was never misreported and if respondents never returned to previous employers then this method would identify job changes without error. They also adopt another set of definitions of job mobility by assuming that a job change occurs whenever the change in tenure between adjacent interviews varies “too much” or “too little” in either direction. In one definition a job change is defined whenever the change in tenure is not exactly equal to the change in calendar time between interviews. This permits no inconsistency in tenure responses within jobs. They also adopt more flexible measures that permit various amounts of inconsistency in reported tenure within jobs. They define another four measures of job mobility when the change in tenure differs from the change in calendar time by more than 6, 12, 18 and 24 months in either direction. As these later definitions define job changes when tenure changes by “too much” as well as by “too little” they are more likely to separate continuing jobs³ but less likely to link jobs that are truly separate⁴ than when job changes are

² The level of inconsistencies in reported tenure in the PSID is described in Section 3.2 where comparisons are made to the LIS data.

³ For example, consider an individual who truly hasn’t changed jobs and who reports tenure of 24 months in one interview and exactly a year later misreports their tenure and says they have been in their job for 45 months when their true tenure is 36 months. When we define a job change as having occurred when reported tenure is less than the time between interviews then we conclude this person

defined as occurring whenever reported tenure is less than the time elapsed since the previous interview. The definition of job mobility that is the most accurate when compared to the NLS data is that a job change has occurred whenever reported tenure is less than the time elapsed between interviews. This is essentially the definition of job change that is adopted in this study.

These types of validation studies are also useful because they provide evidence on the magnitude of the measurement error in tenure data. Bound *et al.* (2001) point to the fact that few studies have investigated the quality of tenure data. Duncan and Hill (1985) present results from a validation study of a large manufacturing company in which administrative records are used to validate survey responses from a sample of workers from the company. Overall they find very little evidence of bias in the interview reports. They find that reported tenure is typically quite accurate; 45 per cent of the sample accurately reported the year they were hired and 90 per cent were able to report year of hire accurately within one year. However, the unit of analysis in the study is defined in terms of years and these types of error margins in a dataset could be problematic if we were to use the measure of tenure to calculate job mobility. As job changes are identified from those who report short tenures the under or over reporting of tenure by a year, in particular by those with short tenures, could lead us to misclassify job changes and vice versa.

Bound *et al.* also cite a study where workers' reported starting dates are compared to employer records. The study by Weiss *et al.* (1961) finds that 71 per cent of jobs in the prior 5 years had reported starting dates within one month of company records. They also find that validity significantly declines as a function of the length of time between the start date and the date of interview. To capture job mobility, tenure, at least for those who have not been in their jobs long, needs to be reported accurately. These validation studies suggest that the quality of tenure data may not be sufficient to do this.

3. Measurement Error

3.1 Dataset and Defining Job Changes

The Living in Ireland Survey (LIS) is used to investigate the determinants of job change. The LIS constitutes the Irish component of the European Community Household Panel (ECHP) which began in 1994 and ended in 2001. It involved an annual survey of a representative sample of private households and individuals aged 16 years and over in each EU member state, based on a standardised questionnaire. A wide range of information on variables such as labour force status, occupation, income and education level is collected. There is also a wealth of data collected on job and firm characteristics.⁵

hasn't changed jobs. However, if we define job mobility as occurring when the change in reported tenure differs from the change in calendar time by more than, say, 12 months then we classify this person as having changed jobs.

⁴ For example, consider an individual who truly has changed jobs and is interviewed 12 months apart. In the first year they report tenure of 5 months and in the subsequent year they report tenure of one month. When we define a job change as having occurred when reported tenure is less than the time between interviews then we conclude this person has changed jobs. Using the other definitions of job mobility we would not classify this person as having changed jobs.

⁵ There was some attrition in the sample in the earlier years, although the representativeness of the sample was improved in 2000 with the addition of new households. These new entrants to the LIS sample have been excluded from the analysis.

To identify those who have changed jobs I make use of the panel dimension of the LIS. A revolving balanced panel of people aged 20 to 60, roughly the prime working age, has been selected from the LIS. This means that individuals are included in the sample in every year they meet this age restriction.⁶ In addition, they must also have completed the interview in each year in question. The reason for choosing a revolving balanced panel over a pure balanced panel is that a balanced panel prevents the entry of younger people into the sample and so over time as the fixed sample ages the proportion of younger people would decline.⁷ Effectively, a revolving balanced panel allows younger people into the sample in later years and lets older people drop out.

There is no explicit question in the LIS about whether or not a person has changed jobs; instead job mobility is inferred from responses to the question about when they started working with their present employer. If a person is employed in two consecutive years and in the second year they report a starting date that falls between the two interview dates we conclude that this person has changed jobs during that period. However, in the absence of exogenous job change information we cannot be certain that this person has changed jobs. For example, a worker may forget their starting date, misunderstand the question or their response could be coded incorrectly. In addition, respondents may consider multiple spells with the same employer differently. These problems could be overcome if the LIS contained a direct question about job mobility or if contained unique employer codes that could be compared across interviews.

Job mobility is defined in terms of employment-to-employment transitions. To capture this in the data workers need to be employed in two consecutive years. For example, someone who is employed in one year and then unemployed for two years and then employed again is not included in the analysis. Even though this person has moved to a new job over the four-year period, they have moved from being employed to being unemployed for two years to being employed again. Restricting the sample to people who are employed in consecutive two-year periods means that this type of case is excluded. I have excluded these types of transitions because the decision to change jobs is different to the decision to move from, say nonparticipation or unemployment to employment. This definition of job mobility only allows people to be unemployed or to not participate in the labour market for a relatively short amount of time between jobs, essentially less than a year (or more precisely less than the amount of time between interviews). In addition, this measure of job mobility may underestimate total job mobility if more than one job change occurs between subsequent interviews. Farber (1999) states that one of the central facts about job mobility is that there is a high hazard of jobs ending within the first year of an employment relationship.

3.2 Consistency of Starting Dates within Jobs

Given the possibility of measurement error or recall error in the data we need to try to ascertain how reliable the information on starting dates is and therefore how useful it is for deducing job changes. If there were no measurement error in the data then

⁶ This approach to selecting a sample is similar to that of Baker and Solon (1999).

⁷ For example, someone who is 20 in 1995 will be 26 in 2001 and if we only considered the same group of people over time (a balanced panel), there would be no one below the age of 26 in the panel by 2001.

starting dates would be constant *within* jobs. By partitioning the dataset into distinct jobs and comparing starting dates across interviews we can investigate how consistent the data is. Table 1 shows the number of workers employed in consecutive two-year periods from the revolving balanced panel and the number of job changes each year.

Table 1: Number of Workers and Job Changes

	1995	1996	1997	1998	1999	2000	2001
Number of workers	1,185	1,229	1,274	1,337	1,406	1,449	1,497
No. Job Changes	77	89	110	146	151	195	159

To convert this data into separate jobs, we begin with the 1995 data. There are 1,185 workers in 1995 but as 77 workers changed jobs a total of 1,262 distinct jobs are observable in that year. For this year alone, the previous jobs of those who changed jobs are excluded from the analysis. We only have one observation on their previous jobs (the starting date in 1994) so we cannot check the consistency of responses whereas we can track the new jobs across subsequent interviews. Therefore, we start the analysis with 1,185 distinct jobs in 1995. In each subsequent year one of four alternatives occurs:

- 1) A worker can stay in their job so the total number of jobs remains the same and we observe the job surviving an additional year.
- 2) A worker can drop out of the sample if they enter a period of unemployment, leave the labour force for more than a year or if they are over the age of 60. In this case, the total number of jobs remains the same but we no longer observe that particular job. Workers who are unemployed or leave the labour force may re-enter the analysis in later years.
- 3) A worker can change jobs and accordingly the total number of jobs increases by one and we stop observing the previous job.
- 4) There can be a new entrant to the sample of workers. This would be someone from the revolving balanced panel who is now 20 and so was excluded in earlier years. This increases the total number of jobs by one. In addition, a worker who was unemployed or out of the labour force may come back into the analysis and this would increase the total number of jobs observed by one.

This results in 2,672 jobs observable for various durations over the period 1995 to 2001. Table 2 shows how many jobs display consistency in starting dates. Of the 2,672 jobs identified, only 301 or 11 per cent have the same reported starting date in each year the job is observed. If we adopt a less stringent definition of consistency such as all starting dates being within 3 months of each other then 27 per cent of jobs meet the criterion. If we relax the criterion further to consider jobs where all starting dates are within 6 months of each other then 31 per cent of jobs display consistent responses. This leaves 1,832 or 69 per cent of jobs where all starting dates do not fall within 6 months of each other.⁸

⁸ The method for partitioning the dataset into distinct jobs uses job changes to identify when one job ends and another one begins. The analysis presented in this section implies that the measure of job change may not accurately identify the true number of job changes i.e. there are probably cases identified as job changes when no change in jobs took place and vice versa. This means we may over or underestimate the true number of jobs and therefore the level of inconsistent starting dates within jobs. For example, consider a person who truly hasn't changed jobs and started working in their current job in January 1993 and in 1995 they report their true starting date but in 1996 they report that they

Table 2: Consistency of Starting Dates within Jobs

Source: Living in Ireland Survey

	All Jobs	Jobs with All Starting Dates			
		Equal	Within 3 months	Within 6 months	Remaining jobs
Number of jobs	2,672	301	715	840	1,832
% of Jobs		11%	27%	31%	69%

Source: Panel Study of Income Dynamics, taken from Brown and Light (1992)

Number of jobs	3,318	246	1,170	1,514	1,804
% of Jobs		7%	35%	46%	54%

This level of inconsistency in starting dates is quite alarming; however it is in line with what has been found in other datasets. Brown and Light (1992) take a sample from the Panel Survey of Income Dynamics (PSID) from 1976 to 1985, partition the data into distinct jobs in an analogous fashion and examine the consistency of reported start dates within jobs. They find that only 7 per cent of the jobs in their sample have identical starting dates in each year the job is observed, while 54 per cent of jobs have starting dates that do not all fall within 6 months of each other.

Brown and Light (1992) point out another aspect of this definition of consistency that may be quite restrictive. Suppose a job is observed every year of the LIS and every starting date given between 1994 and 2001 is equal with the exception of one which is different to the others by 7 months, then this job will not meet any of the measures of consistency define above. They argue most researchers would agree that this outlier could be ‘fixed’ to match up to the other observations for that job. One can further extend the measure of consistency by requiring that only a majority of observations for a given job be in agreement. Table 3 shows how many jobs have a majority of starting dates in agreement. A total of 727 jobs or 27 per cent have a majority of starting dates in perfect agreement, while 64 per cent of all jobs identified have a majority of starting dates that are within 3 months of each other. The bottom panel of the Table reports comparable statistics for the PSID taken from Brown and Light (1992). As before, the magnitudes of the consistency measures are broadly comparable with the Irish data. Given that both datasets display similar discrepancies, it is likely that any study using a similar question to deduce job changes contains measurement error.

started working in January 1996. This case will be erroneously classified as a job change in 1996; two separate jobs will be identified for this person and no inconsistency in starting dates will be recorded. In addition, suppose this person truly changed jobs in January 1996 but they misreport the starting date to be January 1995. This will be recorded as one job with an inconsistent starting date in 1996 when it is truly two jobs with an incorrect starting date in 1996. As the number of job changes (and jobs) is measured imperfectly so too will the amount of inconsistencies evident in the data. If the definition of job change accurately allowed us to define jobs then Table 3 would show the true amount of inconsistencies within jobs. As the definition of job changes only provides us with an imperfect measure of the number of jobs the data presented in the Table reflects both the inconsistencies in starting dates and the fact that we have not accurately identified the true number of jobs.

Table 3: Consistency of the Majority of Starting Dates within Jobs

Source: Living in Ireland Survey

	All Jobs	Jobs with >1/2 Starting Dates			Remaining jobs
		Equal	Within 3 months	Within 6 months	
Number of jobs	2,672	727	1722	1,821	851
% of Jobs		27%	64%	68%	32%

Source: Panel Study of Income Dynamics, taken from Brown and Light (1992)

Number of jobs	3,318	676	2,116	2,471	847
% of Jobs		20%	64%	74%	26%

Tables 2 and 3 focus on the extent and magnitude of inconsistencies evident in the data and it is clear there is substantial measurement error. In this study, the main concern about measurement error is not directly that starting dates are misreported but rather that the misreporting of starting dates may cause cases to be misclassified as job changes and vice versa (that truly distinct jobs would be linked or continuing jobs would be separated).

There are cases where it is very unlikely that we will erroneously assign a case to be a job change, even though there are inconsistencies in starting dates. For example, suppose an individual gives the following starting dates in successive interviews for a job observed from 1995 to 1998: January 1980, February 1975, May 1972 and January 1982. It is unlikely that this person has changed jobs at any point between 1995 and 1998 but rather they found it difficult to recall the starting date of their job. Although the responses for this job would fail to meet any of the consistency measures described in Tables 2 and 3, all of the reported starting dates are sufficiently long ago so as to not cause much concern about erroneously defining recent job changes.

Of particular concern are inconsistencies in short jobs (i.e. where reported tenure is low). For example, suppose we observe a job every year between 1995 and 2001, it is more likely that this person has changed jobs at some point over this period and it has not been captured if the inconsistency in starting dates falls close to or within that period. In addition, suppose we observe a person with two jobs over the period 1995 to 2001, they are be classified as having changed jobs in one year but they may have just misreported their starting date in the year in question. If all inconsistencies in reported starting dates refer to a time period sufficiently far back then it is more likely that this person hasn't changed jobs and just cannot accurately recall when they started working in their current job.

Table 4 examines the timing of inconsistencies in reported starting dates within jobs. It takes the total number of jobs that display any inconsistency in reported starting dates (i.e. $2,672 - 301 = 2,371$) and reports how many of these jobs have the dates of all inconsistencies occurring at least three years prior to the date that we first start

observing the job.⁹ There are 852 jobs where all discrepancies fall reasonably far in the past that so that these are probably truly continuing jobs. However, there are 1,519 jobs where the reported inconsistencies are more recent and it is more likely in these cases that we have linked jobs that are distinct or divided continuing jobs.

Table 4: Timing of Inconsistencies within Jobs

	Total No. Jobs	Equal Starting Dates	All inconsistencies at least 3 years prior to date job is first observed	Remaining jobs
Number of jobs	2,672	301	852	1,519
% of Jobs		11%	32%	57%

If the true number of jobs has been under or overestimated (and therefore the true number of job changes has been under or overestimated), it may be more likely to do so for certain types of worker. As mentioned above, I am more likely to under or overestimate the number of jobs where reported inconsistencies are recent (i.e. jobs that report inconsistent short tenures). As tenure and years of work experience are correlated we may expect to see differences in inconsistencies in starting dates by years of experience. There may also be differences by gender; as women experience more interrupted employment spells than men it may be harder for them to accurately report starting dates.

There are 1,519 jobs (from Table 4) that exhibit some type of inconsistency in starting dates within jobs where the dates associated with any inconsistency are relatively recent. Table 5 examines these jobs by years of work experience and gender. I have labelled these jobs as being ‘problematic’ in the sense that it is more likely that that truly distinct jobs would be linked or continuing jobs would be separated (i.e. that job changes are misclassified and vice versa). This is not to say that workers in short jobs are more likely to be misreport when they started working in their job than workers in long jobs, but rather that the misreporting by those in short jobs is more likely to lead to misclassifying cases as job changes and vice versa. The Table shows that the incidence of problematic jobs declines with years of work experience.¹⁰ For example, 76 per cent of jobs that have less than ten years of work experience associated with them are classified as problematic and this percentage declines as years of experience increases so that only 31 per cent of jobs with more than 30 years of work experience associated with them fall into this category. As there are more of these problematic jobs in low experience categories and job mobility is negatively correlated with experience, this may indicate that we are underestimating the true number of job changes.

There is also some difference when we look at the incidence of these problematic jobs by gender; 53 per cent of all jobs held by men fall into this category while the

⁹ For example, if we observe for the first time in 1995 this measure counts all jobs where each inconsistency refers to dates earlier than or in 1992.

¹⁰ In assigning years of work experience to a worker in a job we use their experience in the first year that the job is observed. For example, if we observe a job each year between 1995 and 2001 the experience assigned to that job when comparing all combinations of starting dates over the period is the years of experience of that person in 1995.

comparable figure for women is almost 10 percentage points higher. We will return to the issue of misclassification varying by different types of worker in Section 5.

Table 5: ‘Problematic’ Jobs by Worker Characteristics

	<i>Years of Work Experience</i>				
	<u>Total</u>	<u><10</u>	<u>10-19</u>	<u>19-29</u>	<u>30+</u>
No. Problematic Jobs	1,519	716	427	236	140
Total Number of Jobs	2,672	944	768	515	445
Problematic jobs as % of total	57%	76%	56%	46%	31%

	<i>Gender</i>		
	<u>Total</u>	<u>Men</u>	<u>Women</u>
No. Problematic Jobs	1,519	865	654
Total Number of Jobs	2,672	1,619	1,053
Problematic jobs as % of total	57%	53%	62%

This section has focussed on examining discrepancies in reported starting dates within jobs. It used jobs as the unit of analysis for looking at measurement error. In the remainder of the paper, the focus will be on how measurement error may lead us to misclassify a worker in a given year as being a job changer and vice versa so the unit of analysis switches to workers. The fact that tenure is misreported does not necessarily affect the measure of job changes. Of the group of workers who misreport tenure those with shorter tenures are more likely to be misclassified as having changed jobs or not then workers with long tenures (unless they significantly underestimate their tenure).

4. Binary Choice Model with Misclassification

A binary choice model can be used to explain the decision to change jobs. Given the level of measurement error in the data, it is likely that incorrect inferences have been made about whether or not a worker has changed jobs so it is essential to control for misclassification. This section begins by outlining the effects of measurement error in a dependent variable in the classical linear regression model and how the effect of measurement error is more serious in a binary choice model. When the dependent variable is dichotomous, misclassification can lead to estimates that are biased and inconsistent. Then a procedure proposed by Hausman et al. (1998) to control for misclassification in the dependent variable in a discrete response model is outlined. This procedure has been used in various empirical applications including investigating fraudulent insurance claims (Artís et al., 2002), smoking data (Brachet, 2005 and Kenkel et al., 2004) and patent applications (Jensen et al., 2008).

4.1 Effect of Measurement Error in the Dependent Variable

In the classical linear regression model measurement error in the dependent variable does not have very serious consequences- the standard errors on coefficients will tend

to be larger than they would have been if there were no measurement error. Consider the following model:¹¹

$$\tilde{y}_i = \beta x_i + \varepsilon_i \quad i=1, \dots, n \text{ where } n \text{ is sample size,}$$

ε_i is i.i.d and all variables are measured as deviations
from sample means

Suppose that \tilde{y}_i is the true dependent variable but that it is measured with error so what we actually observe is:

$$y_i = \tilde{y}_i + v_i \quad \text{where } v_i \text{ is assumed to be independent of the covariates and } \varepsilon_i$$

so $y_i = \beta x_i + \omega_i \quad \text{where } \omega_i = \varepsilon_i + v_i$

The effect of measurement error in y_i is an error term with increased variance since the new error term, ω_i , contains both the original error term, ε_i , and the measurement error, v_i . In this case, the OLS estimates of β will remain unbiased but will be measured with less precision.

In a non-linear regression model, such as a probit model, the effects of measurement error are more severe. If the dependent variable is binary, measurement error takes the form of misclassification errors; some observations where the variable is truly a 1 will be misclassified as a 0 and vice versa. In this case, if the true value is 1(0) but it is misclassified as a 0(1) then the measurement error (observed minus true value of dependent variable) will be negative (positive). In this case the measurement error is negatively correlated with the true variable. This can lead to coefficient estimates that are biased and inconsistent.

4.2 Standard Model of Misclassification

The decision to change jobs can be set in the usual latent-variable specification of the binary choice model. The Hausman *et al.* (1998) model of misclassification can be used for this application as follows:¹²

Let y_i^* be the latent variable that represents the potential or tendency for a worker to change jobs. y_i^* is a continuous variable that is unobservable and is determined by a set of explanatory variables, x_i , in such a way that the larger the value of y_i^* , the greater the probability of changing jobs.

$$y_i^* = x_i \beta + \varepsilon_i \quad \text{where } i=1, 2 \dots n \quad n = \text{sample size} \quad (1)$$

and ε_i is an independently and identically distributed error term. We cannot observe the tendency for a worker to change jobs; instead in the usual binary choice model we observe whether a worker changes jobs or not so for each worker there is a threshold or critical level, y_i^* , at or above which they change jobs otherwise they stay in their

¹¹ Hausman (2001) discusses the effects of measurement error in dependent variables.

¹² The details of the model come from Hausman *et al.* (1998).

jobs. The true response (or what we would observe in the data if there was no measurement error), \tilde{y}_i , is given by:

$$\begin{aligned} \tilde{y}_i &= 1 && \text{if } y_i^* \geq 0 \\ &= 0 && \text{otherwise.} \end{aligned} \quad (2)$$

\tilde{y}_i represents actual job changes, however the classification of workers as having changed jobs or not is observed with error so the model needs to be extended to deal with the possibility of misclassification.

Let y_i denote observed job changes. Two types of misclassification can occur, namely a job change can be misclassified as a non-job change or a non-job change can be classified as a job change. Let the probability that a non-job change is misclassified as a job change is given by α_0 , while the probability that a job change is misclassified as a non-job change is given by α_1 . These probabilities depend on the true value, \tilde{y}_i , so the extent of misclassification depends on how good a proxy y_i is of \tilde{y}_i and the probabilities are assumed to be independent of the covariates, x_i , more formally:

$$\alpha_0 = \Pr(y_i = 1 | \tilde{y}_i = 0), \quad (3)$$

$$\alpha_1 = \Pr(y_i = 0 | \tilde{y}_i = 1). \quad (4)$$

The assumption that the probabilities of misclassification are independent of the covariates is important for identification and this will be discussed later on.

Let $F(\cdot)$ denote the cdf of ε_i . The probability that an observation is truly a job change is given by:

$$\begin{aligned} \Pr(\tilde{y}_i = 1 | x_i) &= \Pr(x_i' \beta + \varepsilon_i > 0) \\ &= \Pr(\varepsilon_i < x_i' \beta) && \text{(if } F(\cdot) \text{ is a symmetric distribution)} \\ &= F(x_i' \beta) \end{aligned} \quad (5)$$

while the probability that it is truly a non-job change is given by:

$$\Pr(\tilde{y}_i = 0 | x_i) = 1 - F(x_i' \beta) \quad (6)$$

As there is a α_0 probability that a non-job change has been incorrectly classified as a job change and a $1 - \alpha_1$ probability that a job change has been correctly classified as such, the probability that an observation has been classified as a job change is given by:

$$\Pr(y_i = 1 | x_i) = (1 - \alpha_1)F(x_i' \beta) + \alpha_0(1 - F(x_i' \beta)) = \alpha_0 + (1 - \alpha_0 - \alpha_1)F(x_i' \beta) \quad (7)$$

Likewise, the probability that an observation is classified as a non-job change is

$$\Pr(y_i = 0 | x_i) = \alpha_1 F(x_i' \beta) + (1 - \alpha_0)(1 - F(x_i' \beta)) = 1 - \alpha_0 - (1 - \alpha_0 - \alpha_1)F(x_i' \beta) \quad (8)$$

The expected value of the observed dependent variable y_i is given by:

$$E(y_i|x_i) = \Pr(y_i = 1|x_i) = \alpha_0 + (1 - \alpha_0 - \alpha_1)F(x_i'\beta) \quad (9)$$

When there is no misclassification ($\alpha_0 = \alpha_1 = 0$), this collapses to usual expression $F(x_i'\beta)$.

If we assume that ε_i are normally distributed then we can use equations (7) and (8) to derive the log-likelihood function for the probit model with misclassification:

$$\ln L = \sum_{i=1}^n \{y_i \ln(\Pr(y_i = 1|x_i)) + (1 - y_i) \ln(\Pr(y_i = 0|x_i))\} = \sum_{i=1}^n \{y_i \ln(\alpha_0 + (1 - \alpha_0 - \alpha_1)\Phi(x_i'\beta)) + (1 - y_i) \ln(1 - \alpha_0 - (1 - \alpha_0 - \alpha_1)\Phi(x_i'\beta))\} \quad (10)$$

where $\Phi(\cdot)$ denotes the cdf of the normal distribution

Maximising the log-likelihood function given in (10) with respect to α_0, α_1 and β yields consistent and efficient estimates of β as well as the probabilities of misclassification.

Identification

The conditions for identification of α_0, α_1 and β are similar to those for the traditional binary choice model. One additional assumption is needed for identification, namely that the misclassification probabilities are not very large, specifically, $\alpha_0 + \alpha_1 < 1$. This assumption is needed because the normal distribution is symmetric ($\Phi(x_i'\beta) = 1 - \Phi(-x_i'\beta)$) and we can define $\hat{\alpha}_0 = 1 - \alpha_1$, $\hat{\alpha}_1 = 1 - \alpha_0$ and $\hat{\beta} = -\beta$ so that:

$$\begin{aligned} \hat{\alpha}_0 + (1 - \hat{\alpha}_0 - \hat{\alpha}_1)\Phi(x_i'\hat{\beta}) &= (1 - \alpha_1) + ((1 - (1 - \alpha_1)) - (1 - \alpha_0))(1 - \Phi(-x_i'\hat{\beta})) \\ &= (1 - \alpha_1) + ((-1 + \alpha_0 + \alpha_1)(1 - \Phi(x_i'\beta))) = \alpha_0 + (1 - \alpha_0 - \alpha_1)\Phi(x_i'\beta) \end{aligned} \quad (11)$$

When the assumption, $\alpha_0 + \alpha_1 < 1$, is not imposed the maximum likelihood estimator cannot distinguish between the parameter values $(\alpha_0, \alpha_1, \beta)$ and $(1 - \alpha_0, 1 - \alpha_1, -\beta)$. The assumption that $\alpha_0 + \alpha_1 < 1$ excludes this situation because $\alpha_0 + \alpha_1 < 1$ implies $(1 - \alpha_1) + (1 - \alpha_0) > 1$. An implication of this assumption is that if $\alpha_0 + \alpha_1 > 1$ but we impose $\alpha_0 + \alpha_1 < 1$ the estimates of β will have the wrong sign. This assumption guarantees that $\alpha_0 + (1 - \alpha_0 - \alpha_1)\Phi(x_i'\beta)$ is strictly increasing in $x_i'\beta$ as $\Phi(\cdot)$ is strictly increasing.

The model parameters are identified from the nonlinearity of $\Phi(\cdot)$. To see this consider the linear probability model where $F(x_i'\beta) = (x_i'\beta)$, then the expected value of y_i is given by:

$$\begin{aligned} E(y_i|x_i) &= \Pr(y_i = 1|x_i) = \alpha_0 + (1 - \alpha_0 - \alpha_1)(x_i'\beta) \\ &= (\alpha_0 + \beta_0) + z_i((1 - \alpha_0 - \alpha_1)\beta_1) \end{aligned} \quad (12)$$

where $x_i = (1, z_i)'$ and $\beta = (\beta_0, \beta_1)'$ (i.e. separating out the constant)

In this case the parameters of the model cannot be separately identified.

Estimating α_0 and α_1 is only possible because they enter (10) additively and are then multiplied by the expression with the normal cdf. This can be seen more clearly by taking limits of $E(y_i|x_i)$ as $x_i'\beta$ tends to $-\infty$ and $+\infty$ in (9).

$$\lim_{x_i'\beta \rightarrow -\infty} E(y_i|x_i) = \alpha_0 \quad \text{and} \quad \lim_{x_i'\beta \rightarrow \infty} E(y_i|x_i) = 1 - \alpha_1 \quad (13)$$

To identify the misclassification probabilities, α_0 and α_1 , $x_i'\beta$ has to get reasonably large in magnitude so as to push $\Pr(\tilde{y}_i = 1|x_i)$ close to 0 and 1 for some i . The intuition behind this is that we have assumed that misclassification rates are constant and depend only on the true value, \tilde{y}_i , so the probability of misclassifying a non-job change as a job change, α_0 , is identified from the group of workers with a near zero probability of truly changing jobs. These are workers for whom $x_i'\beta$ is highly negative and who are therefore very unlikely to be job changers. However, as a constant proportion, α_0 , are misclassified as having changed jobs, $\Pr(y_i = 1|x_i)$ will never fall below α_0 regardless of how negative $x_i'\beta$ is. In a similar fashion, the probability of misclassifying a job change as a non-job change, α_1 , is estimated from the group of workers for whom $x_i'\beta$ is very large and so have a very high probability of truly changing jobs but may be misclassified as not having changed jobs. Therefore $\Pr(y_i = 1|x_i)$ will never rise above $1 - \alpha_1$.

Marginal Effects

From equation (5) we know that the expected value on the true response, \tilde{y}_i , is given by:

$$E(\tilde{y}_i|x_i) = \Pr(\tilde{y}_i = 1|x_i) = F(x_i'\beta) \quad (14)$$

In general, we are usually more interested in the marginal effect of a specific variable, k , which is given by:

$$\frac{\partial E(\tilde{y}_i|x_i)}{\partial x_{ik}} = \frac{\partial \Pr(\tilde{y}_i = 1|x_i)}{\partial x_{ik}} = f(x_i'\beta)\beta_k \quad (15)$$

Equation (9) gives the expected value on the observed response, y_i , and the marginal effect of a variable, k , is:

$$\frac{\partial E(y_i|x_i)}{\partial x_{ik}} = \frac{\partial \Pr(y_i = 1|x_i)}{\partial x_{ik}} = (1 - \alpha_0 - \alpha_1) f(x_i; \beta) \beta_k \quad (16)$$

Comparing equations (15) and (16) shows that when there is misclassification error the marginal effects of interest (from equation (16)) will be biased towards zero (as $(1 - \alpha_0 - \alpha_1) < 1$). The marginal effects on the observed response will always be less than the marginal effects on the true response by a factor of $(1 - \alpha_0 - \alpha_1)$. This result only holds when misclassification is independent of the covariates.

To see the intuition behind this result, consider the following simple example: suppose you have a sample of 20 people, 10 of whom have a high value of some characteristic that makes them more likely to change jobs and the remaining 10 people have a low value of this characteristic that makes them less likely to change jobs. Further suppose that 8 people from the first group and 4 from the second group are identified as job changers. Then the true marginal effect on the characteristic is 0.4 (0.8-0.4). Now further suppose that we introduce misclassification (that does not depend on the particular characteristic) such that 4 out of the 8 true non job changes are misclassified (i.e. $\alpha_0 = .5$) and 3 out of the 12 true job changes are misclassified (i.e. $\alpha_1 = .25$). As the misclassification probabilities are assumed not to depend on the characteristic, this implies that 1 job stayer from the first group and 3 from the second group are misclassified as job changers and 2 job changers from the first group and 1 from the second group are misclassified as job stayers. Then the marginal effect on the characteristic is 0.1 (0.7-0.6) which is a quarter or $(1 - \alpha_0 - \alpha_1)$ of the true marginal effect.

5. Covariate-Dependent Misclassification

The analysis in Section 4 focussed on the case where misclassification is independent of the covariates or where the probabilities of misclassification are constant across all workers. Assuming that the probabilities of misclassification are constant across all types of workers may be quite restrictive. As indicated in Section 3.2 (in particular in Table 5) it is likely that the probabilities of misclassification vary across different types of workers. Hausman *et al.* (1998) consider a simple extension to the model to allow for some covariate dependent misclassification error as follows:

Assume that the misclassification probabilities depends on some or all of the covariates x_i :

$$\alpha_0(x_i) = \Pr(y_i = 1 | \tilde{y}_i = 0, x_i) \quad (17)$$

$$\alpha_1(x_i) = \Pr(y_i = 0 | \tilde{y}_i = 1, x_i) \quad (18)$$

The expected value of the observed dependent variable is:

$$E(y_i|x_i) = \Pr(y_i = 1|x_i) = \alpha_0(x_i) + (1 - (\alpha_0(x_i) - \alpha_1(x_i)))F(x_i'\beta) \quad (19)$$

For example, suppose misclassification only depends on one covariate x_{i1} , then the expression given in (19) becomes:

$$E(y_i|x_i) = \Pr(y_i = 1|x_i) = \alpha_0(x_{i1}) + (1 - (\alpha_0(x_{i1}) - \alpha_1(x_{i1})))F(x_i'\beta) \quad (20)$$

and the likelihood function is similar to equation (10) only the two misclassification probabilities appear as a function of x_{i1} . The model can be identified in a similar manner to what was described in Section 4.2. To see this, first consider the case where x_{i1} is a dummy variable:

$$\alpha_0 = \gamma_0 + \gamma_1 x_1 \quad (21)$$

$$\text{and } \alpha_1 = \delta_0 + \delta_1 x_1 \quad (22)$$

$$\text{Therefore when } x_1 = 1 \Rightarrow \alpha_0 = \gamma_0 + \gamma_1 \text{ and } \alpha_1 = \delta_0 + \delta_1 \quad (23)$$

$$\text{and when } x_1 = 0 \Rightarrow \alpha_0 = \gamma_0 \text{ and } \alpha_1 = \delta_0 \quad (24)$$

The expected value of the observed dependent variable becomes:

$$E(y_i|x_i) = \Pr(y_i = 1|x_i) = \gamma_0 + \gamma_1 x_1 + (1 - (\gamma_0 + \gamma_1 x_1) - (\delta_0 + \delta_1 x_1))F(x_i'\beta) \quad (25)$$

The identification of the misclassification probabilities can be seen more clearly by taking limits of $E(y_i|x_i)$ as $x_i'\beta$ (where $x_i'\beta$ excludes x_{i1}) tends to $-\infty$ and $+\infty$ in (25):

$$\lim_{x_i'\beta \rightarrow -\infty} E(y_i|x_i) = \gamma_0 \text{ when } x_{i1} = 0 \text{ and } \lim_{x_i'\beta \rightarrow -\infty} E(y_i|x_i) = \gamma_0 + \gamma_1 \text{ when } x_{i1} = 1 \quad (26)$$

$$\lim_{x_i'\beta \rightarrow \infty} E(y_i|x_i) = (1 - \delta_0) \text{ when } x_{i1} = 0 \text{ and } \lim_{x_i'\beta \rightarrow \infty} E(y_i|x_i) = (1 - \delta_0 - \delta_1) \text{ when } x_{i1} = 1 \quad (27)$$

Intuitively if misclassification depends on x_{i1} , then the probabilities of misclassification are constant within the two subgroups where $x_{i1} = 1$ and $x_{i1} = 0$. Then γ_0 is identified from the group of workers who truly have a very low probability of changing jobs and who have $x_{i1} = 0$, while γ_1 is identified from the group of workers who truly have a very low probability of changing jobs and who have $x_{i1} = 1$. Similar to before, identification is achieved from the group of workers for whom $x_i'\beta$ is highly negative and who are therefore very unlikely to be job changers but some of them will be misclassified only in this case we effectively divide this group of workers into two subgroups where $x_{i1} = 1$ and $x_{i1} = 0$. A comparable argument can be made for the identification of δ_0 and δ_1 . When x_{i1} is a dummy variable there are four misclassification probabilities to estimate.

If x_{i1} is a discrete variable, then all observations within the subgroups that have the same values for x_{i1} will have constant misclassification probabilities and identification is achieved in a similar way as described above. However, if we have many subgroups it will be harder to estimate $\gamma_0, \gamma_1, \gamma_2$ etc and $\delta_0, \delta_1, \delta_2$ etc as there may not be sufficient workers in the tails of the index with the full range of values of x_{i1} . Finally, if x_{i1} is a continuous variable, we can define cut-off points c_1, c_2, c_3 etc such that if $x_{i1} < c_1$ the misclassification probabilities are constant (or almost constant) within each group.

As before, identification is only possible because the misclassification probabilities enter the likelihood function in an additive way and their sum must be less than one.

6. Estimation Results

6.1 Estimation Results: Misclassification Independent of Covariates

Table 7 shows the estimates from a standard probit regression of the probability of job change and the estimates from the Hausman *et al.* procedure to control for misclassification. This provides estimates of the probabilities of misclassification and allows comparisons to be made on the effect of response error on the estimated coefficients. The data for 1995 to 2001 have been pooled so that there are 9,377 observations from which I have identified 927 job changes. The explanatory variables are defined in Table A1. Zero misclassification probabilities are used as starting values in estimating the model with misclassification.¹³

The estimated probability of misclassification for non-job changers, α_0 is very small at less than one per cent and the estimated probability of misclassification for job changers, α_1 , is high at 54 per cent. Significance tests on α_0 and α_1 can be used as tests of misclassification. Although α_0 is not significant, α_1 is highly significant and so we reject the model without misclassification. Workers who have truly changed jobs are more likely to be misclassified, as α_1 exceeds α_0 . This means that the measure of job change is likely to undercount the true number of job changes.

Hausman *et al.* also apply the procedure to a model of job change using data from the January 1987 Current Population Survey from the Census Bureau. Their study provides an external estimate of the misclassification probabilities. They estimate α_0 to be 6.1 per cent and α_1 to be 30.9 per cent.

¹³ A range of different starting values for α_0 and α_1 were used to check the robustness of the results. If α_0 is given a starting value of 0, the results are robust for any starting value of α_1 between 0 and 1. Similarly if α_1 is given a starting value of 0, the results are robust for any starting value of α_0 between zero and 1. If α_0 and α_1 are given the same starting values, the results are robust up to starting values of 0.32. For starting values between 0.32 and 0.45 the iterations don't make any progress. For higher starting values such as 0.46 the program estimates $(1 - \alpha_0, 1 - \alpha_1, -\beta)$ and for starting values around 0.6 the results are not sensible.

When we allow for misclassification, the estimated coefficients have higher standard errors implying that errors in responses lead to a loss in estimation efficiency. The results also indicate that ignoring response misclassification leads to diminished covariate effects. It is easier to interpret differences in estimates from the two regressions if we look at marginal effects instead of coefficient estimates. These are presented in Table 8.

Table 7: Coefficient Estimates from Models of Job Change*[^]

Variable	Standard Probit Model		Misclassification Model	
	Coefficient Estimate	Standard Errors	Coefficient Estimate	Standard Errors
α_0			0.0091	0.0089
α_1			0.5407	0.1970
Experience	-0.0748	0.0078	-0.1187	0.0465
Experience squared	0.0011	0.0002	0.0018	0.0008
Education- medium	-0.1283	0.0551	-0.1914	0.1017
Education- high	-0.1745	0.0919	-0.3104	0.2003
(Ref: Education – low)				
Public Sector	-0.2068	0.0769	-0.2896	0.1312
Number of Employees > 50	-0.1755	0.0498	-0.2847	0.1363
Overskilled	0.2095	0.0399	0.3192	0.1150
Occupation of Origin:				
(Ref: Elementary Occ's)				
Manager	-0.4953	0.0959	-0.7898	0.3591
Professional	-0.3966	0.0817	-0.5963	0.2276
Clerk	-0.3191	0.0680	-0.5412	0.2794
Skilled	-0.3412	0.0709	-0.5033	0.1757
Sector of Origin:				
(Ref: Non Market Services)				
Agric. & Mining & Utilities	-0.3759	0.1156	-0.5935	0.3181
Manufacturing	-0.2238	0.0985	-0.3733	0.2263
Building	0.3526	0.1098	0.5859	0.3041
Market Services	0.1483	0.0780	0.2338	0.1513
Year Dummies:				
(Ref: 1995)				
1996	0.0225	0.0767	0.0679	0.1301
1997	0.0893	0.0785	0.1733	0.1565
1998	0.2422	0.0734	0.4174	0.2309
1999	0.2246	0.0765	0.3660	0.1872
2000	0.3579	0.0743	0.5995	0.2828
2001	0.2254	0.0748	0.4004	0.2331
Constant	-0.3583	0.1259	0.6036	0.7951
N		9,377		9,377
Wald chi2		481.18		31.67
Prob > chi2		0.000		0.0632
Log pseudolikelihood		-2647.10		-2645.73

* Note: Standard errors are adjusted to take account of the fact that there are multiple observations on the same people

[^] Note: Controls for gender, presence of children, marital status and region were included in the model but dropped from the final specification because they were not significant

Table 8: Marginal Effects from Models of Job Change*^

Variable	<i>Marginal Effect</i>		<i>Marginal Effect</i>		<i>% diff. in Marginal Effects</i>
	<i>P> Z </i>		<i>P> Z </i>		
	<u>Standard Probit Model</u>		<u>Misclassification Model</u>		
α_0			0.0091	0.31	
α_1			0.5407	0.01	
Experience	-0.0104	0.00	-0.0419	0.01	404%
Experience squared	0.0002	0.00	0.0007	0.02	412%
Education- medium	-0.0175	0.02	-0.0435	0.06	249%
Education- high	-0.0220	0.06	-0.0635	0.12	288%
(Ref: Education – low)					
Public Sector	-0.0269	0.01	-0.0627	0.03	233%
Number of Employees > 50	-0.0234	0.00	-0.0630	0.04	269%
Overskilled	0.0294	0.00	0.0747	0.01	254%
Occupation of Origin:					
(Ref: Elementary Occ's)					
Manager	-0.0512	0.00	-0.1271	0.03	248%
Professional	-0.0477	0.00	-0.1173	0.01	246%
Clerk	-0.0392	0.00	-0.1072	0.05	274%
Skilled	-0.0412	0.00	-0.0997	0.00	242%
Sector of Origin:					
(Ref: Non Market Services)					
Agric. & Mining & Utilities	-0.0425	0.00	-0.1023	0.06	241%
Manufacturing	-0.0280	0.02	-0.0762	0.10	272%
Building	0.0602	0.00	0.1693	0.05	281%
Market Services	0.0212	0.06	0.0561	0.12	264%
Year Dummies:					
(Ref: 1995)					
1996	0.0032	0.77	0.0161	0.60	510%
1997	0.0130	0.26	0.0427	0.27	329%
1998	0.0380	0.00	0.1114	0.07	293%
1999	0.0349	0.00	0.0959	0.05	275%
2000	0.0592	0.00	0.1677	0.03	284%
2001	0.0349	0.00	0.1057	0.09	303%
N		9,377		9,377	
Wald chi2		481.18		31.67	
Prob > chi2		0.000		0.0632	
Log pseudolikelihood		-2647.10		-2645.73	

* Note: Standard errors are adjusted to take account of the fact that there are multiple observations on the same people

^ Note: Controls for gender, presence of children, marital status and region were included in the model but dropped from the final specification because they were not significant

Although both models indicate that the same factors determine job mobility the effect of misclassification in the dependent variable on the marginal effects of the various explanatory variables is sizeable.¹⁴ In the theoretical literature on job mobility, years

¹⁴ The marginal effects are evaluated at the sample means of the explanatory variables.

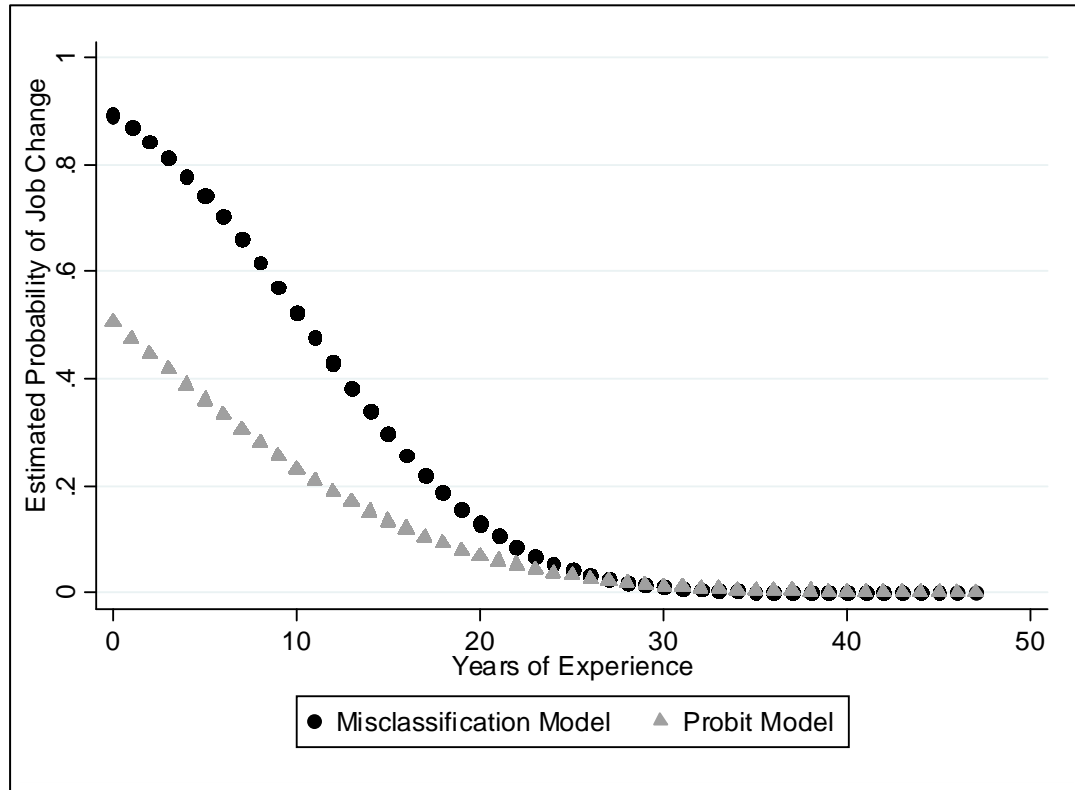
of labour market experience is a key determinant of job change and both models provide findings that are consistent with this. However, in the model that does not allow for misclassification the marginal effect of an additional year of experience is to reduce the probability of changing jobs by 1 percentage point, while the marginal effect in the misclassification model is almost four times larger so an additional year of experience reduces the probability of changing jobs by 4.2 percentage points.

The results also indicate that the negative effect of human capital on the probability of changing jobs is more marked in the misclassification model. For example, general human capital is proxied by education level and in the model incorporating misclassification the marginal effect of third level education almost three times higher than in the probit model, indicating that those with third level education are 6.4 per cent less likely to change jobs than those who have at most Junior Certificate education. The occupation indicator variable is intended to capture more specific human capital and again the marginal effects of higher levels of occupational attainment relative to those in elementary occupations are higher in the misclassification model. Working in the public sector is found to exert a negative effect on the probability of changing jobs and the marginal impact of working in the public sector in the misclassification model is over double the impact than in the model without misclassification.

The effect of the sector a worker was in the previous year (or for job changers the sector they previously worked in) is similar in both models but again the marginal effects are higher in the misclassification model. Finally, the marginal impacts of being overskilled, an indicator of poor match quality, working in a firm with more than 50 employees and the time dummies, which capture factors that vary over time but that affect all people are all bigger in magnitude in the misclassification model.

A useful way to demonstrate the differences between the two models is to graph the marginal effects of the variables. Figure 1 plots the marginal effect of experience from both models. The curves slope down as the probability of job change decreases as years of experience increases (i.e. the marginal effect on experience is negative). The slopes of the curves are steep at lower values of experience and then flatten out at higher years of experience indicating that an additional year of experience reduces the probability of changing jobs but at a declining rate (i.e. the marginal effect on years of experience squared is positive). Overall, the graph shows that the effect of ignoring misclassification error is large, especially at low values of experience.

Figure 1: Marginal Effect of Experience in Models of Job Mobility



6.2 Estimation Results: Covariate-Dependent Misclassification

The results in Section 6.1 show that misclassification has the biggest impact on the marginal effect for experience.¹⁵ Section 3.2 argued that there might be differences in measurement error in the data by experience and gender. Table 9 reports the results for the misclassification model where the misclassification probabilities depend on experience.¹⁶ Table 10 reports the results where the misclassification probabilities depend on gender; even though gender isn't an important determinant of job mobility Section 3.2 argued that there could be gender differences in measurement error.¹⁷

In the model that allows the misclassification probabilities to depend on experience, the estimate of the probability of misclassifying a non-job change remains insignificantly different from zero for each of the experience categories. However, there is some variation within experience groups when we allow misclassification to depend on experience. The probability of misclassifying a job change for a worker with less than 20 years of experience, δ_0 , is 42 per cent. The additional effect for someone with more than 20 years experience is given by δ_1 and the estimate indicates

¹⁵ The effect of misclassification is largest on the marginal effect of the 1996 time dummy although the marginal effect is insignificant. It could be that there were coding errors in the responses to the LIS for that wave.

¹⁶ A categorical experience variable is used to in the model when we allow the misclassification probabilities to depend on experience.

¹⁷ A series of models were run where misclassification was allowed to depend on each of the covariates but none of the estimated probabilities were significant. In addition, a model was run where the probability of misclassifying job changes was allowed to depend on all the covariates but the model failed to achieve convergence.

that these workers are almost 37 percentage points less likely to be misclassified as not having changed jobs than someone with less than 20 years experience. Although the signs of the estimated probabilities are what we would expect none of the estimated probabilities are significant. In general, the coefficient estimates are similar to the model where misclassification is independent of the covariates. In addition, there is some loss in estimation efficiency as the standard errors are a little higher than before.

When the misclassification probabilities depend on gender, the probability of misclassifying a non-job change is not statistically different from zero for men and women. The probability of misclassifying a job change is around 54 per cent for men and the additional effect of misclassifying a job change for women is small and not significant. Once more there is a small loss in estimation efficiency when we allow the misclassification probabilities to differ by gender.

It may be the case the variation in the data is not sufficient to accurately identify covariate dependent misclassification or the results presented in Tables 9 and 10 may indicate that misclassification is independent of the covariates.

Table 9: Covariate Dependent Misclassification Models of Job Change*

Variable	Coefficient Estimate	Standard Error	$P > Z $
Misclassification Depends on:			
		Experience	
γ_0 if experience < 20 years	0.0042	0.0111	0.71
γ_1 if experience \geq 20 years	-0.0011	0.0077	0.88
δ_0 if experience < 20 years	0.4197	0.3524	0.23
δ_1 if experience \geq 20 years	-0.3697	0.3386	0.28
Experience	-0.1067	0.0430	0.01
Experience squared	0.0015	0.0007	0.03
Education- medium	-0.1578	0.0864	0.07
Education- high	-0.2454	0.1851	0.19
(Ref: Education – low)			
Public Sector	-0.2493	0.1121	0.03
Number of Employees > 50	-0.2162	0.1136	0.06
Overskilled	0.2524	0.0974	0.01
Occupation of Origin:			
(Ref: Elementary Occ's)			
Manager	-0.6218	0.2943	0.04
Professional	-0.4710	0.1835	0.01
Clerk	-0.3944	0.2148	0.07
Skilled	-0.4037	0.1561	0.01
Sector of Origin:			
(Ref: Non Market Services)			
Agric. & Mining & Utilities	-0.4497	0.2373	0.06
Manufacturing	-0.3012	0.1882	0.11
Building	0.4291	0.2229	0.05
Market Services	0.1651	0.1204	0.17
Year Dummies:			
(Ref: 1995)			
1996	0.0414	0.1174	0.73
1997	0.1237	0.1420	0.38
1998	0.3113	0.1946	0.11
1999	0.2774	0.1715	0.11
2000	0.4475	0.2454	0.07
2001	0.2986	0.1959	0.13
Constant	0.2993	0.8408	0.72
N		9,377	
Wald chi2		32.98	
Prob > chi2		0.0465	
Log pseudolikelihood		-2641.29	

* Note: Standard errors are adjusted to take account of the fact that there are multiple observations on the same people

Table 10: Covariate Dependent Misclassification Models of Job Change

Variable	Coefficient Estimate	Standard Error	P> Z
Misclassification Depends on:			
	Gender		
γ_0 if male	0.0112	0.0097	0.25
γ_1 if female	-0.0092	0.0061	0.14
δ_0 if male	0.5264	0.2235	0.02
δ_1 if female	0.0161	0.0510	0.75
Experience	-0.1172	0.0476	0.01
Experience squared	0.0018	0.0008	0.03
Education- medium	-0.1812	0.0981	0.07
Education- high	-0.2930	0.1975	0.14
(Ref: Education – low)			
Public Sector	-0.3010	0.1336	0.02
Number of Employees > 50	-0.2776	0.1375	0.04
Overskilled	0.3098	0.1186	0.01
Occupation of Origin:			
(Ref: Elementary Occ's)			
Manager	-0.7745	0.3566	0.03
Professional	-0.5741	0.2291	0.01
Clerk	-0.5026	0.2823	0.08
Skilled	-0.5071	0.1736	0.00
Sector of Origin:			
(Ref: Non Market Services)			
Agric. & Mining & Utilities	-0.6500	0.3360	0.05
Manufacturing	-0.4153	0.2332	0.08
Building	0.5211	0.3099	0.09
Market Services	0.1877	0.1462	0.20
Year Dummies:			
(Ref: 1995)			
1996	0.0582	0.1310	0.66
1997	0.1547	0.1532	0.31
1998	0.4081	0.2321	0.08
1999	0.3562	0.1877	0.06
2000	0.5880	0.2926	0.04
2001	0.3857	0.2366	0.10
Constant	0.6198	0.8265	0.45
N		9,377	
Wald chi2		31.48	
Prob > chi2		0.0661	
Log pseudolikelihood		-2643.81	

* Note: Standard errors are adjusted to take account of the fact that there are multiple observations on the same people

7. Conclusions

This paper finds that there is substantial measurement error in responses to a question about tenure in the LIS; the extent of which is similar to what has been found in other studies. Survey data on tenure are very often used to deduce job mobility. Given the extent of response error evident in the data it is likely that cases are misclassified as having changed jobs when they truly haven't and vice versa. A procedure proposed by Hausman *et al.* is used to control for misclassification. The results indicate that by ignoring misclassification the true number of job changes will be underestimated by around 50 per cent. In addition, ignoring misclassification leads to diminished covariate effects. Future research will extend this analysis to other countries in the ECHP dataset.

Table A1: Explanatory Variables: Definitions and Summary Statistics

<i>Variable</i>	<i>Description</i>	<i>Mean</i>	<i>Std. Dev.</i>
Experience	Number of years in employment	19.3	11.5
Education- low (Reference Category)	Dummy variable that takes the value 1 if highest educational qualification is Junior Certificate and zero otherwise	0.46	0.50
Education- medium	Dummy variable that takes the value 1 if highest educational attainment is above Junior Certificate but below degree level and zero otherwise	0.41	0.49
Education- high	Dummy variable that takes the value 1 if highest educational qualification is a degree or above and zero otherwise	0.13	0.34
Public	Dummy variable that takes the value one if the person was working in the public sector in the previous year and zero otherwise	0.28	0.45
Number of Employees	Dummy variable that takes the value 1 if the number of employees in the firm in the previous year is more than 50 and zero otherwise.	0.35	0.48
Overskilled	Dummy variable that takes the value 1 if the worker reported that they felt they had skills and qualifications to do a more demanding job and zero otherwise.	0.47	0.50
Occupation of Origin:			
Manager	Dummy variable that takes the value 1 if occupation of origin is manager, senior official or legislator and zero otherwise	0.10	0.29
Professional	Dummy variable that takes the value 1 if occupation of origin is professional, technician or associated professionals and zero otherwise	0.25	0.43
Clerk	Dummy variable that takes the value 1 if occupation of origin is clerk, service, shop or sale worker and zero otherwise.	0.23	0.42
Skilled	Dummy variable that takes the value 1 if occupation of origin is skilled agricultural or fishery worker or a skilled craft or trades worker and zero otherwise.	0.22	0.41
Elementary (Reference Category)	Dummy variable that takes the value 1 if occupation in the previous year is plant or machine operator or assembler, or elementary occupation and zero otherwise.	0.20	0.40
Sector of Origin:			
Agriculture & Mining	Dummy variable that takes the value 1 if sector of origin is agriculture, fishing, mining or quarrying , or utilities and zero otherwise.	0.13	0.33
Manufacturing	Dummy variable that takes the value 1 if sector of origin is manufacturing and zero otherwise.	0.19	0.39
Building	Dummy variable that takes the value 1 if sector of origin is building and zero otherwise.	0.08	0.27
Market Services	Dummy variable that takes the value 1 if sector of origin is distribution, hotels and restaurants, transport, storage and communications, financial intermediation, or real estate, renting and business activities and zero otherwise.	0.35	0.48
Non-Market Services (Reference Category)	Dummy variable that takes the value 1 if sector or origin is education, public administration and defence or health and social work and zero otherwise.	0.25	0.43
Year Dummies:			
1995 (Reference Category)	Dummy variable that takes on the value 1 if the year is 1995 and zero otherwise.	0.13	

1996	Dummy variable that takes on the value 1 if the year is 1996 and zero otherwise.	0.13
1997	Dummy variable that takes on the value 1 if the year is 1997 and zero otherwise.	0.14
1998	Dummy variable that takes on the value 1 if the year is 1998 and zero otherwise.	0.14
1999	Dummy variable that takes on the value 1 if the year is 1999 and zero otherwise.	0.15
2000	Dummy variable that takes on the value 1 if the year is 2000 and zero otherwise.	0.15
2001	Dummy variable that takes on the value 1 if the year is 2001 and zero otherwise.	0.16

References

- Artís, M., Ayuso, M. and M. Guillén (2002), “Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims”, *The Journal of Risk and Insurance*, Vol. 69, No. 3, pp. 325-340.
- Baker, M. and G. Solon (1999), “*Earnings Dynamics and Inequality among Canadian Men, 1976-1992: Evidence from Longitudinal Income Tax Records*”, NBER Working Paper No. 7370.
- Bound, J., Brown, C. and N. Mathiowetz (2001), “Measurement Error in Survey Data”, in J. Heckman and E. Leamer, eds, *Handbook of Econometrics*, Vol. 5.
- Brachet, T. (2005), “Maternal Smoking, Misclassification Error and Infant Health”, Working Paper. Available at: <http://works.bepress.com/tbrachet/1>.
- Brown, J. and A. Light (1992), “Interpreting Panel Data on Job Tenure”, *Journal of Labor Economics*, Vol. 10, No. 3 (Jul., 1992), pp. 219-257.
- Farber, S. (1999), “Mobility and Stability: The Dynamics of Job Change in Labor Markets”, in O. Ashenfelter and D. Card, eds, *Handbook of Labor Economics*, Vol.3.
- Hausman, J. (2001), “Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left”, *The Journal of Economic Perspectives*, Vol. 15, No. 4 (Autumn 2001), pp. 57-67.
- Hausman, J., Abrevaya, J. and F.M. Scott-Morton (1998), “Misclassification of the dependent variable in a discrete-response setting”, *Journal of Econometrics*, Vol. 87, pp. 239-269.
- Jensen, P., Palangkaraya, A. and E. Webster (2008), “*Misclassification in Patent Offices*”, Intellectual Property Research Institute of Australia, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Working Paper No. 02/08, May 2008.
- Kenkel, D., Lillard, D. and A. Mathios (2004), “Accounting for misclassification error in retrospective smoking data”, *Health Economics*, Vol. 13, pp. 1031-1044.
- Ureta, M. (1992), “The Importance of Lifetime Jobs in the U.S. Economy, Revisited”, *American Economic Review*, Vol. 82, No. 1, pp. 322-335.
- Weiss, D., Dawis, R., England, G. and L. Lofquist (1961), “Validity of Work Histories Obtained by Interview”, Industrial Relations Center, University of Minnesota.