HEDG
HEALTH, ECONOMETRICS AND DATA GROUP

THE UNIVERSITY of York

WP 13/16

# Point identification in the presence of measurement error in discrete variables: application – wages and disability

Eirini-Christina Saloniki & Amanda Gosling

August 2013

york.ac.uk/res/herc/hedgwp

# POINT IDENTIFICATION IN THE PRESENCE OF MEASUREMENT ERROR IN DISCRETE VARIABLES: APPLICATION - WAGES AND DISABILITY[*]

**Paper prepared for the 22th European Workshop on 'Econometrics and Health Economics' September 2013[†]**

Eirini-Christina Saloniki[‡] and Amanda Gosling[§]
University of Kent

## Abstract

This paper addresses the problem of point identification in the presence of measurement error in discrete variables; in particular, it considers the case of having two "noisy" indicators of the same latent variable and without any prior information about the true value of the variable of interest. Based on the concept of the fourfold table and creating a nonlinear system of simultaneous equations from the observed proportions and predicted wages, we examine the need for different assumptions in order to obtain unique solutions for the system. We show that by imposing a simple restriction(s) for the joint misclassification probabilities, it is possible to measure the extent of the misclassification error in that specific variable. The proposed methodology is then used to identify whether people misreport their disability status using data from the British Household Panel Survey. Our results show that the probability of underreporting is greater than the probability of overreporting disability.

**Keywords:** measurement error; discrete; misclassification probabilities; identification; disability.
**JEL classification:** C14, C35, J14, J31.

[*] We would like to thank Sylvain Barde for his technical advice and Melanie Jones, Sarah Brown and Andy Dickerson for their insightful comments and suggestions.

[†] The full paper can be found at http://ideas.repec.org/p/ukc/ukcedp/1214.html.

[‡] School of Economics, University of Kent, Email: es343@kent.ac.uk.
[§] School of Economics, University of Kent, Email: A.Gosling@kent.ac.uk.

## I. INTRODUCTION

Measurement error in survey data has received much attention from statistical analysts and economists over the last three decades. It is known that individuals when taking part in a survey have to answer some questions about their behaviour or other personal characteristics; however, it is commonly observed that these respondents tend to not truly report their characteristics. As a result, the estimated coefficients of the relevant variables used in the regression analysis are often biased (towards zero) leading to inconsistent interpretation of the results.

Although most survey questions are related to variables that are thought of as continuous such as age, earnings or wages, a considerable part of the data is in the form of variables taking values in finite sets (discrete). Examples of the latter include educational qualifications, race, marital status, employment status and health/functional status.

When measurement error in discrete variables occurs, it is normal to think of the problem in terms of misclassification error. Misclassification can occur in one of the following ways: the respondent indicates having a specific characteristic when he truly does not (false positive response - Type I error) or indicates not possessing the characteristic when he really does so (false negative response - Type II error). In that way, they either overestimate or underestimate respectively the quantity being requested. Possible sources of the misclassification error are associated with the respondents' difficulty to understand or remember the question being asked or his willingness to say what the interviewer, and the society in general want to hear (socially desirable).[1] Also, it may be present due to the fact that certain words mean different things to different people or because the respondent gives different answers to different interviewers depending on his mood, attitude or the day the survey is taking place (Lohr 1999). An example of misclassification error in discrete variables is related to how people report their disability status. There are strong economic (Bound and Burkhauser 1999) and psychological (Myers 1982, Bowe 1993, Hale 2001) incentives to misreport disability status. In particular, the respondents in a survey tend to either overreport being disabled so that they can take advantage of any social benefits such as claiming a disability or incapacity benefit or underreport being disabled in the fear of stigma and social exclusion. A possible social stigma in the long-run might reduce the possibilities of finding a job for those who are currently unemployed or alternatively, impact negatively the wages of those who are already in the labour market.

In the literature, a vast array of methods has been developed to accommodate measurement error and essentially misclassification in model analysis. These methods can be classified into two broad categories; the first one is related to setting bounds for the misclassification probabilities which can be then used to bound the corresponding estimates of the mismeasured variables. Interestingly, that approach is not limited to the number of the mismeasured variables, although in most cases they cannot pick up the same latent variable.

If the regressor is binary, assuming that the reporting process is exogenous, in other words the error process generating the mismeasured variable is independent from the residual error in the structural equation, and that the researcher has knowledge of the

[1] The difficulty to understand or remember the question being asked is known as telescoping or recall bias.

true value of the mismeasured variable, then it is possible to *bound* the corresponding coefficients away from zero (Bollinger 1996).[2] Moreover, when stronger prior information about the relationship between the two misclassification probabilities is available - $p$ and $q$ stand for the false positive and false negative responses respectively - such as $p+q<1$ or equivalently $p \leq 1/2$ and $q \leq 1/2$ these bounds can be made tighter. A different approach proposed initially by Horowitz and Manski (1995) and extended from Kreider and Pepper (2007 and 2008), based once again on a nonparametric bounding methodology but imposing further assumptions provides similar bounds for the estimated coefficients of the mismeasured variable.[3] The new sets of assumptions imposed are mainly of distributional and functional form; firstly, a certain subgroup of respondents provides fully accurate reports, hence there is a *known* lower bound on the accurate reporting rate. However, it should not be disregarded that without any prior information on the nature and degree of that accurate reporting rate the bounds can be very wide to impossible. Secondly, there is a *monotonic* relationship between the true value of the mismeasured variable and specific observed covariates such as age. For example, in case the mismeasured variable is the disability status it is usually assumed that the population disability rate is non-decreasing with age or that the employment rate of the disabled is non-increasing with age.

The second way of accommodating misclassification error is mostly found in epidemiology studies and used to determine whether a new method for testing a disease is as effective as an existing one.[4] It is based on exacting *unique solutions* for the misclassification probabilities using algebraic analysis and the concept of the *fourfold table* or a *maximum likelihood procedure*.

In particular, Bross (1954) assuming that there exists a unique method with no measurement error available to the researcher, a “gold standard” as it is usually denoted, and that both misclassification probabilities should be equal and each less than 0.5 showed that they can be calculated directly. Knowing the exact value of the misclassification probabilities, he also proposed a way to derive the bias between the “gold standard” method and the other method with the measurement error. An extension of his proposed model came later on from Tenenbein (1972) who allowed the two methods to follow a multinomial distribution. In comparison to the above studies, Hui and Walter (1980) claimed that *none* of the two methods for disease testing needed to be a “gold standard”; if the two methods were applied simultaneously to the same individuals from two populations with different disease prevalence, and assuming conditional independence of the errors of these methods (in the same way as with the non-parametric methodology), then it is possible to derive closed form solutions for the estimates of the error rates of both methods, and the true prevalence in both populations by using a *maximum likelihood procedure*.

[2] The exogenous reporting process is known as “non-differential” misclassification error.
Bollinger (1996) also provides bounds for the misclassification probabilities.

[3] This approach is known as monotone instrument variable bound (MIV).
Kreider and Pepper (2007) found that under relatively weak nonparametric assumptions non-workers appear to systematically overreport being disabled.

[4] The particular method of accommodating measurement error in epidemiology studies is known in the literature as finite fixture models.

Considering the above ways proposed to deal with the measurement error, an important question then arises: Is it possible to achieve point identification in the presence of misclassification error if: i) there are more than one (two) mismeasured variables that identify the *same* (unobserved) characteristic; ii) *no* “gold standard” sample or accurate reporting rate exists; and iii) the concept of the fourfold table is used as a baseline framework?

The aim of the paper is to answer that question; we show that we can exact identify the extent of measurement error in discrete variables by solving a nonlinear system of simultaneous equations from observed (expected sample) proportions and available predicted wages - without having any information about the true value of the variable of interest - and imposing a simple restriction for the joint misclassification probabilities.

We then use our proposed methodology to measure the extent of misclassification on reporting disability status using data from the British Household Panel Survey (BHPS). Assuming that we have two indicators of the disability status, our results show that the probability of underreporting is greater than the probability of overreporting disability. Depending on whether disability limits their daily activities or the type and amount of work they can do, truly non-disabled seem to underreport more that their disability limits their daily activities while the truly non-disabled underreport more that it limits the type or amount of work they can perform.

The rest of the paper is organized as follows. In Section II we present the model we use to exactly identify the extent of classification error in case we have two variables measured with error. In Section III we provide an overview of the different ways of measuring disability and introduce our data to empirically test our methodology. In Section IV we present the final solutions for the system of simultaneous equations, focusing on whether people misreport their disability status, as well as how the results differ if not correcting for the misclassification error. The fifth and final section concludes.

## II. THE MODEL

We start by introducing a fully general model which is directed towards finding unique solutions for each of the misclassification probabilities. Our model differs from the ones known in the literature since it is based on the strong assumption that we have two variables - “noisy” indicators - measured with error instead of having only one, in the sample that pick up the same latent variable and without having any prior information on the true value of the variable under consideration. Therefore, the coefficients for the two indicators are assumed to be the same.

To formalize the problem, suppose that we observe two mismeasured discrete variables, $M_1$ and $M_2$ that can take only two integer values $\{0,1\}$. The value of each of these variables will be equal to zero if the individual has reported not possessing a specific characteristic and will be equal to one if he has reported doing so. In addition, suppose that the *unobserved* true variable $T$ takes two values in the same way as the reported ones.

In order to build up the mathematical model, we first need to define the following misclassification probabilities:[5]

- $p_1$ : probability the person does not report possessing the attribute $M_1$ given that the true variable $T$ suggests possessing the attribute $\left[\Pr(M_1 = 0 \mid T = 1)\right]$;
- $p_2$ : probability the person does not report possessing the attribute $M_2$ given that the true variable $T$ suggests possessing the attribute $\left[\Pr(M_2 = 0 \mid T = 1)\right]$;
- $q_1$ : probability the person reports possessing the attribute $M_1$ given that the true variable $T$ suggests not possessing the attribute $\left[\Pr(M_1 = 1 \mid T = 0)\right]$;
- $q_2$ : probability the person reports possessing the attribute $M_2$ given that the true variable $T$ suggests not possessing the attribute $\left[\Pr(M_2 = 1 \mid T = 0)\right]$.

Also, we denote the probability the true variable is equal to one as $\lambda$ which is unknown.[6]

Then, in total we can obtain 8 joint misclassification probabilities for the two reported measures - for the benefit of notation, the first subscript in each probability denotes whether there is misclassification error in the $M_1$ variable while the second one whether there is misclassification error in the $M_2$ variable (see *Appendix, Annex A* for more information on the definitions of each one of them) - denoted as follows;[7] $\{p_{mm}, p_{Mm}, p_{mM}, p_{MM}, q_{mm}, q_{Mm}, q_{mM}, q_{MM}\}$.

Indicatively,

- $p_{mm}$ : probability the individual is not misclassified in none of the two variables (measures) given that he truly possesses the characteristic $\left[\Pr(M_1 = 1, M_2 = 1 \mid T = 1)\right]$;
- $p_{Mm}$ : probability the individual is misclassified only in one of the two variables - indicators $(M_1)$ given that he truly possesses the characteristic $\left[\Pr(M_1 = 0, M_2 = 1 \mid T = 1)\right]$;
- $q_{MM}$ : probability the individual is misclassified in both reported variables given that he does not truly possess the characteristic $\left[\Pr(M_1 = 1, M_2 = 1 \mid T = 0)\right]$.

Also, from the probability theory it holds that

$$p_{mm} + p_{Mm} + p_{mM} + p_{MM} = 1, \tag{1}$$

$$q_{mm} + q_{Mm} + q_{mM} + q_{MM} = 1. \tag{2}$$

[5] The subscript in each of these probabilities denotes the respective reported variable measured with error.

[6] The probability the true variable is equal to zero $\left[\Pr(T = 0)\right]$ is $1 - \lambda$ .

[7] A capital subscript letter indicates the existence of misclassification error whereas a little one indicates the non-existence of misclassification error.

Taking into account the above probabilities, we can create a fourfold table containing the expected sample (observed) proportions in each combination of the two reported variables (see *Table 1*). Each of these proportions will be just a weighted sum, by $\lambda$ and $(1-\lambda)$ accordingly, of the joint misclassification probabilities.

Table 1: Expected sample proportions

| | ***Reported variable 2 ($M_2$)*** | |
|---|---|---|
| ***Reported variable 1 ($M_1$)*** | **"Yes"** | **"No"** |
| **"Yes"** | $\lambda p_{mm}+(1-\lambda)q_{MM}$ | $\lambda p_{mM}+(1-\lambda)q_{Mm}$ |
| **"No"** | $\lambda p_{Mm}+(1-\lambda)q_{mM}$ | $\lambda p_{MM}+(1-\lambda)q_{mm}$ |

From Table 1 we obtain four equations, (3)-(6) each of which is equal to the observed proportions $A,B,C,D$ respectively.[8]

$$\lambda p_{mm}+(1-\lambda)q_{MM}=A, \tag{3}$$
$$\lambda p_{mM}+(1-\lambda)q_{Mm}=B, \tag{4}$$
$$\lambda p_{Mm}+(1-\lambda)q_{mM}=C, \tag{5}$$
$$\lambda p_{MM}+(1-\lambda)q_{mm}=D. \tag{6}$$

The set of equations (3)-(6) can be considered as a system of **4** simultaneous equations with **9** unknowns $\{\lambda, p_{mm}, p_{Mm}, p_{mM}, p_{MM}, q_{mm}, q_{Mm}, q_{mM}, q_{MM}\}$ for which we want to find unique solutions. However, equation (6) is collapsed since the two joint misclassification probabilities $p_{MM}$ and $q_{mm}$ can be easily calculated using conditions (1) and (2). Therefore, we end up with a system of **3** equations with **7** unknowns; we need 4 more equations to achieve full identification.

It is worthy to mention that even if we assume that the misclassification probabilities for each mismeasured variable are identical, in other words the probabilities of lying are absolutely the same, conditions (3)-(5) are *not* sufficient to reveal the prevalence of the probability of the true variable from the available data. In order to see why that happens we should examine the following alternatives:

*Case 1: One mismeasured variable*

If we have only one mismeasured variable $M$ and a "gold standard" $G$ the expected sample proportions cannot be determined directly from the existing sample. The available expected sample proportions are the number of people who have answered either "Yes" or "No" only to the reported variable $M$ (See *Appendix, Annex A* for more

[8] Note that $A+B+C+D=1$.

information regarding the definitions of the observed proportions and the misclassification probabilities).[9]

Table 2: Expected sample proportions with only one mismeasured variable

| | ***Reported variable (*** $M$ ***)*** | |
|---|---|---|
| ***"Gold standard" (*** $G$ ***)*** | **"Yes"** | **"No"** |
| **"Yes"** | $\lambda(1-p)$ | $\lambda p$ |
| **"No"** | $(1-\lambda)q$ | $(1-\lambda)(1-q)$ |

$$\lambda(1-p)=A', \tag{3i}$$
$$\lambda p=B', \tag{4i}$$
$$(1-\lambda)q=C', \tag{5i}$$
$$(1-\lambda)(1-q)=D'. \tag{6i}$$

The system of equations (3i), (4i) and (5i) - equation (6i) is collapsed for the same reason as mentioned earlier in the analysis for equation (4) - is under identified since the number of the available equations is less than the number of the unknowns $\{\lambda, p, A', B'\}$.

*Case 2: Two mismeasured variables*

If we have two mismeasured variables we do not observe any considerable difference regarding the identification of the true variable. In particular, assuming that $p_{mm}=q_{mm}$, $p_{mM}=q_{mM}$, $p_{Mm}=q_{Mm}$, and $p_{MM}=q_{MM}$, then equations (3)-(5) can be transformed as shown below;

$$\lambda p_{mm}+(1-\lambda)p_{MM}=A, \tag{3ii}$$
$$\lambda p_{mM}+(1-\lambda)p_{Mm}=B, \tag{4ii}$$
$$\lambda p_{Mm}+(1-\lambda)p_{mM}=C. \tag{5ii}$$

Similarly to the case of having only one mismeasured variable, the system of equations (3ii), (4ii) and (5ii) is again under identified since the available equations are less than the number of the unknowns $\{\lambda, p_{Mm}, p_{mM}, p_{mm}\}$.

The extra equations needed for full identification of the system will be taken from the fourfold table of expected (predicted) wages in every combination of the two reported variables (see *Table 3* below). The expected wages will be a sum of the baseline wage $w$ - simply the wage of not misreporting - and the penalty of belonging on each specific cell. It should be noted that $\beta$ stands for the coefficient which by assumption is the same for both measures. Thus, the predicted wages are basically a function of the expected sample proportions (See *Appendix, Part A* for a representation).

[9] Only the sums $(A'+C')$ and $(B'+D')$ are known from the existing data.

Table 3: Predicted wages

| | ***Reported variable 2 (*** $M_2$ ***)*** | |
|---|---|---|
| ***Reported variable 1 (*** $M_1$ ***)*** | **“Yes”** | **“No”** |
| **“Yes”** | $w+\frac{\beta\lambda p_{mm}}{\lambda p_{mm}+(1-\lambda)q_{MM}}$ | $w+\frac{\beta\lambda p_{mM}}{\lambda p_{mM}+(1-\lambda)q_{Mm}}$ |
| **“No”** | $w+\frac{\beta\lambda p_{Mm}}{\lambda p_{Mm}+(1-\lambda)q_{mM}}$ | $w+\frac{\beta\lambda p_{MM}}{\lambda p_{MM}+(1-\lambda)q_{mm}}$ |

Therefore, from Table 3 we obtain 4 more equations (7)-(10) but with two more unknowns, $w$ and $\beta$. Each of these equations is set equal to the average log gross weekly wages denoted by $E,F,G$ and $H$ respectively.[10]

$$w+\frac{\beta\lambda p_{mm}}{\lambda p_{mm}+(1-\lambda)q_{MM}}=E\,, \tag{7}$$

$$w+\frac{\beta\lambda p_{mM}}{\lambda p_{mM}+(1-\lambda)q_{Mm}}=F\,, \tag{8}$$

$$w+\frac{\beta\lambda p_{Mm}}{\lambda p_{Mm}+(1-\lambda)q_{mM}}=G\,, \tag{9}$$

$$w+\frac{\beta\lambda p_{MM}}{\lambda p_{MM}+(1-\lambda)q_{mm}}=H\,. \tag{10}$$

Overall, from the observed proportions and the predicted wages we obtain a nonlinear system of **7** simultaneous equations, (3)-(5) and (7)-(10), with **9** unknowns $\{p_{mm},p_{mM},p_{Mm},q_{mM},q_{Mm},q_{MM},\lambda,w,\beta\}$ so it is still under identified. The last equation for full identification will come from a restriction for the joint misclassification probabilities.

Interestingly, we can find an unbiased and inconsistent estimator of $\beta$ from the appropriate estimation of wages and by adding exogenous regressors that also have a key effect on wages. These regressors could be personal characteristics other than disability status such as age, qualifications and marital status. The coefficient we will get from the regression analysis will be very close to $\beta$.

### a. *Restrictions for the joint misclassification probabilities*

#### *A. No Bias/Monotonicity*

No bias essentially arises from the definition of the misclassification error; the respondents in the survey may report possessing a characteristic when they truly do not so, thus the variable of interest is upwards biased, in order to receive any social and

[10] Note that in comparison to the observed proportions $E+F+G+H\neq 1$.

economic benefits. Alternatively, they may report not possessing the characteristic when they truly do so causing a downwards bias to the variable of interest so that they avoid possible social exclusion and discrimination observed in the labour market. In both cases the respondent's intention is to act in a way that usually *improves* his position than if he had acted differently. For example, suppose that the specific characteristic is how people report their disability status; the participants are usually more willing to report an "excellent" or a "very good" health status than reporting a "bad" or a "very bad" one. By doing so, they considerably try to avoid receiving lower wages compared to those who reported having a "bad" or a "very bad" health - for those who are currently employed -, or avoid having difficulties in entering the labour market for those who are unemployed. Both facts can occur due to employer's prejudice or statistical discrimination against them.[11]

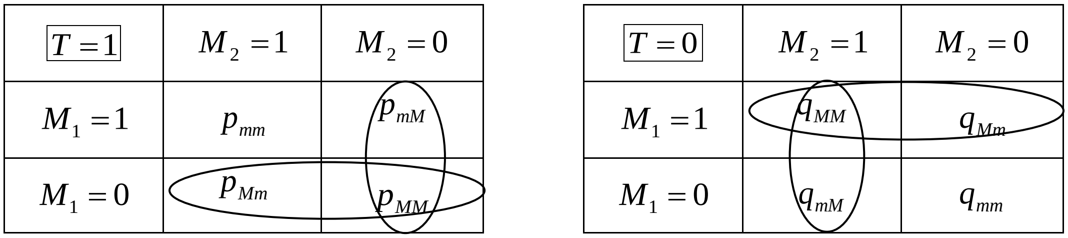
In algebraic terms, no bias assumption is translated as follows;

$$p_{Mm} + p_{MM} = q_{Mm} + q_{MM}\,, \tag{11}$$

$$p_{mM} + p_{MM} = q_{mM} + q_{MM}\,. \tag{12}$$

In a simpler way, no bias indicates that the probability of misreporting is the same for each reported variable $M_1$ and $M_2$ so it should be that $p_1 = q_1$, and $p_2 = q_2$. A graphical representation of the no bias assumption is shown in the next figure.

Figure 1: No bias assumption



| $\boxed{T=1}$ | $M_2=1$ | $M_2=0$ |
|---|---|---|
| $M_1=1$ | $p_{mm}$ | $p_{mM}$ |
| $M_1=0$ | $p_{Mm}$ | $p_{MM}$ |

| $\boxed{T=0}$ | $M_2=1$ | $M_2=0$ |
|---|---|---|
| $M_1=1$ | $q_{MM}$ | $q_{Mm}$ |
| $M_1=0$ | $q_{mM}$ | $q_{mm}$ |

By imposing only the no bias assumption we achieve full identification of the system of simultaneous equations since we have **9** equations with **9** unknowns $\{p_{mM}, p_{Mm}, p_{mm}, q_{Mm}, q_{mM}, q_{MM}, \lambda, w, \beta\}$.

Interestingly, a weaker assumption of no bias, *monotonicity*, rules out the possibility that $p$ is strictly less or more than $q$.

## *B. No correlation*

No correlation assumption refers to the case where the misclassification error in the one reported variable is independent of the misclassification error in the other (reported) variable. Therefore, each of the joint misclassification probabilities should be a product of the respective misclassification probabilities.

[11] Statistical discrimination occurs if employers do not have sufficient information to access the productivity of minority applicants accurately and they may use membership in the minority group as a signal of lower productivity thus, discriminate on the basis of wages against this group (Phelps 1972).

For that reason, according to the no correlation assumption it should hold that

$$p_{MM} = p_1 p_2 \,, \tag{13}$$
$$p_{mm} = (1-p_1)(1-p_2) \,, \tag{14}$$
$$p_{Mm} = p_1(1-p_2) \,, \tag{15}$$
$$p_{mM} = (1-p_1)\, p_2 \,, \tag{16}$$
$$q_{MM} = q_1 q_2 \,, \tag{17}$$
$$q_{mm} = (1-q_1)(1-q_2) \,, \tag{18}$$
$$q_{Mm} = q_1(1-q_2) \,, \tag{19}$$
$$q_{mM} = (1-q_1) q_2 \,. \tag{20}$$

For instance, if we consider the joint misclassification probability $p_{Mm}$, then according to the no correlation assumption those who lied on the first measure have the same probability of lying on the second measure as those that told the truth.

As with the no bias case, by imposing only the no correlation assumption we get a fully identified system with **7** equations and **7** unknowns $\{p_1, p_2, q_1, q_2, \lambda, w, \beta\}$.

*C. No bias and no correlation*[12]

In general, the combination of the above assumptions is close to the definition of the "classical" measurement error; the measurement error in the reported variable is assumed to be independent of the true level of that variable and all other variables used in the specific regression model. The error in the measurement of that variable will produce a downward biased (attenuated) and an inconsistent parameter estimate of its effect.

It should not be disregarded that when we impose at the same time no bias and no correlation assumptions, we end up with an over identified system of **7** equations with **5** unknowns $\{p_1, p_2, \lambda, w, \beta\}$.

Overall, we can conclude that it is possible to achieve either full or over identification of the system of simultaneous equations, and hence measure the extent of misclassification error in a specific variable, either by imposing only one or a combination of restrictions for the joint misclassification probabilities.[13]

[12] In that case it should hold that $p_{MM} = q_{MM} = p_1 p_2 = q_1 q_2$, $p_{Mm} = q_{Mm} = p_1(1-p_2) = q_1(1-q_2)$ and so on.

[13] For an analytic representation of the system of equations in each of the three possible cases ("no correlation", "no bias and no correlation" and "no bias"), see *Appendix, Annex A*.

## III. CLASSIFICATION ERROR AND DISABILITY STATUS

### *a. Measurement of disability*

There are different ways of inferring whether someone is thought of as disabled or not; either by using subjective, usually known as self-reported data (Baldwin and Schumacher 2002, Madden 2004, Jones, Latreille, and Sloane 2006a, Jones and Sloane 2009) or more objective measures of health (Burchardt 2000, Jones 2009) or finally indicators such as registered as disabled with Social Services or with a green card.

In subjective data, an individual through questions being asked assesses their own condition and whether that condition affects their capacity to undertake work. The survey questions are designed as closed as possible to the definition of disability stated in the legislative reforms. As a consequence, they usually take the form: *"Do you have any long-standing (at least six months) illness, disability or infirmity?"*; and in order to capture for work limitations: *"Does this health problem in any way affect your work capacity?"*. Nevertheless, it should not be disregarded that when determining whether an individual has a long-term health problem and whether that problem is work limiting there may be social and economic incentives to misreport disability status.

On the contrary, in objective data the information is on specific health conditions; either through a body mass index or a composite measure of mental well-being generated from the 12-item version of the General Health Questionnaire (GHQ12), information on subsequent mortality or even through doctor's reports. Therefore, the questions are narrower, more concrete and are not closely related to employment behaviour, as in the subjective data.

### *b. Data*

The data used in this study are from the British Household Panel Survey (BHPS) covering 4 waves from 2005-2009. For the purpose of our study the sample is restricted to employed men and women aged 21 to 60 years old. Each individual in the household has a unique identification number and every year is asked almost the same questionnaire. In some cases, a member of a household may be absent throughout the field period or too old to complete the interview themselves. As a consequence, a proxy respondent is administered to answer the questionnaire which is usually another member of the household, with preference shown for the spouse or adult child. Since in the proxy schedule the questionnaire is a much shortened version of the individual questionnaire proxy respondents are also excluded from our sample.

In addition, so as to allow for different health problems, a set of 13 dummies were conducted from a categorical health variable.[14] Then, making use of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), each of these dummies was further classified into three broad categories: physical, mental and mixed health problems.[15] Indicatively, physical health problems refer to problems such as

[14] The dummies for the different health problems are: "arm/leg/hands", "sight", "hearing", "skin", "chest/breathing", "heart/blood", "stomach/digestion", "diabetes", "anxiety/depression", "alcohol/drugs", "epilepsy", "migraine", and "other".

[15] More information about the ICD-10 can be found in *Appendix, Annex B*.

migraine and diabetes whereas the mixed category includes health problems such as epilepsy, cancer and stroke.

In order to define whether someone is thought of as disabled or not we focus on the latest definition of disability stated in the legislative reforms for the UK - Equality Act 2010 - and we make use of two measures to infer disability status.[16] For both measures, the respondent is classified as disabled if: he has initially responded **"Yes"** to the general question ***"Do you have any of the health problems or disabilities listed on this card?"***, accompanied by **"Yes"** in <u>*one*</u> of the specific questions - depending on the measure used - ***"Does your health limit the type of work or the amount of work you can do?"*** for the first measure, and ***"Does your health in any way limit your daily activities compared to most people of your age?"*** for the second measure respectively. In any other case, for example if the respondent has reported "Yes" to the general question and "No" to the second specific question or "No" to the general question and "Yes" to the first specific question and so on, he is classified as non-disabled.

Having made the appropriate changes and excluding all observations which still had missing values we ended up with a total sample of 19,048 individuals consisting of 9,235 men and 9,813 women. Regarding how many of them are classified as disabled, 1,187 are considered as disabled, covering 6.23 percent of the total sample using the first measure while using the second measure the disabled account for slightly less, 6.08 percent of the total sample (1,158 individuals). In other words, more people have reported that their health problem/disability limits the type or amount of work they can perform rather than their daily activities such as doing the housework, climbing stairs or walking for at least 10 minutes.

Distinguishing by type of health problem and based on the ICD-10 classification, Table 4 provides some insightful findings about how individuals report possessing at least one type of health problem (physical, mental or mixed) are distributed in the sample. Indicatively, using the second measure to infer disability status, more people - 1,063 individuals accounting for about 5.6% of the total sample - tend to report having physical health problems like arthritis and heart/blood pressure related problems while quite a few, only 46 individuals (covering 0.24 percent of the total sample) report having mental health problems such as anxiety and depression. Diversifying by gender and independently of the disability measure used, women tend to report slightly more suffering from mental and mixed health problems compared to men; almost 0.09% (0.03%) more women than men suffer from mental (mixed) problems.

[16] According to **Equality Act (2010)** an individual is thought of as disabled if "A physical or mental impairment has a substantial and long-term adverse effect on his or her ability to carry out normal day-to-day activities".

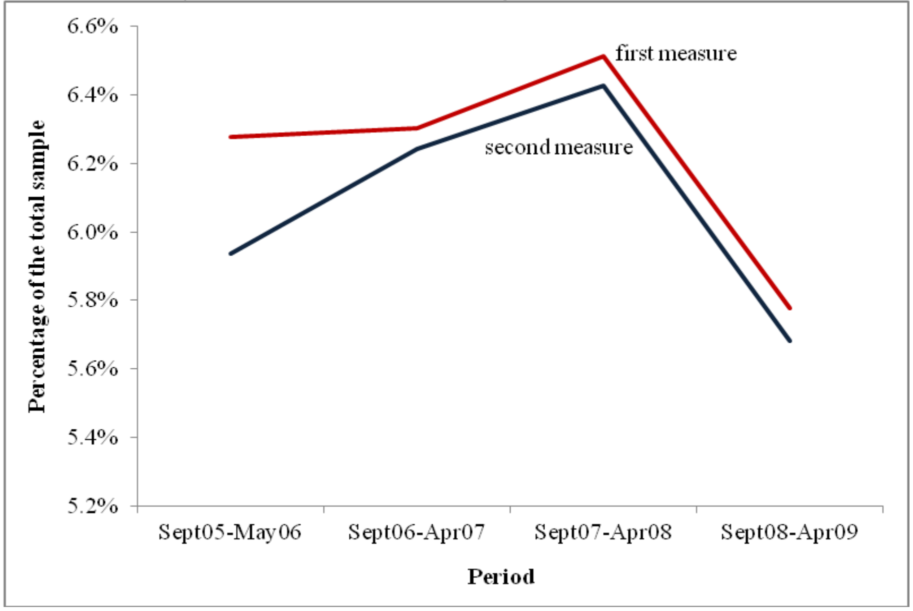
Table 4: Prevalence of different types of health problems by gender[17]

| ICD-10 | Males | | Females | |
|---|---|---|---|---|
| | ***First measure***[18] | ***Second measure***[19] | ***First measure*** | ***Second measure*** |
| ***Physical*** | 535 (2.81%) | 488 (2.56%) | 553 (2.90%) | 575 (3.02%) |
| ***Mental*** | 11 (0.06%) | 13 (0.07%) | 25 (0.13%) | 33 (0.17 %) |
| ***Mixed*** | 28 (0.15%) | 23 (0.12%) | 35 (0.18%) | 26 (0.14%) |
| ***Total*** | **574 (3.01%)** | **524 (2.75%)** | **613 (3.22%)** | **634 (3.33%)** |

### *c. Descriptive statistics*

Considering a more detailed analysis of the sample, between 2005 and 2009 disability rates in the United Kingdom have mainly decreased; using the different measures to denote disability status and starting from the first measure, disability rates remain stable till the middle of the second wave of our study, increase significantly in the next wave accounting for about 6.5 percentage points of the sample and reach a low of 5.8 percentage points in the last period between September 2008 and April 2009. With the second measure, though disability rates increase substantially in the first three waves, similarly to the first measure, they considerably decrease in the final wave accounting for almost 5.7% of the sample (See *Figure 2*).

Figure 2: Disability rates in the United Kingdom 2005-2009 for those at work[20]



Turning to the wages, as shown in Figure 3, at first glance disabled people seem to have lower earnings than the non-disabled ones; specifically, taking the average across

[17] Each of the numbers in parentheses is a percentage of the total sample.

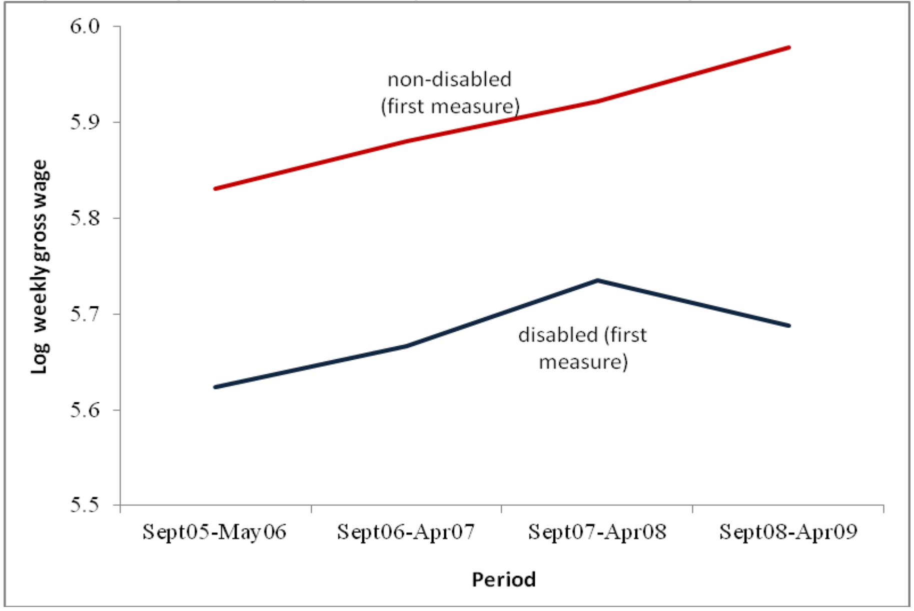
[18] *First measure* refers to those who have reported "Yes" to the first measure [ $M_1 = 1$].

[19] *Second measure* refers to those who have reported "Yes" to the second measure [ $M_2 = 1$].

[20] Each of the disability rates is expressed as a percentage of the total sample.

all waves, those whose disability limits the type or amount of work they can do (first measure) are paid overall the 96.2% of the log weekly gross wage of their non-disabled counterparts. Allowing for gender differences, the above percentage varies with men being slightly worse-off relative to women, 96% and 96.5% respectively.[21] In general, the results do not vary dramatically year by year for the disabled implying that there is no evidence that their position has considerably changed over time. It is worthy to note that there are no significant differences to the following figure in case we use the second indicator of disability status (see *Appendix, Annex C, Figure C1* for more details).

Figure 3: Log weekly gross wages in the United Kingdom 2005-2009



The means of the main variables, some of which are used later in the regression analysis, are presented in Table 5 and important differences among the disabled and non-disabled are in accordance with the expectations. Firstly, the non-disabled are typically younger than the disabled, but both disabled and non-disabled have mainly either an A-level or a higher degree qualification.[22] Furthermore, for each of the two indicators used and independently of gender, the disabled seem to be mostly not married or divorced. Considering any differences in race, men whose disability limits their daily activities are mostly Pakistani or tend to belong to the any other ethnic group category while women of the same group are mainly Indian. Gender differences are also observed regarding the hours of work with disabled men working almost 9 hours more than disabled women independently of the measure used to denote disability status. Finally, and confirming again our expectations the disabled, as defined using either of the two measures, prefer to work less hours and more part-time than their non-disabled counterparts for both genders.

[21] For more information on log weekly gross wage variations by gender see *Appendix, Annex C, Figures C2 and C3*.

[22] Non-disabled people are on average 2.5 years younger than the disabled who have reported that their disability limits the amount or type of work they can do (first measure).

Non-disabled people are on average 1.4 years younger than the disabled who have reported that their disability limits their daily activities (second measure).

Table 5: Descriptive Statistics[23]

| **Means[24]** | ***Males*** | | | | ***Females*** | | | |
|---|---|---|---|---|---|---|---|---|
| | ***First measure*** | ***Second measure*** | ***Both*** | ***None*** | ***First measure*** | ***Second measure*** | ***Both*** | ***None*** |
| **Variables** | | | | | | | | |
| **Log weekly gross wage** | 5.94 | 5.98 | 5.91 | 6.19 | 5.43 | 5.48 | 5.40 | 5.63 |
| **Age** | 43.5 | 42.7 | 43.3 | 41.0 | 44.0 | 43.6 | 44.7 | 41.6 |
| **Black** | 0 | 0 | 0 | 0.34 | 0.49 | 0.63 | 0.80 | 0.46 |
| **Indian** | 0.52 | 0.38 | 0.29 | 0.80 | 1.79 | 1.74 | 2.92 | 0.62 |
| **Pakistani** | 0.87 | 0.76 | 1.14 | 0.26 | 0 | 0 | 0 | 0.06 |
| **Bangladeshi** | 0 | 0 | 0 | 0.01 | 0 | 0 | 0 | 0.01 |
| **Chinese** | 0 | 0 | 0 | 0.14 | 0 | 0.16 | 0 | 0.11 |
| **Any other ethnic group** | 0.70 | 0.76 | 0.57 | 0.75 | 0.82 | 1.10 | 1.06 | 1.11 |
| **Separated** | 2.96 | 3.44 | 3.43 | 2.18 | 2.94 | 3.63 | 3.18 | 3.12 |
| **Divorced** | 8.89 | 9.54 | 9.43 | 7.05 | 14.68 | 13.09 | 13.26 | 1.89 |
| **Widowed** | 0.52 | 0.76 | 0.86 | 0.58 | 3.43 | 3.79 | 3.71 | 1.62 |
| **Never married** | 25.44 | 27.10 | 27.14 | 26.77 | 15.66 | 15.93 | 14.06 | 22.10 |
| **In a civil partnership** | 0.35 | 0.38 | 0.29 | 0.64 | 0 | 0.16 | 0 | 0.51 |
| **Graduates and higher degree** | 14.46 | 17.75 | 15.71 | 23.16 | 18.92 | 19.09 | 18.83 | 23.78 |
| **Higher school** | 32.23 | 29.58 | 30.29 | 35.01 | 32.46 | 34.54 | 35.28 | 28.41 |
| **None of these** | 18.82 | 18.70 | 18.86 | 11.48 | 14.52 | 15.46 | 15.38 | 12.12 |
| **Part-time** | 1.07 | 1.06 | 1.07 | 1.03 | 1.37 | 1.35 | 1.39 | 1.34 |
| **Hours of work** | 37.85 | 38.02 | 37.52 | 38.96 | 29.02 | 29.41 | 28.50 | 30.24 |
| | | | | | | | | |
| ***Observations*** | 574 | 524 | 350 | 8,487 | 613 | 634 | 377 | 8,943 |
| | 9,235 | | | | 9,813 | | | |

[23] Each column refers to a different group of people:

*First measure*: thought of as disabled according to the first measure $[M_1 = 1]$;

*Second measure*: thought of as disabled according to the second measure $[M_2 = 1]$;

*Both*: thought of as disabled according to both measures: $[M_1 = 1$ and $M_2 = 1]$;

*None*: thought of as non-disabled according to both measures: $[M_1 = 0$ and $M_2 = 0]$.

[24] The means of all variables except for those of log weekly gross wage, age, part-time and hours of work are multiplied by 100.

## IV. RESULTS

Using the British Household Panel Survey (BHPS) sample, we empirically test our proposed methodology and more specifically, whether and if so to what extent the respondents in a survey tend to misreport their disability status.

*The procedure*

We start by taking the BHPS data and obtaining the expected sample proportions and predicted wages for the four different groups of people: “First measure”, “Second measure”, “Both” and “None” as defined previously in the descriptive statistics. Then, we input these results into MATLAB and solve for the unknowns of the system of equations but each time imposing the different restriction(s) for the joint misclassification probabilities (“no correlation”, “no bias and no correlation” and “no bias”).

*The final solutions*

Initially and without correcting for the misclassification error the majority is thought of as non-disabled, using both measures to infer disability status, and accounting for about 91.5 percentage points. It is useful to note that the disabled report slightly more that their disability limits the type or amount of work they can do and less their daily activities - 6.2 and 6.1 percent of the total sample respectively. Considering any wage differentials for the four different groups, those who have reported “Yes” to both measures seem to face significantly lower wages, by almost 4.6 percentage points, than their non-disabled counterparts.

Table 5: Obtained values from the available BHPS data without correcting for the misclassification error

| **Expected sample proportions** | | | |
|---|---|---|---|
| ***First measure*** | ***Second measure*** | ***Both*** | ***None*** |
| 0.062 | 0.061 | 0.038 | 0.915 |
| **Predicted wages** | | | |
| ***First measure*** | ***Second measure*** | ***Both*** | ***None*** |
| 5.677 | 5.704 | 5.645 | 5.902 |

After imposing the no correlation assumption and correcting for the misclassification error in reporting disability status, we get the results presented in the table below. Noticeably, truly disabled tend to lie more when they report their disability status (underreport disability) compared to the truly non-disabled (overreport disability) since always the $p's > q's$. Also, truly disabled tend to underreport more that their disability limits their daily activities rather than the type or amount of work they can perform $(p_2 > p_1)$, whereas the truly non-disabled overreport more that their disability limits their daily activities $(q_2 > q_1)$. Finally, wage differences between the disabled and non-disabled are still apparent, with the truly disabled - as expected - facing lower wages than truly non-disabled (the coefficient $\beta$ is negative) and the baseline log weekly gross wage being approximately 5.90.

Table 6: Final solutions for the system - “no correlation”

| $t$ | $p_1$ | $p_2$ | $q_1$ | $q_2$ | $w$ | $\beta$ |
|---|---|---|---|---|---|---|
| 0.067 | 0.187 | 0.303 | 0.008 | 0.015 | 5.903 | -0.258 |

It is important to note that when assuming “no bias and no correlation”, we get a set of final solutions similar in magnitude - for most of the unknowns except for the two misclassification probabilities that are significantly lower - to the ones presented in Table 6. Though the truly non-disabled in this case are assumed to misreport disability by the same amount as the truly disabled for each of the two indicators used, both seem to lie less when reporting that their disability limits the type or amount of work they can do $(p_1 < p_2$ and $q_1 < q_2)$.

Table 7: Final solutions for the system - “no bias and no correlation”

| $t$ | $p_1$ | $p_2$ | $w$ | $\beta$ |
|---|---|---|---|---|
| 0.051 | 0.009 | 0.017 | 5.901 | -0.260 |

When imposing only the “no bias” assumption, the algorithm crashes and that is because in the way we have defined the system of equations it makes it very difficult, if not impossible for the programme to identify some of the unknowns, and more importantly the probability the respondent lies on both measures $p_{MM}$ (that probability is tiny).

As mentioned earlier in the analysis, in the presence of misclassification error the estimates of the variables of interest are biased (towards zero). In fact, we can directly check the above statement by running a simple OLS regression of the log weekly gross wages on the disability status, separately for each of the different measures used to infer disability. Thus, the model we estimate takes the form

$$lpayglw_i = a + bd_i + \varepsilon_i\,, \tag{21}$$

where $lpayglw_i$ refers to the log weekly gross wage of each individual $i$, $a$ is a constant, $d_i$ is a discrete variable taking two values $\{0,1\}$ if the respondent has reported being non-disabled and disabled respectively, $b$ is the associated parameter vector and $\varepsilon_i$ is the error term following a normal distribution $(0, \sigma_\varepsilon)$. For the purpose of this analysis we do not take into account the different factors that can determine wages such as age, gender, race, marital status and qualifications.

The estimated coefficients from the above regression are shown in Table 8; essentially, the constant in our regression corresponds to the wage the truly non-disabled receive or simply the baseline wage $w$.

Table 8: OLS estimates for each of the different measures used to infer disability status

| | $\_cons$ | $\hat{\beta}$ |
|---|---|---|
| **First measure** | 5.899 | -0.223 |
| **Second measure** | 5.897 | -0.193 |

Comparing the results from Table 6, Table 7 and Table 8 we can conclude that without correcting for the misclassification error, the constant and the estimated coefficients (in absolute values) differ, they are biased downwards, independently of the indicator used showing the extent of the measurement error in that specific variable. At last, comparing together the two measures, the difference for both estimates is more acute in case where the second measure is used to infer disability.

*Issues for future studies*

It would be useful for future studies to consider the case where the two measures used to infer disability status pick up something different such as short or long-term disability. So does it mean that the bounds methodology commonly used in the literature should be used again to measure the extent of misclassification error in that case? *Not necessarily*, unless there is a way to achieve point identification of a similar system of simultaneous equations when the $\lambda's$ (probabilities the true value of each of the two measures are equal to one) for each of the two measures are different and so the coefficients of the relevant variables are not by assumption equal.

## V. CONCLUSION

The paper using the concept of the fourfold table and obtaining a system of simultaneous equations from the available sample proportions and expected wages together with a restriction for the misclassification probabilities, examines whether it is possible to achieve point identification in the presence of measurement error in discrete variables. Considering the case of having *two* reported variables measured with error that pick up the *same* latent variable, thus the coefficients of the two indicators are assumed to be the same, it provides unique solutions for the system.

Empirically, using the British Household Panel Survey sample, we examine the extent of classification error in health related variables and in particular, on whether and if so to what extent people tend to misreport their disability status. Our results suggest that the probability of underreporting is greater than the probability of overreporting disability. Also, truly disabled tend to underreport more that their disability limits their daily activities rather than the type or amount of work they can perform, whereas the truly non-disabled overreport more that their disability limits their daily activities. Comparing the coefficients and the baseline wage before and after correcting for the misclassification error, we can conclude that they differ - in accordance with the definition of the measurement error - and mainly affecting those who report that their disability has work limitations.

## REFERENCES

Aigner, D. J. (1973) ‘Regression with a binary independent variable subject to errors of observation’, *Journal of Econometrics*, 2, p.49-59.

Baldwin, M. and Schumacher, E. (2002) ‘A note on job mobility among workers with disabilities’, *Industrial Relations*, 41, p.430-441.

Barron, B. (1977) ‘The effects of misclassification on the estimation of relative risk’, *Journal of Biometrics*, 33, p.414-418.

Berzofsky, M., Biermer, P. and Kalsbeek, W. (2008) ‘A brief history of classification error models’, Joint statistical meetings, American Statistical Association.

Bollinger, C. (1996) ‘Bounding the effects of measurement error in regressions involving dichotomous variables’, *Journal of Econometrics*, 73, p.387-399.

Bollinger, C. (2003) ‘Measurement error in human capital and the black-white wage gap’, *The Review of Econometrics and Statistics*, 85, p.578-585.

Bound, J. and Burkhauser, R. (1999) ‘Economic analysis of transfer programs targeted on people with disabilities’, *Handbook of Labour Economics*, 3C, p.3417-3528.

Bound, J., Brown, C. and Mathiowetz, N. (2001) ‘Chapter 59: Measurement error in survey data’, *Handbook of Econometrics*, 5, p.3705-3843.

Bowe, F. (1993) ‘Statistics, politics, and employment of people with disabilities’, *Journal of disability policy studies*, 4, p.84-91.

Bross, I. (1954) ‘Misclassification in 2x2 tables’, *Biometrics*, 10, p.478-486.

Burchardt, T. (2000) ‘The dynamics of being disabled’, Working paper, Centre for Analysis of Social Exclusion.

Diamond, E. and Lilienfield A. (1962) ‘Misclassification errors in 2x2 tables with one margin fixed: some further comments’, *American Journal of Public Health and the Nations Health*, 52, p.2106-2110.

Equality Act (2010), chapter 15, www.legislation.gov.uk.

Goldberg, J. (1975) ‘The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table’, *Journal of the American Statistical Association*, 70, p.561-567.

Greenland, S. and Kleinbaum, B. (1983) ‘Correcting for misclassification in two-way tables and matched-pair studies’, *International Journal of Epidemiology*, 12, p.93-97.

Hale, T. (2001) ‘The lack of a disability measure in today’s current population survey’, *Monthly Labour Review*, 124, p.37-58.

Healthy People (2010) ‘Summary measures of population health: Report of findings on methodologic and data issues’, U.S Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Hui, S. L. and Walter, S. D. (1980) ‘Estimating the error rates of diagnostic tests’, *Biometrics*, 36, p.167-171.

Höfler, M. (2005) ‘The effect of misclassification on the estimation of association: a review’, *International Journal of Methods in Psychiatric Research*, 14, p.92-101.

Horowitz, J. and Manski, C.F. (1995) ‘Identification and robustness with contaminated and corrupted data’, *Econometrica*, 63, p. 281-302.

Jones, M., Latreille, P. and Sloane, P. (2006) ‘Disability, gender, and the British labour market’, *Oxford Economic Papers*, 58, p.407-449.

Jones, M., Latreille, P. and Sloane, P. (2007) ‘Disability and work: A Review of the British Evidence’, *Estudion de Economia Aplicada*, 25, p.473-498.

Jones, M. (2009) ‘The employment effect of the disability discrimination act: evidence from the health survey for England’, *Labour*, 23, p.349-369.

Jones, M., and Sloane, P. (2009) ‘Disability and skill mismatch’, Discussion Paper, IZA.

Keys, A., and Kihlberg, J. (1963) ‘Effect of misclassification on estimated relative prevalence of a characteristic: Part I. Two populations infallibly distinguished. Part II. Errors in two variables’, *American Journal of Public Health and the Nations Health*, 53, p.1656-1665.

Klepper, S. (1984) ‘Bounding mean regressions when a binary regressor is mismeasured’, *Journal of Econometrics*, 37, p.343-359.

Kreider, B., (2006) ‘Partially identifying the prevalence of health insurance given contaminated sampling response error’, Working Paper, Department of Economics, Iowa State University.

Kreider, B. and Pepper, J. (2007) ‘Disability and employment: Reevaluating the evidence in light of reporting errors’, *Journal of the American Statistical Association*, 102, p.432-441.

Kreider, B. and Pepper, J. (2008) ‘Inferring disability status from corrupt data’, *Journal of Applied Econometrics*, 23, p.329-349.

Lohr, S. L. (1999) *Sampling: Design and analysis*, Boston: Houghton Mifflin.

Madden, D. (2004) ‘Labour market discrimination on the basis of health: an application to UK data’, *Applied Economics*, 36, p.421-442.

McGarry, K. (2002) ‘Health and Retirement: Do changes in health affect retirement expectations?’ Working Paper, National Bureau of Economic Research.

Molinary, F. (2008) ‘Partial identification of probability distributions with misclassified data’, *Journal of Econometrics*, 144, p.81-117.

Morrissey, M. J. and Spiegelman, D. (1999) ‘Matrix methods for estimating odds ratios with misclassified exposure data: Extensions and comparisons’, *Biometrics*, 55, p.338-344.

Myers, R. (1982) ‘Why do people retire from work early?’, *Social Security Bulletin*, 45, p.10-14.

Newell, D. (1962) ‘Errors in the interpretation of errors in epidemiology’, *American Journal of Public Health and Nations Health*, 52, p.1925-1928.

Phelps, E. (1972) ‘The statistical theory of racism and sexism’, *American Economic Review*, 62, p.659-661.

Rodgers III, W. (2006) *Handbook on the Economics of Discrimination*, Edward Elgar Publishing Limited.

Savoca, E. (2000) ‘Measurement errors in binary regressors: An application to measuring the effects of specific phychiatric diseases on earnings’, *Health Services and Outcomes Research Methodology*, 1, p.149-164.

Tenenbein, A. (1972) ‘A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection’ *Technometrics*, 14, p.187-202.

The Equality and Human Rights Commission (EHRC) (2010) *Equality Act 2010*, 1, Chapter 15.

Wallace, R. and Herzog, A. (1995) ‘Overview of the health measures in the health and retirement study’, *The Journal of Human Resources*, 30, p.S84-S107.

World Health Organization. (2008) *International Statistical Classification of Diseases and Related Health Problems*, 1, p.1-1207 (10th edition).

# APPENDIX

## *Annex A*

### 1. Useful definitions

*a. Two mismeasured variables*

- $p_{mM}$ : probability the individual is misclassified only in one of the two measures $(M_2)$ given that he truly possesses the characteristic $\left[\Pr(M_1 = 1, M_2 = 0 \mid T = 1)\right]$;
- $p_{MM}$ : probability the individual is misclassified in both measures given that he truly possesses the characteristic $\left[\Pr(M_1 = 0, M_2 = 0 \mid T = 1)\right]$;
- $q_{Mm}$ : probability the individual is misclassified only in one of the two measures $(M_1)$ given that he does not truly possess the characteristic $\left[\Pr(M_1 = 1, M_2 = 0 \mid T = 0)\right]$;
- $q_{mM}$ : probability the individual is misclassified only in one of the two measures $(M_2)$ given that he does not truly possess the characteristic $\left[\Pr(M_1 = 0, M_2 = 1 \mid T = 0)\right]$;
- $q_{mm}$ : probability the individual is not misclassified in none of the two measures given that he does not truly possess the characteristic $\left[\Pr(M_1 = 0, M_2 = 0 \mid T = 0)\right]$.

*b. One mismeasured variable*

- $\lambda$ : probability the "gold standard" is equal to one $\left[\Pr(G = 1)\right]$;
- $(1-\lambda)$ : probability the "gold standard" is equal to zero $\left[\Pr(G = 0)\right]$;
- $p$ : probability the individual reports not possessing the characteristic when he truly possesses it $\left[\Pr(M = 0 \mid G = 1)\right]$;
- $(1-p)$ : probability the individual reports possessing the characteristic when he truly possesses it $\left[\Pr(M = 1 \mid G = 1)\right]$;
- $q$ : probability the individual reports possessing the characteristic when he does not truly possess it $\left[\Pr(M = 1 \mid G = 0)\right]$;
- $(1-q)$ : probability the individual reports not possessing the characteristic when he does not truly possess it $\left[\Pr(M = 0 \mid G = 0)\right]$;
- $A'$ : the number of people that have reported "Yes" to $G$ and "Yes" to $M$ ;
- $B'$ : the number of people that have reported "Yes" to $G$ and "No" to $M$ ;
- $C'$ : the number of people that have reported "No" to $G$ and "Yes" to $M$ ;
- $D'$ : the number of people that have reported "No" to $G$ and "No" to $M$ .

*c. Expected wages are simply a function of the expected sample proportions…*

| | ***Reported variable 2 (*** $M_2$ ***)*** | |
|---|---|---|
| ***Reported variable 1*** <br> ***(*** $M_1$ ***)*** | **“Yes”** | **“No”** |
| **“Yes”** | $w+\frac{\beta\lambda p_{mm}}{\underbrace{\lambda p_{mm}+(1-\lambda)q_{MM}}_{A}}$ | $w+\frac{\beta\lambda p_{mM}}{\underbrace{\lambda p_{mM}+(1-\lambda)q_{Mm}}_{B}}$ |
| **“No”** | $w+\frac{\beta\lambda p_{Mm}}{\underbrace{\lambda p_{Mm}+(1-\lambda)q_{mM}}_{C}}$ | $w+\frac{\beta\lambda p_{MM}}{\underbrace{\lambda p_{MM}+(1-\lambda)q_{mm}}_{D}}$ |

### 2. Nonlinear systems of simultaneous equations depending on the different assumptions imposed for the joint misclassification probabilities

a. *No bias (full identification)*

$$\lambda p_{mm}+(1-\lambda)q_{MM}=A,$$
$$\lambda p_{mM}+(1-\lambda)q_{Mm}=B,$$
$$\lambda p_{Mm}+(1-\lambda)q_{mM}=C,$$
$$w+\frac{\beta\lambda p_{mm}}{\lambda p_{mm}+(1-\lambda)q_{MM}}=E,$$
$$w+\frac{\beta\lambda p_{mM}}{\lambda p_{mM}+(1-\lambda)q_{Mm}}=F,$$
$$w+\frac{\beta\lambda p_{Mm}}{\lambda p_{Mm}+(1-\lambda)q_{mM}}=G,$$
$$w+\frac{\beta\lambda\left(1-p_{Mm}-p_{mM}-p_{mm}\right)}{\lambda\left(1-p_{Mm}-p_{mM}-p_{mm}\right)+(1-\lambda)\left(1-q_{Mm}-q_{mM}-q_{MM}\right)}=H,$$
$$1-p_{mM}-p_{mm}=q_{Mm}+q_{MM},$$
$$1-p_{Mm}-p_{mm}=q_{mM}+q_{MM}.$$

*b. No correlation (full identification)*

$$\lambda(1-p_1)(1-p_2)+(1-\lambda)q_1q_2=A,$$
$$\lambda(1-p_1)p_2+(1-\lambda)q_1(1-q_2)=B,$$
$$\lambda p_1(1-p_2)+(1-\lambda)(1-q_1)q_2=C,$$
$$w+\frac{\beta\lambda(1-p_1)(1-p_2)}{\lambda(1-p_1)(1-p_2)+(1-\lambda)q_1q_2}=E,$$
$$w+\frac{\beta\lambda(1-p_1)p_2}{\lambda(1-p_1)p_2+(1-\lambda)q_1(1-q_2)}=F,$$

$$w+\frac{\beta\lambda p_1(1-p_2)}{\lambda p_1(1-p_2)+(1-\lambda)(1-q_1)q_2}=G\,,$$
$$w+\frac{\beta\lambda p_1 p_2}{\lambda p_1 p_2+(1-\lambda)(1-q_1)(1-q_2)}=H\,.$$

*c. No bias and no correlation (over identification)*

$$\lambda(1-p_1)(1-p_2)+(1-\lambda)p_1 p_2=A,$$
$$\lambda(1-p_1)p_2+(1-\lambda)p_1(1-p_2)=B,$$
$$\lambda p_1(1-p_2)+(1-\lambda)(1-p_1)p_2=C\,,$$
$$w+\frac{\beta\lambda(1-p_1)(1-p_2)}{\lambda(1-p_1)(1-p_2)+(1-\lambda)p_1 p_2}=E\,,$$
$$w+\frac{\beta\lambda(1-p_1)p_2}{\lambda(1-p_1)p_2+(1-\lambda)p_1(1-p_2)}=F\,,$$
$$w+\frac{\beta\lambda p_1(1-p_2)}{\lambda p_1(1-p_2)+(1-\lambda)(1-p_1)p_2}=G\,,$$
$$w+\frac{\beta\lambda p_1 p_2}{\lambda p_1 p_2+(1-\lambda)(1-p_1)(1-p_2)}=H\,.$$

## 3. System of nonlinear simultaneous equations using a simpler notation (I)

*a. General case*

We simply redefine the joint misclassification probabilities as follows;

$$p_{MM}=\Delta p\,,$$
$$p_{Mm}=p_1-\Delta p\,,$$
$$p_{mM}=p_2-\Delta p\,,$$
$$p_{mm}=(1-p_1-p_2+\Delta p)\,.$$

and similarly for the $q's$ :

$$q_{MM}=\Delta q\,,$$
$$q_{Mm}=q_1-\Delta q\,,$$
$$q_{mM}=q_2-\Delta q\,,$$
$$q_{mm}=(1-q_1-q_2+\Delta q)\,.$$

Therefore, our system becomes

$$\lambda(1-p_1-p_2+\Delta p)+(1-\lambda)\Delta q=A,$$
$$\lambda(p_2-\Delta p)+(1-\lambda)(q_1-\Delta q)=B,$$
$$\lambda(p_1-\Delta p)+(1-\lambda)(q_2-\Delta q)=C\,,$$

$$w+\frac{\beta\lambda(1-p_1-p_2+\Delta p)}{\lambda(1-p_1-p_2+\Delta p)+(1-\lambda)\Delta q}=E\,,$$
$$w+\frac{\beta\lambda(p_2-\Delta p)}{\lambda(p_2-\Delta p)+(1-\lambda)(q_1-\Delta q)}=F\,,$$
$$w+\frac{\beta\lambda(p_1-\Delta p)}{\lambda(p_1-\Delta p)+(1-\lambda)(q_2-\Delta q)}=G\,,$$
$$w+\frac{\beta\lambda\Delta p}{\lambda\Delta p+(1-\lambda)(1-q_1-q_2+\Delta q)}=H\,.$$

*b. No bias (full identification)*
Since under no bias assumption; $p_1=q_1$, and $p_2=q_2$, then

$$\lambda(1-p_1-p_2+\Delta p)+(1-\lambda)\Delta q=A\,,$$
$$\lambda(p_2-\Delta p)+(1-\lambda)(p_1-\Delta q)=B\,,$$
$$\lambda(p_1-\Delta p)+(1-\lambda)(p_2-\Delta q)=C\,,$$
$$w+\frac{\beta\lambda(1-p_1-p_2+\Delta p)}{\lambda(1-p_1-p_2+\Delta p)+(1-\lambda)\Delta q}=E\,,$$
$$w+\frac{\beta\lambda(p_2-\Delta p)}{\lambda(p_2-\Delta p)+(1-\lambda)(p_1-\Delta q)}=F\,,$$
$$w+\frac{\beta\lambda(p_1-\Delta p)}{\lambda(p_1-\Delta p)+(1-\lambda)(p_2-\Delta q)}=G\,,$$
$$w+\frac{\beta\lambda\Delta p}{\lambda\Delta p+(1-\lambda)(1-p_1-p_2+\Delta q)}=H\,.$$

Thus, we finally have a system of 7 equations with 7 unknowns $\{p_1,p_2,\Delta p,\Delta q,\lambda,w,\beta\}$.

*c. No correlation (full identification)*
As with no correlation assumption it should hold that $\Delta p=p_1p_2$, and $\Delta q=q_1q_2$, then

$$\lambda(1-p_1-p_2+p_1p_2)+(1-\lambda)q_1q_2=A\,,$$
$$\lambda(p_2-p_1p_2)+(1-\lambda)(q_1-q_1q_2)=B\,,$$
$$\lambda(p_1-p_1p_2)+(1-\lambda)(q_2-q_1q_2)=C\,,$$
$$w+\frac{\beta\lambda(1-p_1-p_2+p_1p_2)}{\lambda(1-p_1-p_2+p_1p_2)+(1-\lambda)q_1q_2}=E\,,$$
$$w+\frac{\beta\lambda(p_2-p_1p_2)}{\lambda(p_2-p_1p_2)+(1-\lambda)(q_1-q_1q_2)}=F\,,$$
$$w+\frac{\beta\lambda(p_1-p_1p_2)}{\lambda(p_1-p_1p_2)+(1-\lambda)(q_2-q_1q_2)}=G\,,$$

$$w+\frac{\beta\lambda p_1 p_2}{\lambda p_1 p_2+(1-\lambda)(1-q_1-q_2+q_1q_2)}=H\ .$$

Hence, we end up with a system of 7 equations with 7 unknowns $\{p_1, p_2, q_1, q_2, \lambda, w, \beta\}$.

*d. No bias and no correlation (over identification)*

Since under no bias and no correlation assumptions $\Delta p = p_1 p_2$, $\Delta q = q_1 q_2$, $p_1 = q_1$, and $p_2 = q_2$, it should hold that $\Delta p = \Delta q$, so

$$\lambda(1-p_1-p_2)+(1-\lambda)p_1p_2=A,$$
$$\lambda(p_2-p_1)+p_1-p_1p_2=B,$$
$$\lambda(p_1-p_2)+p_2-p_1p_2=C,$$
$$w+\frac{\beta\lambda(1-p_1-p_2+p_1p_2)}{\lambda(1-p_1-p_2)+p_1p_2}=E\ ,$$
$$w+\frac{\beta\lambda(p_2-p_1p_2)}{\lambda(p_2-p_1)+p_1-p_1p_2}=F\ ,$$
$$w+\frac{\beta\lambda(p_1-p_1p_2)}{\lambda(p_1-p_2)+p_2-p_1p_2}=G\ ,$$
$$w+\frac{\beta\lambda p_1p_2}{(1-p_1-p_2+p_1p_2)-\lambda(1-p_1-p_2)}=H\ .$$

Therefore, we have a system of 7 equations with 5 unknowns $\{p_1, p_2, \lambda, w, \beta\}$.

## 4. System of nonlinear simultaneous equations using a simpler notation (II)

*a. General case*

We redefine the different groups of people as:

- Those who have reported “Yes” to the first measure $(M_1)$;
- Those who have reported “Yes” to the second measure $(M_2)$;
- Those who have reported “Yes” to both measures;
- Those who have reported “No” to both measures.

$$\lambda(1-p_1)+(1-\lambda)q_1=SP1,$$
$$\lambda(1-p_2)+(1-\lambda)q_2=SP2,$$
$$\lambda(1-p_1-p_2+p_{12})+(1-\lambda)q_{12}=SP3,$$
$$w+\frac{\beta\lambda(1-p_1)}{\lambda(1-p_1)+(1-\lambda)q_1}=EW1,$$
$$w+\frac{\beta\lambda(1-p_2)}{\lambda(1-p_2)+(1-\lambda)q_2}=EW2,$$

$$w+\frac{\beta\lambda(1-p_1-p_2+p_{12})}{\lambda(1-p_1-p_2+p_{12})+(1-\lambda)q_{12}}=EW3,$$
$$w+\frac{\beta\lambda p_{12}}{\lambda p_{12}+(1-\lambda)(1-q_1-q_2+q_{12})}=EW4.$$

Hence, using the primary notation, each of the observed proportions and predicted wages correspond to the following;

$$SP_1=A+B,$$
$$SP_2=A+C,$$
$$SP_3=A,$$
$$SP_4=D,$$
$$EW_1\cong\frac{E+F}{2},$$
$$EW_2\cong\frac{E+G}{2},$$
$$EW_3=E,$$
$$EW_4=H.$$

*b. No bias (full identification)*

$$\lambda-2\lambda p_1+p_1=SP1,$$
$$\lambda-2\lambda p_2+p_2=SP2,$$
$$\lambda(1-p_1-p_2+p_{12})+(1-\lambda)q_{12}=SP3,$$
$$w+\frac{\beta\lambda(1-p_1)}{\lambda(1-p_1)+(1-\lambda)q_1}=EW1,$$
$$w+\frac{\beta\lambda(1-p_2)}{\lambda(1-p_2)+(1-\lambda)q_2}=EW2,$$
$$w+\frac{\beta\lambda(1-p_1-p_2+p_{12})}{\lambda(1-p_1-p_2+p_{12})+(1-\lambda)q_{12}}=EW3,$$
$$w+\frac{\beta\lambda p_{12}}{\lambda p_{12}+(1-\lambda)(1-q_1-q_2+q_{12})}=EW4.$$

*c. No correlation (full identification)*

$$\lambda(1-p_1)+(1-\lambda)q_1=SP1,$$
$$\lambda(1-p_2)+(1-\lambda)q_2=SP2,$$
$$\lambda(1-p_1-p_2+p_1p_2)+(1-\lambda)q_1q_2=SP3,$$
$$w+\frac{\beta\lambda(1-p_1)}{\lambda(1-p_1)+(1-\lambda)q_1}=EW1,$$

$$w + \frac{\beta\lambda(1-p_2)}{\lambda(1-p_2)+(1-\lambda)q_2} = EW2\,,$$

$$w + \frac{\beta\lambda(1-p_1-p_2+p_1p_2)}{\lambda(1-p_1-p_2+p_1p_2)+(1-\lambda)q_1q_2} = EW3\,,$$

$$w + \frac{\beta\lambda p_1p_2}{\lambda p_1p_2+(1-\lambda)(1-q_1-q_2+q_1q_2)} = EW4\,.$$

*d. No bias and no correlation (over identification)*

$$\lambda - 2\lambda p_1 + p_1 = SP1,$$

$$\lambda - 2\lambda p_2 + p_2 = SP2\,,$$

$$\lambda(1-p_1-p_2)+p_1p_2 = SP3\,,$$

$$w + \frac{\beta\lambda(1-p_1)}{t-2\lambda p_1+p_1} = EW1,$$

$$w + \frac{\beta\lambda(1-p_2)}{\lambda-2\lambda p_2+p_2} = EW2\,,$$

$$w + \frac{\beta\lambda(1-p_1-p_2+p_1p_2)}{\lambda(1-p_1-p_2)+p_1p_2} = EW3,$$

$$w + \frac{\beta\lambda p_1p_2}{(1-\lambda)(1-p_1-p_2)+p_1p_2} = EW4\,.$$

## *Annex B*

### Classification of specific health problems in the sample (based on ICD-10)

a. Physical

- *arm/leg/hands* (arthritis, rheumatism etc.);
- *sight* (other than needing glasses to read normal size print);
- *hearing*;
- *skin/allergies*;
- *chest/breathing* (asthma, bronchitis);
- *heart/blood pressure* or blood circulation problems;
- *stomach/liver/kidney* or digestive problems;
- *diabetes*;
- *migraine* or frequent headaches.

b. Mental

- *anxiety/depression* or bad nerves;
- *alcohol/drugs*.

c. Mixed

- *epilepsy*;
- *other* (cancer, stroke, etc.).

## *Annex C*

### 1. Useful figures

Figure C1: Log weekly gross wages in the United Kingdom 2005-2009

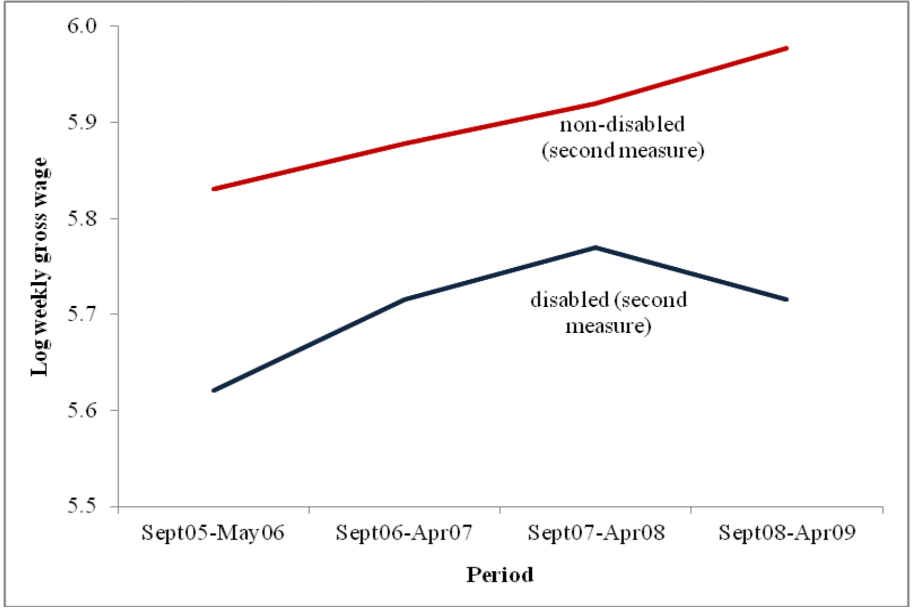


Figure C2: Log weekly gross wages in the United Kingdom 2005-2009 by gender using the 1st measure to infer disability status

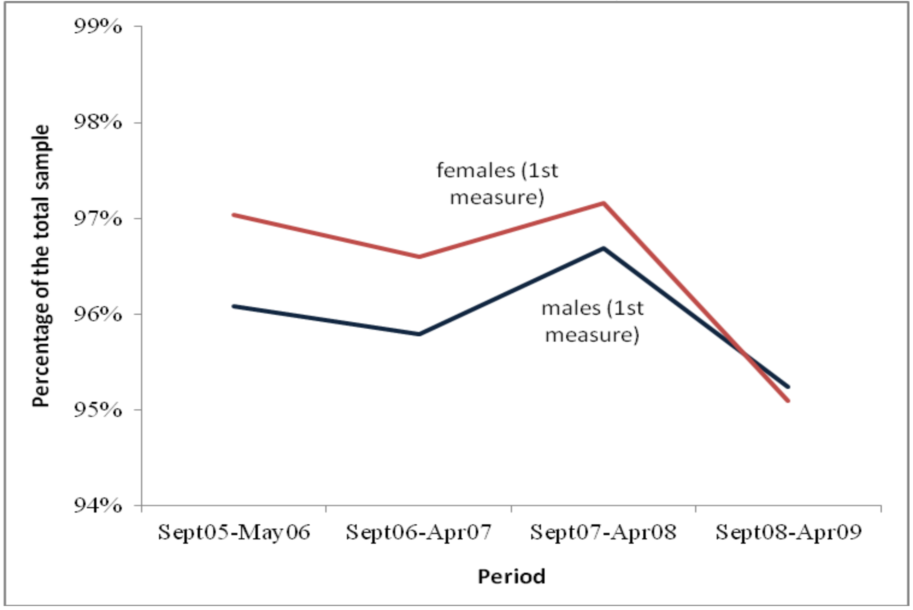
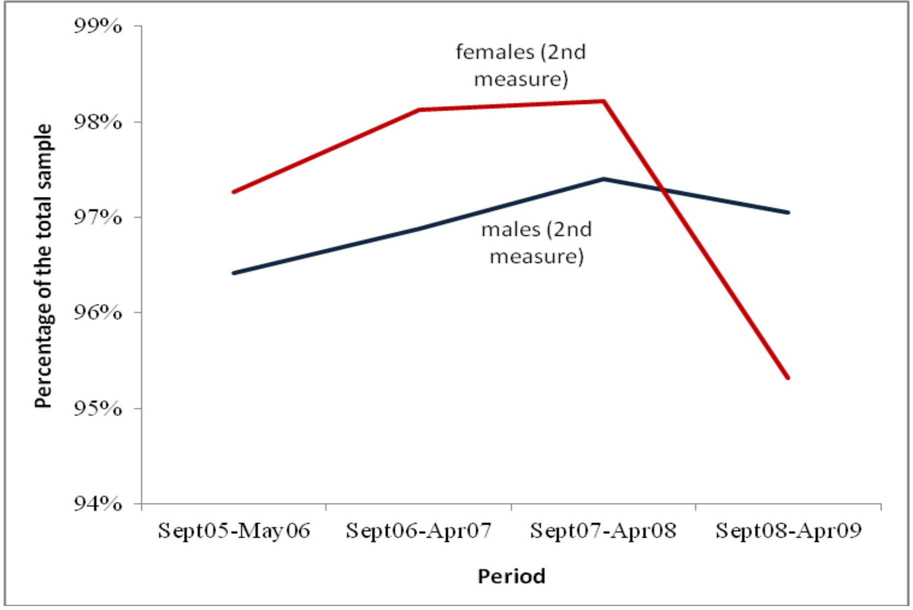


Figure C3: Log weekly gross wages in the United Kingdom 2005-2009 by gender using the 2nd measure to infer disability status


Percentage of the total sample
females (2nd measure)
males (2nd measure)
Sept05-May06
Sept06-Apr07
Sept07-Apr08
Sept08-Apr09
Period