

Kristensen, Nicolai; Westergård-Nielsen, Niels Christian

Working Paper

A large-scale validation study of measurement errors in longitudinal survey data

IZA Discussion Papers, No. 2329

Provided in Cooperation with:

IZA – Institute of Labor Economics

Suggested Citation: Kristensen, Nicolai; Westergård-Nielsen, Niels Christian (2006) : A large-scale validation study of measurement errors in longitudinal survey data, IZA Discussion Papers, No. 2329, Institute for the Study of Labor (IZA), Bonn, <https://nbn-resolving.de/urn:nbn:de:101:1-20090825339>

This Version is available at:

<https://hdl.handle.net/10419/33904>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

IZA DP No. 2329

A Large-Scale Validation Study of Measurement Errors in Longitudinal Survey Data

Nicolai Kristensen
Niels Westergaard-Nielsen

September 2006

A Large-Scale Validation Study of Measurement Errors in Longitudinal Survey Data

Nicolai Kristensen

Aarhus School of Business and CCP

Niels Westergaard-Nielsen

*Aarhus School of Business, CCP
and IZA Bonn*

Discussion Paper No. 2329
September 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit company supported by Deutsche Post World Net. The center is associated with the University of Bonn and offers a stimulating research environment through its research networks, research support, and visitors and doctoral programs. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ABSTRACT

A Large-Scale Validation Study of Measurement Errors in Longitudinal Survey Data^{*}

In this paper, we analyze measurement and classification errors in several key variables, including earnings and educational attainment, in a matched sample of survey and administrative longitudinal data. The data, spanning 1994-2001 and covering all sectors in the Danish economy, are much more comprehensive than usually seen in validation studies. Measurement errors in earnings are found to be much larger than reported in previous studies limited to one single firm. Individuals who attrite from the panel report their earnings significantly less accurate than individuals who are observed throughout the entire sampling period. Furthermore, females are found to report their earnings significantly more precise than males, part-time workers report significantly less accurate than full-time workers and low-income workers report significantly less accurate than workers with relatively higher income. Classification errors in categorical variables are found to be of about the same magnitude as previously found in the literature. We analyze whether response error in one variable makes it more likely that the same respondent will report other variables with error but do not find support for this hypothesis.

JEL Classification: J24, J31, I2, J28

Keywords: measurement error, classification error, validation

Corresponding author:

Niels Westergaard-Nielsen
Aarhus School of Business
Department of Economics
Prismet, Silkeborgvej 2
DK-8000 Aarhus C
Denmark
E-mail: nwn@asb.dk

^{*} We thank Rasmus Andersen for excellent research assistance; Hans Bay for providing the Danish ECHP data; Søren Leth-Sørensen and Jørn Schmidt Statistics Denmark for providing a matched survey-register panel; Susan Stilling for carefully reading the manuscript, and Paul Bingley for helpful comments.

1 Introduction

The aim of this paper is to contribute to the important but often neglected field of measurement and classification errors. With the availability of greater and less costly computational power, the number of empirical studies based on survey data has increased almost exponentially. Although the empirical literature on measurement errors has grown in recent years, still only relatively few and limited studies exist on the extent and importance of measurement errors in survey data.

To date, the empirical literature on measurement and classification errors is limited by the lack of good validation data; both in terms of quality and quantity. For instance, several seminal papers turn around validation data from one single US company with two cross-section data sets, collected four years apart. These data are based on company records of some 450 employees and on a panel with two waves for about 275 of these employees (Duncan and Hill, 1985, Bound et al., 1994, Pischke, 1995). Bound and Krueger (1991) analyze much more comprehensive data from the CPS matched with Social Security Administration data, and the same data set was later expanded to include women who were not head of households, cf. Bollinger (1998). Nevertheless, these data only cover two waves.

All the studies mentioned above are based on data from the US. Some studies of European data do exist though. Battistin and Sianesi (2005) analyze classification errors in educational attainment using data from the British National Child Development Survey, a panel with repeated measures of individual educational attainment which is supplemented with administrative school files.

Several unpublished validation studies exist for the Scandinavian countries which combine administrative records with survey data, cf. Epland and

Kirkeberg (2002) for Norway, Jørgensen (1998) for Denmark and Nordberg et al. (2001) for Finland.²

Most existing studies focus either on measurement errors in earnings or on education; not on both in the same study, and they almost never analyze measurement or classification errors in other important variables. Furthermore, they rarely cover all sectors in the economy and are usually also limited to cross-section data or very short panels.

The purpose of this paper is to contribute along exactly these dimensions. We analyze measurement error and classification error in *several* key variables, including earnings and educational attainment, in a matched sample of survey and administrative longitudinal data. The data span the period 1994-2001 and cover all sectors in the Danish economy. We limit the sample to employees only.

Throughout this study, we consider the administrative records as validation data as we believe these data have a very high quality. However, despite their high quality they too may be prone to error. This issue is discussed in detail later on. Hence, the validation data are from administrative records, while the survey data are the Danish component of the ECHP.³ The fact that ECHP data have been so widely applied in empirical studies during the last decade makes this study even more relevant.

Participants in the Danish ECHP survey were randomly selected for personal interviews, and all responses from the ECHP interviews have been matched with administrative records from Statistics Denmark. This means that, contrary to other studies (e.g. Jäckle et al., 2004), respondents did not have to give their consent to the matching of survey and validation data which is something that otherwise, potentially, could have biased the match towards lower measurement

² The papers mentioned here constitute only a small fraction of studies on measurement and validation errors. See Bound et al. (2001) for a survey.

³ European Community Household Panel (ECHP), collected for 15 countries 1994-2001, and administered by Eurostat; see <http://epp.eurostat.cec.eu.int>.

error. Hence, the validation data allow us to make a so-called "complete record check" (Bound et al. 2001, p. 3741).⁴

We find measurement errors in earnings to be much larger than usually found in studies limited to one single firm. On the other hand, we confirm previous results stating that there is substantial mean-reversion in earnings (e.g. Bound and Krueger, 1991), which means that measurement errors in earnings are non-classical, cf. Bound et al. (2001).

Individuals who attrite from the panel are found to report their earnings significantly less accurately than individuals who are observed throughout the entire sampling period. This finding corroborates previous results found by Bollinger (1998). We also find that females report significantly better than males; part-time workers report significantly less accurately than full-time workers, and low-income workers report significantly less accurately than workers with relatively higher income. Lastly, measurement errors in earnings are found to be stable over the business cycle and across sectors of the economy.

Classification errors in categorical variables are found to be of about the same magnitude as previously found in the literature. We analyze whether response errors in one variable make it more likely that the same respondent will report other variables with errors as well but do not find support for this.

This paper proceeds as follows: In Section 2, we describe our two sources of data and compare with population averages in order to assess how representative the data are. In Section 3, we analyze measurement errors in earnings. In Section 4, we turn to categorical variables, notably educational attainment, firm size and industry, and analyze the extent of classification errors in these important variables that often play a key role in applied micro-econometric studies. A conclusion follows in Section 5.

⁴ As noted by Bound et al. (2001, p. 3742): "A complete record check [...] provides the best means for assessing both under-reporting as well as over-reporting. However, such studies are rare..."

2 Data

As mentioned in the introductory section, the survey data we seek to validate in this study originate from the Danish version of the ECHP data. These data were collected by the Survey Department at the Danish National Institute of Social Research. This department is highly specialized and collects and processes data for use by researchers, public authorities and private organizations and enterprises, conducting almost 90,000 interviews per year.

The mode of collection was face-to-face interviews using pencil and paper. The respondent is always the person about whom the interview is concerned, i.e. family members can not answer on behalf of each other. All interviewers are aged 30 or more and are trained and experienced enumerators. An attempt is made to allocate an interviewer to the same household during consecutive years of the panel. Some attrition exists. We explore whether or not this attrition is random.

Validation data are from the IDA administrative database maintained by Statistics Denmark. The extract of IDA that we have access to is a longitudinal database that contains information about *all* individuals aged 15 to 74 (demographic characteristics, education, labor market experience, tenure and earnings) and employees in *all* workplaces in Denmark during the period 1980-2001. This information has been collected by merging information from several registers in Statistics Denmark with the help of unique identification numbers for individuals and workplaces. Persons and workplaces are matched at the end of November each year. Consequently, only changes of employment November-to-November are accounted for, not intervening changes. We have only included workers who have their main occupation with an employer. The background data for IDA consist of various registers supplemented with data from the latest census in 1970. Thus, data on education come from the census in 1970 and after that from reports from all educational institutions on their current population of students and their completion degree.

Table 1 (overleaf) shows the sample size in the matched administrative-survey panel data set, and how it diminishes as various restrictions are imposed. The

final sample size has 16,748 observations. The restrictions are mainly imposed in order to ensure that discrepancies between the two data sets can be attributed to measurement errors. Some of the restrictions imposed, such as an annual income above USD 1,500, may not be strictly necessary but concur with a data cleaning procedure that eliminates observations that clearly cannot be true. As revealed by Table 1, we also limit the sample to employees only, and hence we exclude students, retirees, self-employed and individuals outside the labor force. Without the restrictions imposed in Table 1 we would be uncertain of the source of error.

In the analysis of categorical variables (education, firm size and industrial classification), the sample size diminishes further due to missing observations and/or difficulties in matching the two sources. This is discussed in more detail in Section 4.

In order to evaluate the quality of this matched sample, we compare the sample means with population means for various key variables, cf. Appendix B. The mean age and share of males in the sample is slightly over-represented, which also results in a slightly higher mean annual income in the sample vis-à-vis the population. Furthermore, Jutland, the main part of Denmark (geographically), is slightly over-represented. In order to evaluate whether attrition is an issue to be concerned about, the sample means are also matched to population means for 1994 (education 1998) and 2001, respectively. This exercise reveals that the sample mean age is 6.5 years higher in 2001 vis-à-vis the population, while the 1994 sample age distribution largely mirrors the population age distribution in 1994. This development is a direct consequence of sample attrition and lack of additional respondents included in the survey sample.

However, the overall picture is that the sample fairly well represents the overall population.

Table 1 Evolution of the sample

	Sample size remaining	Percent of previous row	Cumulative percentage elimination
1 Starting point	36,795	.	0.00%
2 Monthly (administrative) income * 12 > DKK 10,000 (USD 1,500)	35,666	96.93%	3.07%
3 Annual income > DKK 10,000 (USD 1,500)	24,372	68.33%	33.76%
4 Age between 18-67	23,872	97.95%	35.12%
5 Administrative hours above ($\frac{1}{2}$ * full time) and below (2 * full time)	20,053	84.00%	45.50%
6 Employees only, i.e. no self-employed, students or retired	18,330	91.41%	50.18%
7 Require that the respondent works more than 15H per week (ECHP)	17,633	96.20%	52.08%
8 Require that the respondent has provided a monthly wage	17,332	98.29%	52.90%
9 Eliminating respondents who have changed job after November 30th	17,323	99.95%	52.92%
10 Eliminating respondents who do not classify themselves as employees	16,748	96.68%	54.48%

3 Measurement Errors in Earnings Data

3.1 Earnings Comparison

Information on earnings measures is often considered to be notoriously flawed with errors. While the reasons may be manifold the main reason probably is that questions about income and earnings are, by most people, considered to be rather sensitive. Furthermore, respondents are likely to report their income in round numbers or may simply have difficulties in remembering the exact amount. This may be of particular concern in cases where the respondent is asked, in retrospect, to provide total income last year.

In the data at hand, both the survey and the validation data, inform about monthly earnings in the main occupation at the time of the interview as well as the total annual income last year is available. Details about these variables are given in Appendix A.

Figure 1, overleaf, shows the distribution of measurement errors in earnings for all employees in the sample as well as for selected sub-groups. In line with the literature, we compute measurement errors in earnings as the difference in logarithmic earnings, thus asserting that the error is a relative measure.

The spike in the distribution of monthly gross earnings is around zero (plot 1), but the distribution still has slightly more density to the left of zero meaning that the administrative records tend to be slightly higher than the survey records (median= -0.013). This asymmetry largely disappears when we condition on the time of the interview. The administrative records include annual earnings due to employment as of end-of-November while the survey is carried out throughout the year. Hence, although the sample is conditioned on the respondents having the same job, they may have received a wage increase in-between the two observations. When we only include observations from interviews in November-December (plot 4), the tendency of administrative records to be higher than survey records disappears (median difference = -0.003). Still, the variance is left almost unchanged (0.042 for all and 0.047 for November-December).

By splitting the sample into part-time workers (plot 6) and full-time workers (plot 5) and leaving out workers who work significantly more than full time, we obtain a sub-sample of workers who are expected to have a higher-than-average stable work- and earnings path and who may well be in a better position to report their earnings.

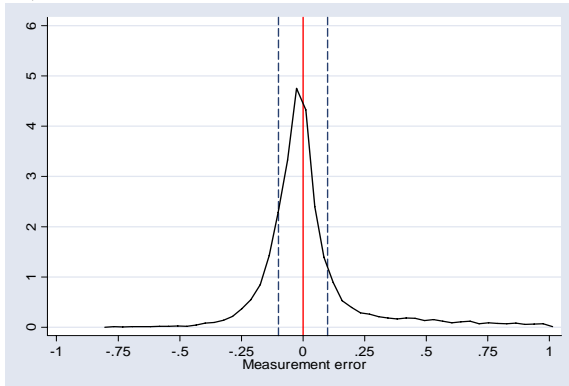
The bottom part of Figure 1 indicates that this is the case. Now, not only is the mean almost at zero (although we haven't conditioned on the month of the interview) but in addition, the variance is 0.019, i.e. a reduction of more than 50 percent.

Part-time workers, on the other hand, have a variance in the measurement error of 0.085, i.e. about twice the size of the variance on the overall sample. The density distribution has a large mass to the right, indicating that among part-time workers survey earnings are over-reported (median= 0.056).

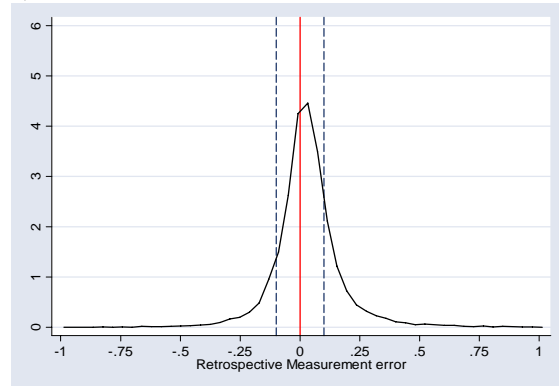
Finally, we look at measurement errors in annual total income, reported in retrospect (plot 2). The density is skewed to the right, which again indicates that respondents have a tendency of over-reporting their retrospective income. However, the median value at 0.025 is not alarming and the variance (=0.021) is actually smaller than for monthly earnings.

Figure 1 Measurement errors in gross earnings

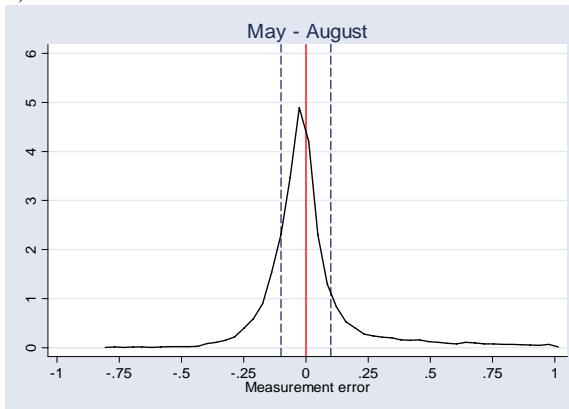
PLOT 1
*Monthly earnings *12*
 16,182 observations



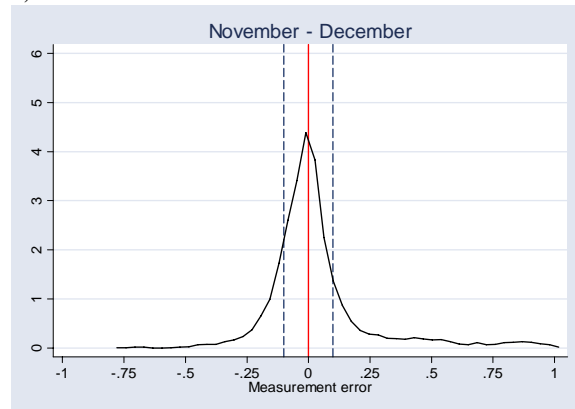
PLOT 2
Annual income, retrospect
 8,152 observations



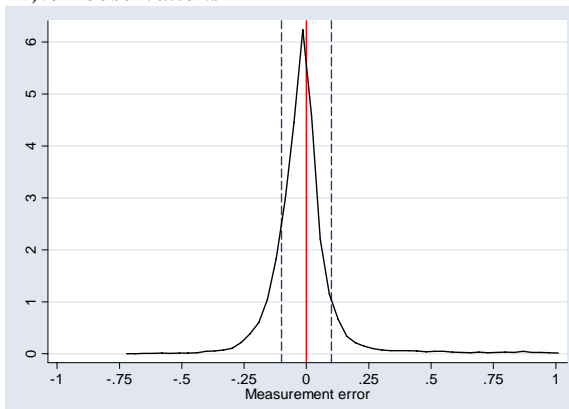
PLOT 3
*Monthly earnings*12, time of interview May-August*
 8,323 observations



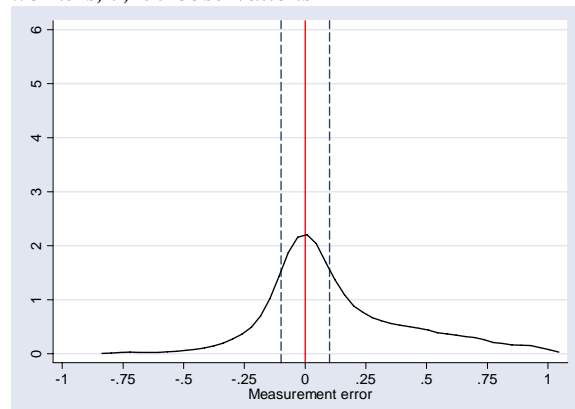
PLOT 4
*Monthly earnings*12, time of interview Nov-Dec*
 2,559 observations



PLOT 5
*Monthly earnings*12, Full time workers*
 10,091 observations



PLOT 6
*Monthly earnings*12, "Less than full time" workers, 3,273 observations*



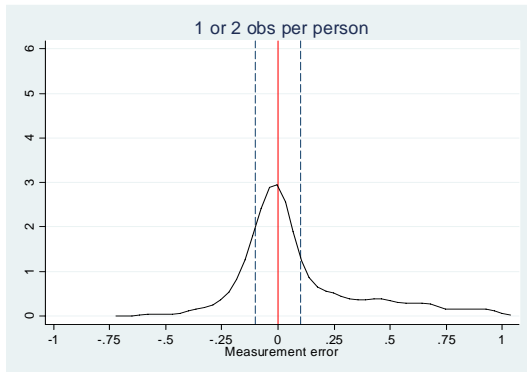
Note: The measurement error is calculated as $\ln(\text{survey gross earnings}) - \ln(\text{administrative gross earnings})$. Plot 2 (retrospective measurement error) is calculated using total annual gross *income*. Confidence bounds (not shown) are very tight.

Some degree of attrition exists in most longitudinal surveys. Usually, panel models are estimated on unbalanced data sets and attrition is therefore generally not considered particularly troublesome. However Bollinger (1998) finds, using the same data and extending the work of Bound and Krueger (1991), those respondents that are present in both waves of his two-wave panel respond with greater accuracy than respondents only observed once, i.e. who are observed only in a cross-section. A similar finding emerges from Figure 2.

Figure 2 Measurement errors in gross earnings, by number of observations per individual

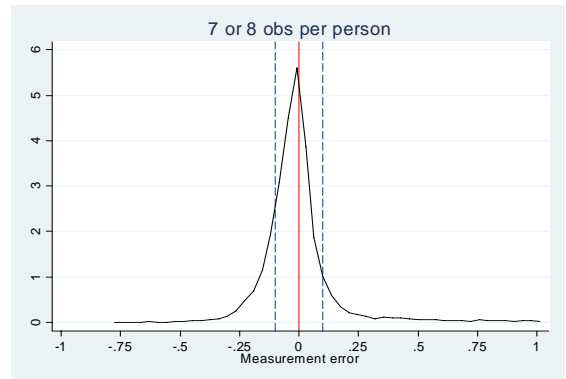
Plot 1

Monthly earnings*12, 1 or 2 observations per individual
7,641 observations



Plot 2

Monthly earnings*12, 7 or 8 observations per individual
1,878 observations



Note: The measurement error is calculated as $\ln(\text{survey gross earnings}) - \ln(\text{administrative gross earnings})$.

Clearly, the measurement error in earnings is more severe for individuals who only enter the survey panel once or twice compared to individuals who are observed 7 or 8 times during the 1994-2001 period. Similar densities emerge if the sample is limited to 1 observation (8 observations) per individual. The reason for the seemingly higher reliability of responses by panel-participants is not clear. For instance, it could simply be related to age or some other covariate. We explore this issue through simple OLS-regressions below.

Formal tests of equal distribution and zero skewness and kurtosis (and joint test of normality) of the differences shown in Figure 1 and 2, all reject the null hypothesis of equal distributions as well as reject the null hypotheses of zero

skewness and zero kurtosis. A test of the null hypothesis that the median of the differences is zero cannot reject the null hypothesis when the sample is restricted to interviews taking place in November-December.

A part of the measurement error is expected to be due to rounding errors since respondents are likely to report their monthly, and in particular annual, earnings in round numbers, while the administrative records report the exact numbers. We find that 96 percent of the respondents in the survey report their monthly earnings in multiples of DKK 100, while in the administrative records only 0.9 percent receives earnings of modulus 100.⁵ Earnings in multiples of DKK 1,000 or DKK 10,000 are naturally smaller, albeit still substantive (79 percent and 22 percent, respectively). Pischke (1995) makes a similar exercise and finds comparable results. Following Pischke, we round the administrative income to nearest DKK 100 (and DKK 1,000 and 10,000), but only find exact earnings match (measurement error=0) in 3-4 percent of the cases (when rounding to DKK 10,000) while Pischke, rounding to nearest USD 1,000, finds an exact match in about 20 percent of his single-firm sample.

An OLS-regression of errors cleaned for rounding errors on the true (total) error results in a parameter estimate slightly but significantly below 1 and an R-squared value very close to 1. This means that although rounding is massive and statistically important, it does not account for a large proportion of the error, and other (systematic) components in the measurement error must exist.

Nevertheless, rounding is likely to have a much more severe impact on the distribution of wage *changes* since it will strongly increase the percentage of nominal wage rigidity.⁶ Biscourp et al. (2004) investigate this issue and find that rounding has an essential influence on wage changes.

Previous studies have generally found that errors are non-classic. For instance, Bound and Krueger (1991) report a significant AR(1) component in the measurement error in panel data, i.e. they find that individuals tend to make the

⁵ In the 1990s, the USD/DKK exchange rate fluctuated around DKK 7 to 8 for one USD.

⁶ While individual measurement errors cancel out if they are serially correlated, rounding errors do not. For an in-depth study of the potential impact of rounding errors, with an application to Finnish data, see Hanisch and Rendtel (2002).

same kind of error (under- or over-reporting) in consecutive surveys over time. Column (6) in Table 2 (overleaf) shows a similar pattern in our data with subsequent measurement errors significantly correlated; correlation coefficients between 0.3-0.4, and rather constant over the 1994-2001 period. This is very much in line with the correlation of 0.40 found by Bound and Krueger.

Columns (2)-(4) of Table 2 show the variance of measurement errors, variance of true earnings and variance of reported (survey) earnings. The share of the variance in the measurement errors to the variance in the true earnings is about 0.5 and constant over the 1994-2001 period, which spans almost an entire business cycle. Likewise, the reliability ratio, $\frac{Var(adm)}{Var(survey)}$, that plays a prominent role in regression analysis in the classical measurement error literature (Bound et al. 2001), is stable over the business cycle - but stable at a very high level around 1.9 (column 5). This indicates that standard regression may result in severely biased parameter estimates.⁷

However, the ratio $\frac{Var(adm)}{Var(survey)}$ can only be considered as the reliability ratio if the administrative earnings records are without errors. Although they are likely to be of a very high quality, this need not be the case. We discuss this issue further in Section 3.3. We also note that when 3 percent of the observations with the most extreme difference in earnings between survey and administrative records are excluded from the sample, the reliability ratio, as defined here, decreases from a value around 2 to a value around 1.3 – in some cases even lower.

Apart from errors being correlated over time, previous studies generally also find mean-reversion in measurement errors. By mean-reversion we mean that individuals with true high earnings under-report their level of earnings and individuals with true low earnings over-report their level of earnings. This study is no exception. Column (6) in Table 2 shows a negative and significant correlation between measurement errors and true earnings.

⁷ Unbiased estimates would appear if the reliability ratio was 1. The further from 1 the larger the bias. Notice that a reliability ratio > 1 only can occur when the covariance between true value and measurement error is non-negligible (as opposed to the classic assumption) and negative.

Table 2 Basic facts about measurement errors (log monthly earnings)

	obs	var(m)	var(adm)	var(survey)	reliability ratio	corr(m, adm)	corr(m(t), m(t-1))	b(m,adm)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel 1994-2001	16,748	0.123	0.249	0.134	1.858	-0.360	0.372	-0.478 (0.009)
1994	2,621	0.123	0.257	0.137	1.876	-0.362	na	-0.474 (0.023)
1995	2,320	0.120	0.236	0.122	1.934	-0.347	0.427	-0.495 (0.025)
1996	2,169	0.121	0.236	0.123	1.919	-0.367	0.356	-0.497 (0.028)
1997	2,104	0.106	0.215	0.115	1.870	-0.354	0.297	-0.478 (0.024)
1998	2,006	0.127	0.232	0.127	1.827	-0.442	0.306	-0.498 (0.024)
1999	1,947	0.125	0.250	0.121	2.066	-0.419	0.387	-0.508 (0.025)
2000	1,778	0.131	0.262	0.128	2.047	-0.385	0.416	-0.507 (0.025)
2001	1,793	0.131	0.250	0.121	2.066	-0.380	0.422	-0.520 (0.027)
Panel 1994-2001, full time work	10,153	0.050	0.143	0.111	1.288	-0.362	0.372	-0.288 (0.014)
Panel 1994-2001, Nov.-Dec. interview	1,168	0.156	0.291	0.132	2.205	-0.372	0.403	-0.542 (0.029)
Panel 1994-2001, adm. wage > median	8,369	0.016	0.065	0.075	0.867	-0.082	0.423	-0.047 (0.006)
Panel 1994-2001, adm. wage < median	8,369	0.207	0.199	0.080	2.488	-0.456	0.397	-0.820 (0.011)
Panel 1994-2000, annual wage, retrospect	8,263	0.026	0.118	0.116	1.017	-0.207	0.358	-0.117 (0.008)
<i>By sector</i>								
Panel 1994-2001, manufacturing	4,258	0.138	0.249	0.122	2.041	-0.379	0.342	-0.532 (0.018)
Panel 1994-2001, service	11,431	0.113	0.246	0.138	1.783	-0.348	0.397	-0.449 (0.011)

Note: m=measurement error. Columns (3) and (4) report variance in logarithmic earnings. Column (8) reports heteroscedasticity-consistent standard errors in parentheses. All correlations are significantly different from zero at the 5% significance level (Spearman test).

Another test of mean-reversion is to regress true earnings on the measurement error. The OLS parameter estimate from such a regression is reported in column (8) and the results show the same negative relationship between measurement errors and true earnings, and hence corroborate the findings in column (6). Compared to previous studies, e.g. Pischke (1995), the mean-reversion found here is strong and persistent over the business cycle.

The overall sample has been split into several sub-samples: full-time workers; respondents who were interviewed in November-December (close to the administrative record update); above or below the median income; and respondents who have replied to the question about annual, retrospective, income. The general impression from comparing sub-samples with the overall sample is that there are very important differences between subgroups. High-income earners have a much lower mean-reversion in reported earnings. Similarly, the reliability ratio is closer to one for full-time workers and high-income earners than for the entire sample. The same is the case for annual, retrospective income. Hence, using this latter variable or limiting the sample to full-time workers will give regression results with less biased parameter estimates.

Analyses by the manufacturing and service sectors reveal that these sectors are very much alike and, as the basic facts shown for 1994-2001 are stable over the business cycle (not shown), there does not appear to be any difference in business cycle sensitivity between sectors. Pischke (1995) argues that, as the economy as a whole is less cyclical than the manufacturing sector which he analyzes, we should find measurement errors to be more volatile over the business cycle in the manufacturing sector than in the service sector. We find, in line with Bound and Krueger (1991), that the measurement error is roughly constant over the business cycle; in the economy at large as well as in both the manufacturing and service sectors.

Table 3 Results from OLS-regressions of measurement errors

	1995			1998			2001	
	Coef.	Std. Err.		Coef.	Std. Err.		Coef.	Std. Err.
age	-0.028	0.013	**	0.019	0.015		0.015	0.022
agesq	0.000	0.000		0.000	0.000		0.000	0.000
Top socio-group	0.059	0.058		0.067	0.057		0.241	0.082 ***
<i>Education (ref=advanced, 5 years)</i>								
primary I	0.104	0.155		-0.112	0.154		-0.087	0.296
primary II	0.032	0.042		0.034	0.048		0.238	0.068 ***
secondary	-0.028	0.086		-0.012	0.087		-0.037	0.127
gymnasium	-0.151	0.109		-0.065	0.108		-0.089	0.152
vocational education	-0.017	0.082		0.012	0.080		0.043	0.114
short advanced (1-2 years)	0.004	0.106		-0.025	0.101		0.049	0.146
middle advanced (3-4 years)	0.055	0.080		0.113	0.077		0.136	0.107
Female	-0.270	0.040	***	-0.205	0.042	***	-0.178	0.060 ***
<i>Earnings quantile (ref=highest)</i>								
lowest	0.764	0.065	***	0.865	0.065	***	1.421	0.094 ***
second	0.092	0.060		0.170	0.060	***	0.240	0.089 ***
third	0.049	0.058		0.048	0.053		0.130	0.073 *
<i>Number of waves observed (1=ref)</i>								
observed in 2 waves	-0.080	0.108		-0.266	0.158	*	-0.225	0.166
observed in 3 waves	-0.187	0.109	*	-0.408	0.145	***	-0.365	0.156 **
observed in 4 waves	-0.150	0.111		-0.400	0.138	***	-0.388	0.154 **
observed in 5 waves	-0.154	0.114		-0.429	0.137	***	-0.361	0.159 **
observed in 6 waves	-0.091	0.110		-0.436	0.135	***	-0.305	0.151 **
observed in 7 waves	-0.114	0.108		-0.445	0.137	***	-0.313	0.147 **
observed in 8 waves	-0.130	0.102		-0.413	0.136	***	-0.322	0.144 **
constant	0.916	0.297	***	0.211	0.320		-0.014	0.473
Regional dummies	yes			yes			yes	
Number of observations	2,320			2,006			1,793	

Note: The dependent variable is the $\text{abs}\{\ln(\text{gross earnings survey}) - \ln(\text{gross earnings administrative records})\} / \ln(\text{gross earnings administrative records})$.

Next, by using simple OLS-regressions of observables on measurement errors in monthly earnings using cross-section data, we investigate whether any observable variables can explain measurement errors in monthly earnings. The dependent variable is the absolute error (or absolute deviance from the administrative records) in percent. To the extent that females have a higher propensity to be part-time workers, we should expect to find more misreporting among females (if not controlling for hours). However, OLS regressions reveal that females make significantly *fewer* errors than men, see Table 3. This finding confirms results reported by Bound and Krueger (1991) and Bollinger (1998).

Epland and Kirkeberg (2002) find the opposite result for Norway, i.e. that females mis-report more than men.

Low-income earners (the two lowest quartiles) are found to misreport significantly more than high-income earners. This may to some extent mirror the finding by Nordberg et al. (2001) that low-educated individuals make more errors than highly educated; although parameter values for the education indicator variables are generally insignificant in this study.

The OLS-regressions confirm the impression from Figure 2 that individuals who attrite from the panel, i.e. individuals who are observed in a cross section or in a few waves only, report earnings with a larger error than individuals who appear more often or who do not attrite at all, even when we control for age and other covariates. This may have important implications for analyses using unbalanced panels. While most analyses based on panel data allow for an unbalanced panel, to the best of our knowledge they never control for asymmetric differences in measurement error with a systematically larger error among individuals who are only present in some years of the panel.

3.2 Estimation of the Conditional Means Function

In this section, we take the analysis a step further by estimating the relationship between the survey earnings and administrative records of earnings in a nonparametric fashion.

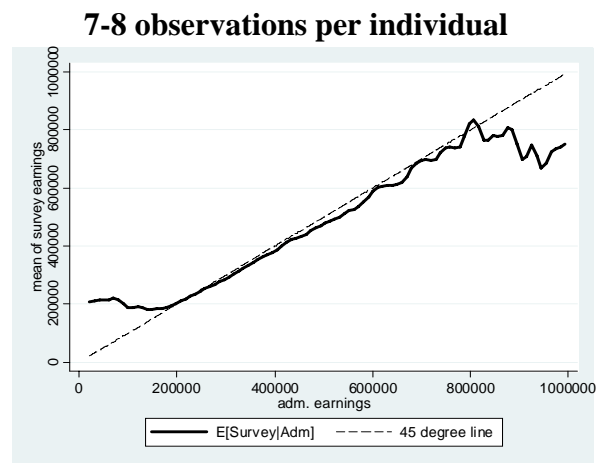
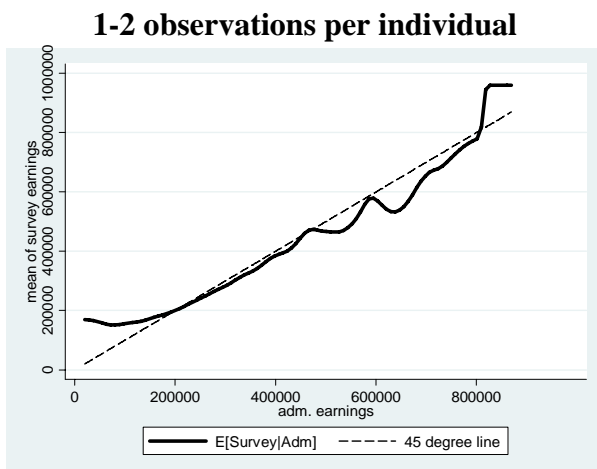
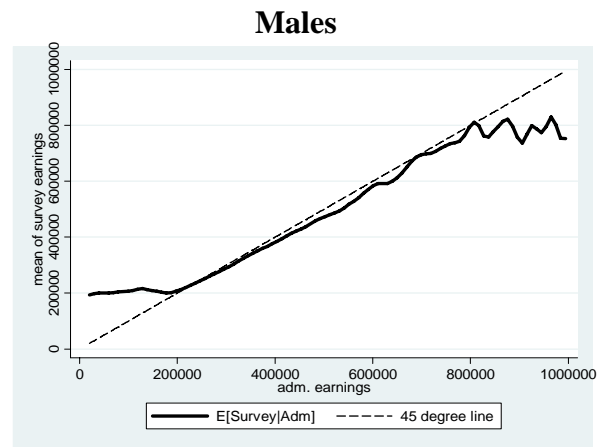
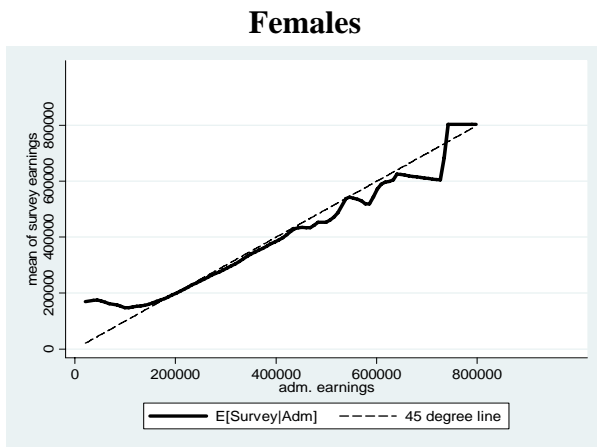
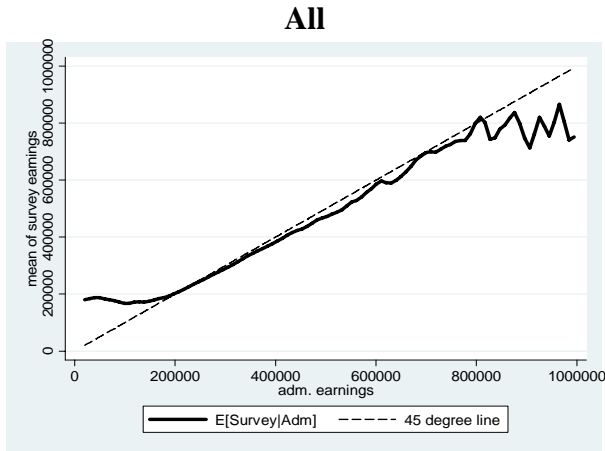
The nonparametric approach is well suited to problems where the data set is large and where focus is on the true relationship between the two sources of earnings since this method is able to capture non-linearities that would go undetected in an OLS-regression or some other linear parametric regression method.

The nonparametric kernel regressions presented in Figure 3 are based on the Nadayadara-Watson estimator (see Bollinger, 1998, for further details). The graphs present the estimation of $E[\text{survey earnings} \mid \text{adm. earnings}]$. In each graph, the solid line represents the estimate of the conditional expectation of the survey earnings response given the level of earnings in the administrative records.

In the absence of measurement errors, the solid line would coincide with the dashed 45° line, which would imply that the conditional expectation of survey earnings would equal the administrative earnings. As previously mentioned, formal tests indicate that the distributions are not equal and hence we should not expect the solid line to follow the 45° line. Indeed, irrespective of which sub-sample we condition on the figures show a tendency of regression towards the mean. This is the case since the conditional expectation generally is found above the 45° line for low levels of earnings and below the 45° line for high levels of earnings.

The vast proportion of individuals (93 percent) have annual earnings in the interval [100,000; 600,000], and only 1 percent of the observations is found above this interval. Thus, in the relevant earnings range the income levels are almost identical, except for individuals who are observed only once or twice. Generally, the figures suggest that a linear model specification may be appropriate after all.

Figure 3 Nonparametric regression of survey earnings conditional on administrative record earnings



3.3 Discussion of Quality of Administrative Earnings Data

So far we have treated the administrative earnings and income records as error-free validation data. Although we have made every effort to optimize the quality of this information, it will inevitably also be contaminated with its own errors. The register data are considered to have a high quality because they originate from an administrative database, which is updated and revised on a continuous basis by various administrative agencies.

However, the administrative earnings data rely on employers' reports to tax authorities and this information may potentially be underreported. The tax registration system is organized so that an employer only deducts labor costs if he/she declares the tax. Although this does not prevent tax fraud altogether, it does provide a clear economic incentive for the employer to declare all wage costs. Furthermore, the measurement errors are virtually symmetric around zero and the analysis so far does not indicate any systematic bias. Hence, underreporting in the validation data appears not to be a (serious) problem.

The match between Danish administrative records and survey data on earnings found here appears much better than the corresponding 1996 cross-section match for Finland, cf. Nordberg et al. (2001), who find that net wages from survey data tend to be systematically lower than administrative records. In a Norwegian study, Epland and Kirkeberg (2002) match the 1997 Survey of Living Conditions with administrative records and find that two thirds of the respondents underreport their previous annual gross income. A comprehensive analysis of Danish 1994 cross-section ECHP data shows that gross wages are substantially overreported, Jørgensen (1998).

Why do we find a much better match in income and earnings? An important difference is that we limit our sample to employees only, i.e. excluding self-employed, students, retirees and others out of the labor force. Self-employed, for instance, have much more complex accounts. Furthermore, we limit the sample

to 18-67-year-olds while other studies have a slightly broader age group, e.g. including 16-17-year-olds.

4 Measurement Errors in Categorical Variables

Next we turn to measurement errors in categorical variables, which are usually termed classification errors. With a long list of covariates available, we are able to compare a large number of categorical variables. However, we limit our analysis to some of the most important covariates that are found in a large number of econometric studies: education, industry and firm size.

We are not concerned with whether or not the classification error is classic or not since, by definition, it cannot be classic (Aigner, 1973). Categorical variables are bounded from below and above, which inevitably makes any classification error vary with the true level of the categorical variable, i.e. there will be mean-reversion in these variables if measured with error.

As was the case for the validation earnings data, we have made every possible effort to ensure correctness in the validation data. Nevertheless, errors in administrative records may exist and we discuss these potential errors for each variable in turn.

4.1 Sector and Industry

Classification errors into type of sector or industry are fairly low (see Table 4 for classification into primary, secondary and tertiary sectors and Appendix C for a more detailed classification with 15 industries). The more aggregated the group the lower classification error. The percent of correct classifications is given (in bold) along the diagonal. Among respondents in the tertiary (service) sector, less than 5 percent make a classification error, and since employment in this sector accounts for more than 70 percent of all workers in the sample, it means that, confining the analysis to the 3-sector classification, 94 percent of the respondents make a correct classification.

Table 4 Sector categorization

Admin. Record	Survey			Total	Number of obs
	Primary	Secondary	Tertiary		
Primary	80.7%	8.8%	10.5%	100.0%	171
Secondary	0.3%	90.5%	9.2%	100.0%	4,221
Tertiary	0.4%	4.2%	95.5%	100.0%	11,323
Number of obs	195	4,306	11,214	100.0%	15,715

Note: Absolute numbers and column percentages.

The more detailed grouping (Appendix C) shows more discrepancy and the percentage of correct reports drops to 68 percent. One of the major industries, health and social work, with about 2,700 observations in the administrative records, only has about 46 percent correct classifications. An almost equally large proportion (44 percent) of respondents working in the health and social work sector classify themselves as 'public administration employees'. While this type of mistake is understandable, it may have a very large impact on analyses based on survey records compared to administrative records.

Still, the magnitude of the classification errors in sector/industry is moderate when compared to existing studies. Mellow and Sider (1983) report correct classification in about 84-92 percent of their observations and other studies are in the same range, cf. Bound et al. (2001).

Some of the errors may be due to failure in coding the survey responses. Furthermore, it is conceivable that firms change their primary industry over time and that the administrative records are adjusted only partly and slowly. Hence, it may be the case that the survey data are more accurate than the administrative data because the latter source under-estimates the occurrence of firms changing industrial classification over time. On the other hand, mis-reporting may also exaggerate the occurrence of changes in industry when estimates of such changes are based on comparing reports of industry obtained at two different points in time.

4.2 Firm Size

The validity of firm (plant) size categorization is not impressive, cf. Table 5. While 70-80 percent of the respondents in very large or very small firms categorize the size of their company correctly, the shares are much lower for intermediate firm sizes. This difference is expected as it reflects mean-reversion. But the relatively low level of correct categorization appears somewhat surprising. Overall, a correct match is found in about 70 percent of the observations. Some consolation is found in the fact that most classification errors are made to the immediately adjacent category.

Existing studies on validation of reports of establishment and firm size are sparse. Brown and Medoff (1996) report the correlation between employee and employer records on firm size to be 0.86. In comparison, we find a correlation in firm size between survey and administrative records of 0.76.

How exact do the administrative records capture firm size? Since the records include a personal identification number for every individual aged 16-74, we are able to sum across these numbers and hence, in principle, obtain highly accurate information on firm size. However, the administrative records only provide a snapshot of employees at the end of November. As most of the errors are to the immediately adjacent category, it is difficult to ascertain the exact amount of misreporting but it is likely to be lower than the 30 percent we find here. Another reason for this is that some workers captured in the administrative records may work very few hours and/or be loosely connected to the firm. However, if this was generally the case, we should expect to find a relatively larger share in the lower triangle of Table 5 but the reverse is actually the case (18 percent of respondents are in the upper diagonal and 12 percent in the lower diagonal).

Lastly, although it was made clear to the enumerators that they should inquire about plant size and not overall firm size (in case of several plants within the same firm), this is still a potential source of error. The fact that 18 percent are found in the upper diagonal (vs. 12 percent in the lower diagonal) suggests that this potential confusion is real and present in the data.

4.3 Education

The categorization of education changed code in 1997 and we therefore limit this part of the analysis to the 1998-2001 period, in order to make a good match between completed education in the two data sources. Furthermore, the survey data include highest education attended, while the validation data have information on on-going education as well as completed education. By restricting the sample to individuals who in the start year of the register data, 1984, were 18 years or younger, we can follow their on-going education and in this manner construct highest education attended. These restrictions limit the sample to 2,053 observations on highest education attended.⁸

The extent of classification errors in educational attainment appears pronounced, cf. Table 6. Correctly classified individuals are again found along the diagonal (in bold), which only includes 71 percent of the total number of observations. In particular there seems to be a lot of confusion about whether the respondent's highest education attended is vocational training/education or a short or medium length further education.

The question about education is formulated in the following manner

*To date what is the highest education you have received,
when disregarding vocational courses and further
education after vocational courses.*⁹

It seems likely to conjecture that some respondents interpret this question as referring to highest education completed, not attended. When we make the best possible match between the survey data on education attended and the administrative records on attended or completed - whichever fits best for each individual in question - the overall fit increases from 71 percent to 82 percent, cf. Table 7.

⁸ The register data actually start in 1980 but education codes were not fully captured before 1984. After 1984, all educations are included in the register.

⁹ The underlining is included in the questionnaire. Follow-up questions are also posed. See Appendix A for details.

Table 5 Firm size

	Survey						Total	Number of obs
	1-4 employees	5-19 employees	20-49 employees	50-99 employees	100-499 employees	500+ employees		
Administrative records								
1-4 employees	69.9%	17.3%	5.4%	0.6%	5.4%	1.4%	100.0%	700
5-19 employees	11.1%	71.1%	8.2%	3.6%	2.9%	3.2%	100.0%	2,712
20-49 employees	1.8%	23.5%	55.9%	9.3%	5.2%	4.2%	100.0%	2,062
50-99 employees	1.6%	5.8%	23.6%	46.1%	16.6%	6.3%	100.0%	1,803
100-499 employees	1.3%	4.4%	6.5%	11.1%	62.0%	14.7%	100.0%	2,798
500+ employees	0.9%	4.8%	4.6%	3.3%	9.9%	76.5%	100.0%	1,757
Number of obs	906	2,845	2,103	1,493	2,433	2,052	100.00%	11,832

Table 6 Highest education attended

Administrative records	Survey						Total	Number of obs
	Primary education	Secondary education	Vocational training	Short further education	Medium further education	Long further education		
Primary education	55.5%	4.3%	35.4%	2.7%	1.5%	0.5%	100.0%	1,361
Secondary education	1.8%	43.5%	27.5%	7.5%	9.0%	10.8%	100.0%	400
Vocational training	1.6%	0.7%	90.4%	5.3%	1.9%	0.1%	100.0%	2,975
Short further education	1.5%	2.3%	53.0%	31.1%	9.2%	2.9%	100.0%	479
Medium further education	0.1%	0.6%	26.2%	10.0%	58.4%	4.7%	100.0%	1,604
Long further education	0.3%	0.2%	2.0%	1.1%	3.8%	92.6%	100.0%	651
Number of obs	821	273	3,969	542	1,120	745	100.0%	7,470

Table 7 Highest education attended or completed (best possible match per individual)

Administrative records	Survey						Total	Number of obs
	Primary education	Secondary education	Vocational training	Short further education	Medium further education	Long further education		
Primary education	91.6%	1.4%	7.0%	0.0%	0.0%	0.0%	100.0%	328
Secondary education	0.0%	80.4%	16.4%	1.6%	0.5%	1.1%	100.0%	239
Vocational training	0.9%	2.1%	90.5%	6.2%	0.3%	0.0%	100.0%	1,370
Short further education	0.9%	2.8%	28.7%	61.1%	5.6%	0.9%	100.0%	164
Medium further education	0.6%	0.6%	26.1%	6.7%	62.2%	3.9%	100.0%	557
Long further education	0.9%	1.3%	4.5%	4.5%	11.2%	77.7%	100.0%	342
Number of obs	210	183	1,048	162	259	191	100.0%	2,053

The large discrepancies between the two sources with respect to vocational training may to some extent be classification differences, not errors, see Appendix A for details. A key problem is that some training courses (vocational, theoretical or mixed) may have a relatively short duration and do not increase the level of formal education as measured by the rather crude classification system applied in the survey. At the same time, some (vocational) training courses may raise the formal education level from "vocational" to "further education" and the limit here is not clearly defined.¹⁰ This may also explain why the percentage of respondents with "long further education" is lower in the diagonal in Table 7 compared to Table 6.

4.4 Correlations in Measurement Errors between Different Variables

From the point of view of an applied econometrician, it is of interest to investigate whether reporting errors tend to be highly correlated between variables, i.e. does misreporting on, say, earnings imply an increased risk that the same respondent will make classification errors in other variables?

The answer is not clear. Naturally, errors in sector and industry are, by definition, closely related and have a relatively high correlation of around 0.4. Some of the other correlations are also significantly different from zero but most are completely insignificant and very small - a few correlations are even negative, see Table 8. A measurement error in earnings of 5 percent or more is here classified as an error. Altering the definition to a 10 percent error in earnings has only a negligible impact on the result.

¹⁰ That is, in the administrative records it is very well defined and education is divided into about 500 different types of education. But, respondents in a survey are not likely to be able to answer such questions in great detail.

Table 8 Correlation between measurement errors

	sector (3 groups)	industry (15 groups)	firm size	earnings (error>5%)	education
sector (3 groups)	1.000				
industry (15 groups)	0.392 (0.000)	1.000			
firm size	0.092 (0.000)	0.081 (0.000)	1.000		
earnings (error>5%)	0.051 (0.010)	0.054 (0.006)	0.024 (0.228)	1.000	
education	0.041 (0.035)	-0.004 (0.826)	0.007 (0.726)	-0.025 (0.196)	1.000

Note: Based on 2,587 observations. P-values in parentheses.

We also included an indicator variable for errors in educational attainment in the OLS-regression analysis of measurement errors in earnings reported in Table 3. The results (not shown) reveal that misclassified education does not significantly affect the likeliness of measurement errors in earnings.

Looking at the number of errors per respondent within the same cross section (see Table 9) shows that most individuals make one or two errors but only about 6 percent misreport more than three times out of the five possible.

Table 9 Number of errors by the same respondent

Number of errors	Frequency	Percent	Cumulative percentage
0	297	11.48	11.48
1	877	33.90	45.38
2	822	31.77	77.16
3	435	16.81	93.97
4	141	5.45	99.42
5	15	0.58	100.00
Total	2,587	100.00	100.00

The overall picture is that errors do not seem to be prolific within individual cross-section responses. The immediate implication of this finding is that fixed

effects estimates are not a relevant tool for mitigating problems with classification errors.¹¹

5 Conclusion

This paper investigates the extent of measurement and classification errors in a large sample of matched ECHP survey data and administrative records. We show that despite a careful and well-organized survey set-up carried out by a highly professional agency, the measurement errors in earnings found here are generally more substantive than found in previous studies. Furthermore, classification errors with respect to firm size, industry and education group are of about the same magnitude as previously found in the literature.

The main reason why we find more pronounced errors probably is that many of the important contributions to the literature on measurement errors (e.g. Duncan and Hill, 1985, Duncan and Mathiowetz, 1985, Bound et al., 1994) are based on analyses of one single firm. Incidentally, the employees in this firm have a relatively high level of education and long tenure. The authors also mention that the measurement errors from a large scale validation studies are likely to be larger. Our analysis confirms this point. On a more positive note, the median absolute error in earnings is not significantly different from zero when the sample is limited to the November-December interviews.

The finding that respondents who attrite from the sample report significantly worse than respondents who are present throughout all panel waves is important and rather disturbing. An immediate consequence of this finding is that it is not enough simply to allow for unbalanced samples in panel data analyses; differences in estimates based on cross-section data and panel data, respectively, may be due to differences in response errors. Hence, with this type of measurement error heteroskedasticity is not the only problem we encounter in models based on panel data hampered with attrition.

¹¹ Bound and Krueger (1991) find that measurement errors can seriously bias fixed-effects estimates. We shall not discuss this issue further here.

A key conclusion that emerges from this paper is that the possibility of non-classical measurement errors should be taken much more seriously by those who analyze survey data. Implicitly working under the assumption of classical measurement errors is not likely to have much basis in reality, and in order to base ones econometric analysis on firm ground it is advisable to be explicit about the error structure in the data.

Although we believe the validation data applied in this study have a high quality, they are not without errors, and this is something that has been discussed throughout this paper. There are, of course, also limits to how much the results found here can be generalized to apply to other contexts, other countries or other error-ridden data sets. Nevertheless, the mere magnitude of the sample and the fact that it represents all sectors in the economy makes this study a relevant contribution to the literature. We urge further validation studies in order to establish cross-country and/or cross-sample patterns of measurement error problems.

6 References

- Aigner, D.J. (1973) Regression with a Binary Independent Variable Subject to Errors of Observation, *Journal of Econometrics*, **1**, 49-60.
- Battistin, E. and Sianesi, B. (2005), Misreported Schooling and Returns to Education: Evidence from the UK, Draft Manuscript, Institute for Fiscal Studies, UK.
- Biscourp, P., Dessy, O. and Fourcade, N. (2004), Downward wage rigidity: a micro-level empirical analysis for France in the `90s, Manuscript, *CREST-INSEE*, France.
- Bollinger, C. R. (1998), Measurement Error in the Current Population Survey: A Nonparametric Look, *Journal of Labor Economics*, **16** (3), 576-594.
- Bound, J., Brown, C. and Mathiowetz, N. (2001), Measurement error in survey data, Ch. 59 in Heckman, J.J. and Leamer, E. (eds.) *Handbook of Econometrics*, **5**, 3705-3843.

- Bound, J., Brown, C. Duncan, G. J. and Rodgers, W. L. (1994), Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data, *Journal of Labor Economics*, **12** (3), 345-368.
- Bound, J. and Krueger, A.B. (1991), The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?, *Journal of Labor Economics*, **9** (1), 1-24.
- Brown, C. and Medoff, J.L. (1996), Employer Characteristics and Work Environment, *Annales D'Économie et de Statistique*, 41/41.
- Duncan, G. J. and Hill, D.H. (1985), An Investigation of the Extent and Consequences of Measurement Error in Labor-economic Survey Data, *Journal of Labor Economics*, **3** (4), 508-532.
- Duncan, G.J. and Mathiowetz, N.A. (1985), A Validation Study of Economic Survey Data, *Ann Arbor, MI: Institute for Social Research*.
- Epland, J. and Kirkeberg, M.I. (2002), Comparing Norwegian income data in administrative registers with income data in the Survey of Living Conditions, Paper prepared for The International Conference on Improving Surveys (ICIS), Copenhagen 2002.
- Hanisch, J. and Rendtel, U. (2002), Quality of Income Data from Panel Surveys with Respect to Rounding, *CHINTEX WP 6*.
- Jäckle, A. Jenkins, S. P., Lynn, P. and Sala, E. (2004), Validating Survey Data: Experiences Using Employer Records and Governmental Benefit Data in the UK, Unpublished Manuscript, *University of Essex*.
- Jørgensen, J.M. (1998), Comparison of Income Variables between the ECHP and Danish Statistical Registers, Manuscript, *Statistics Denmark*.
- Mellow, W. and Sider, H. (1983), Accuracy in Response in Labor Market Surveys: Evidence and Implications, *Journal of Labor Economics*, **1** (4), 331-344.
- Nordberg, L. Penttilä, I. and Sandström, S. (2001), A study on the effects of using interview versus register data in income distribution analysis with an application to the Finnish ECHP-Survey in 1996, *CHINTEX WP 5*, Statistics Finland/Working Paper 1.

Pischke, J.-S. (1995), Measurement Error and Earnings Dynamics: Some Estimates From the PSID Validation Study, *Journal of Business & Economic Statistics*, **13** (3), 305-314.

Appendix A: Description of Key Variables

Monthly earnings:

In the survey, the question about monthly earnings is phrased as follows:

What is your normal gross income from your main occupation? (gross income per month, i.e. before tax, pension and other deductions?)

The corresponding variable in the administrative records gives the gross income for the main occupation in November each year.

Annual earnings:

In the survey, questions about annual earnings appear after an introduction where the interviewer states: The following questions are concerned with various types of income last year. The phrasing of the question about annual earnings reads:

How large do you think your total gross earnings were last year?

This question is followed up by the following questions:

*Did you receive any additional payment for overwork, gratuity, tip or other? [yes/no]
Shall these additional payments be added to the annual income you mentioned or are they already included? [are included/are not included].*

Education:

The codes for completed education are very detailed in the administrative data with up to 500 sub-groups of education code, which are re-coded into 6 sub-groups. In the survey the highest completed education is divided into several questions:

1. To date what is the highest education you have received, when discarding vocational courses and further education after vocational courses?

The enumerator is here explicitly asked to code people who have attended a certain level as the right choice - irrespective of whether the person has completed that education and hence received an exam. Clearly, this will in itself account for an important part of the upward bias in educational attainment in the survey data. This means that all classifications in the upper triangular may be correct.

2. Have you completed any form of vocational education? [yes/no]

If yes:

3. Which education?

- a) Further education after ending vocational course
- b) Vocational education with duration of one year or more, exclusively taking place in a school

- c) Vocational education with duration of one year or more, taking place in a school and at a workplace
- d) Vocational education lasting less than a year

It is not clear how exactly to define the "vocational training" group (used in Table 5) from answers to these questions. Respondents in group a) may belong to "short further education" or they may belong to "Medium further education" - both may be right. Answers falling into group b) and c) are coded as "vocational training" while group d) is not considered to be enough to lift the respondents highest (completed) education up. The education level for a respondent is only re-coded if the education level is below the level achieved by taking further vocational education. This manner of re-coding is the most exact we can make but as it is not always clearly defined, it may cause discrepancies between the survey and administrative records that are in fact inflicted by our re-coding.

Firm size:

The question reads:

How many permanent employees are there at your work place (the company where you work)?¹²

The answer is pre-coded into the groups shown in Table 5.

It is spelled out for the enumerators that they should ask in to the plant size, not the firm size. In the administrative records, there are identifiers for both overall firm and plant and hence the numbers shown in Table 5 are related to the plant size.

Sector/Industry

The question reads:

What is the main activity of the firm where you work (the branch)?

The respondent was shown a card with 23 pre-coded options and a 24th possibility "other".

¹² Parenthesis not in questionnaire. This reflects the meaning of the question.

Appendix B: Comparison of Population and Sample Means and Distributions

	1994-2001		1994		2001	
	Sample	Population	Sample	Population	Sample	Population
<i>Age</i>	40.94 (10.42)	37.48 (10.59)	40.16 (10.62)	39.49 (10.87)	42.06 (10.44)	35.67 (10.05)
<i>Gender</i>						
Men	9,006 53.77%	2,470,002 47.39%	1,417 54.06%	313,840 48.70%	945 52.70%	298,889 45.80%
Women	7,742 46.23%	2,742,049 52.61%	1,204 45.94%	330,562 51.30%	848 47.30%	353,778 54.20%
<i>Region</i>						
Sjaelland and Islands	7,457 44.52%	2,599,055 49.87%	1,130 43.11%	324,304 50.33%	46.07 46.07%	325,387 49.85%
Fyn island	1,514 9.04%	431,935 8.29%	253 9.65%	53,264 8.27%	155 8.64%	53,851 8.25%
Jylland	7,777 46.44%	2,181,061 41.85%	1,238 47.23%	266,834 41.41%	812 45.29%	273,429 41.89%
<i>Income</i>	244,313 (111,849)	240,562 (120,859)	214,533 (99,584)	211,738 (106,120)	281,129 (124,456)	274,927 (137,067)
<i>Education</i>						
	1998-2001		1998		2001	
Primary education	1,361 18.22%	557,255 21.45%	399 20.05%	148,528 22.72%	293 16.44%	130,120 20.14%
Secondary education	400 5.35%	145,958 5.62%	107 5.38%	40,237 6.15%	99 5.56%	32,872 5.09%
Vocational training	2,975 39.82%	1,025,534 39.48%	799 40.15%	257,719 39.42%	709 39.79%	255,395 39.53%
Short further education	479 6.41%	136,716 5.26%	128 6.43%	33,770 5.17%	116 6.51%	34,515 5.34%
Medium further education	1,605 21.48%	501,839 19.32%	397 19.95%	119,828 18.33%	405 22.73%	131,477 20.35%
Long further education	651 8.71%	230,531 8.87%	160 8.04%	53,684 8.21%	160 8.98%	61,698 9.55%

Note: Standard deviation in parentheses.

Appendix C: Detailed Sector Categorization

Administrative record	Survey							Total	Number of obs
	Agriculture	Mining and quarrying	Manufacturing	Electricity, gas and water supply	Construction	Wholesale and retail trade 2)	Hotels and restaurants		
Agriculture ¹⁾	80.7%	0.0%	4.1%	0.0%	4.7%	0.0%	0.6%		
Mining and quarrying	0.0%	83.7%	2.0%	10.2%	2.0%	0.0%	2.0%		
Manufacturing	0.4%	0.2%	85.5%	0.3%	3.2%	2.6%	0.1%		
Electricity, gas and water supply	0.0%	0.0%	1.7%	91.4%	0.0%	0.0%	0.0%		
Construction	0.0%	0.1%	6.9%	8.8%	78.3%	0.2%	0.0%		
Wholesale and retail trade	0.9%	0.0%	10.3%	0.7%	1.1%	67.1%	0.8%		
Hotels and restaurants	0.0%	0.0%	1.3%	0.0%	0.6%	0.6%	72.4%		
Transport ²⁾	0.1%	0.0%	1.6%	0.1%	1.1%	2.1%	0.5%		
Financial intermediation	0.0%	0.0%	0.1%	0.4%	0.6%	1.6%	0.0%		
Real estate ³⁾	0.8%	0.0%	0.0%	0.0%	3.1%	7.7%	0.0%		
Public adm. ⁴⁾	0.7%	0.0%	2.9%	0.2%	4.5%	1.7%	0.3%		
Education	0.3%	0.3%	0.2%	0.0%	0.0%	0.2%	0.0%		
Health and social work	0.0%	0.0%	0.1%	0.0%	0.0%	0.1%	0.1%		
Other community ⁵⁾	0.4%	0.0%	1.6%	0.6%	1.6%	3.0%	0.0%		
Other	0.0%	0.0%	0.0%	0.0%	0.0%	1.7%	0.0%		
Number of obs	195	53	3,093	224	936	1,412	148		

	Transport	Financial intermediation	Real estate	Public adm.	Education	Health and social work	Other community	Other	Total	Number of obs
Agriculture ¹⁾	1.2%	0.0%	0.0%	1.2%	2.3%	0.0%	2.9%	2.3%	100.0%	171
Mining and quarrying	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	49
Manufacturing	0.2%	0.0%	0.0%	0.2%	0.1%	0.1%	1.5%	5.4%	100.0%	3,197
Electricity, gas and water supply	0.0%	0.0%	0.0%	6.0%	0.9%	0.0%	0.0%	0.0%	100.0%	116
Construction	2.4%	0.0%	0.0%	0.2%	0.0%	0.0%	0.3%	2.6%	100.0%	859
Wholesale and retail trade ²⁾	1.4%	0.4%	0.0%	0.3%	0.3%	0.8%	6.3%	9.7%	100.0%	1,800
Hotels and restaurants	2.6%	0.0%	0.0%	2.6%	11.5%	0.0%	0.6%	7.7%	100.0%	156
Transport ³⁾	51.3%	0.2%	0.0%	5.5%	0.0%	0.0%	3.2%	34.4%	100.0%	1,021
Financial intermediation	0.0%	93.4%	0.0%	1.7%	0.0%	0.1%	0.7%	1.3%	100.0%	698
Real estate ⁴⁾	0.8%	2.3%	18.5%	0.8%	0.0%	0.0%	1.5%	64.6%	100.0%	130
Public adm. ⁵⁾	0.9%	0.7%	0.0%	60.2%	1.4%	1.1%	3.8%	21.6%	100.0%	2,153
Education	0.1%	0.0%	0.0%	5.6%	85.1%	0.6%	6.1%	1.8%	100.0%	1,597
Health and social work	0.1%	0.0%	0.0%	43.5%	2.7%	46.2%	1.3%	5.6%	100.0%	2,695
Other community ⁵⁾	0.4%	1.5%	0.0%	8.8%	2.1%	1.5%	56.4%	22.3%	100.0%	1,014
Other	1.7%	0.0%	0.0%	6.8%	0.0%	15.3%	0.0%	74.6%	100.0%	59
Number of obs	613	695	24	2,746	1,515	1,320	995	1,746	100.0%	15,715

1) including hunting, forestry and fishing

2) including storage and communication

3) including renting and business activities

4) including defence and compulsory social security

5) including social and personal activities, private households with employed persons, extra territorial organizations and bodies, and reserach.