Errors and Lies about Educational Attainment.*

Francesco Avvisati
Paris School of Economics

This version: January 2010

Nous gagnerions plus de nous laisser voir tels que nous sommes, que d'essayer de paraître ce que nous ne sommes pas.
François de la Rochefoucauld (1613-1680)

Abstract

Using a French employer-employee data set, I document large misclassification errors in both employer reports and self-reports of workers' educational attainment. I present a model for the data-generating process under which the precision of each report is identified. When major misclassification errors are considered only, estimates indicate that employer reports are correct for about 84% of the sample, while self-reports are accurate for almost 94% of observations.

Using the same model to test whether wages are related to employers' mistakes, I find that overstatements of worker education pay a positive wage-premium, whereas under-statements are not linked to significant differences in wages. I propose a simple interpretation of this pattern.
Keywords: Misclassification, Returns to education
JEL codes: J31, I20, C31, C51, D82, D83, D84

*First draft June 2008. This paper benefited from helpful discussions with Eric Maurin, Marc Gurgand, Milo Bianchi, Anders Stenberg, Claudio Lucifora, Maëlys de la Rupelle, Yanos Zylberberg, Thomas Breda and comments by the audiences at Paris-Jourdan Sciences Economiques (lunch seminar, June 2008), the XXIII Italian Conference of Labour Economics (September 2008), the Augustin Cournot Doctoral Days (April 2009), the COST workshop "Wages and Firms" in Budapest (May 2009), the EALE annual conference in Tallinn (September 2009).
PhD Candidate, Paris School of Economics, 48 Boulevard Jourdan, 75014 Paris. Email: francesco.avvisati@ens.fr.

Contents

Educational reports obtained for the same person from two separate sources display a surprisingly high number of discrepancies. This finding has been documented for twin data (self-reported vs. twin-reported educational attainment), recall data at different dates, self-reported vs. administrative data[1]. Despite the increasing availability of firm-level data on workforce qualification, there is little evidence about employers' knowledge of workers' degrees, and the accuracy of employers in reporting this information. Using the French Structure of Earnings Survey 2002, I document frequent discrepancies between self- vs. employer-reported educational attainment, and investigate their relative precision.

The Structure of Earnings Survey 2002 can be used to match the employer's report about workers' qualifications with their own self-reports. Different reported levels of qualification, within a given employer-employee pair, signal reporting error by at least one respondent. Moreover, the proportion of inconsistent reports poses a lower bound on the fraction of error-ridden reports. Inconsistencies are present in over 30% of observations from this survey; even after aggregating information into 4 main levels of educational qualifications (less than high school, high school, college and graduate degree), that is after dropping minor disagreements from the count, by this measure over 20% of observations are plagued by either a false report by the worker or an error by the employer.

Educational certificates constitute standardized and easily recorded information, and are used in the hiring process to screen workers' applications. However, getting consistent information on the level of these certificates from the employer and the worker appears to be more difficult than expected. Comparisons with other countries and surveys are limited, as contexts differ in many important aspects and there are no existing assessments of employers' knowledge of workers' degrees. The basic finding of significant error rates nevertheless echoes assessments of the accuracy of self-reports on education [Card, 1999, p. 1816] and administrative transcripts [Kane *et al.*, 1999; Bound *et al.*, 2001; Battistin & Sianesi, 2006a]. It also relates to comparisons of employer and employee responses on other items, such as industry, occupation, union coverage, hours worked and wages [Mellow & Sider, 1983; Mathiowetz, 1992].

Discordant reports on educational certificates are central to the methodology that I develop in order to estimate the reporting precision of workers and firms: under the hypotheses which define a model for the data generating process, information on the content of truth of each report is elicited from the conditional correlation of a third imperfect indicator of educational attainment with each report. This methodology builds on the work of [Kane *et al.*, 1999]; results indicate that about 17% of firms' reports are misclassified in the Structure of Earnings Survey; the error-rate for workers is only 7%.

1 Data

This paper uses the French Structure of Earnings Survey (SES) from the year 2002. This is a representative survey for French private or public firms, operating in the market sector of the economy and employing at least 10 workers. The

[1]Several studies are summarized in Bound *et al.* [2001]; more recent studies with newly available data include papers by Battistin & Sianesi [2006a,b].

smallest sampling unit is the individual worker employed at one of these firms; in 2002, over 120'000 employees were sampled, in about 13'600 establishments.

The French SES 2002 is both a business and a household survey; its objective is to correlate worker and firm characteristics with the level and structure of wage earnings. Firms are asked to complete a detailed wage-bill for each sampled worker, and to answer questions about employee characteristics that influence wages along with questions about their own wage setting practices. The workers' descriptions are completed by a questionnaire directly submitted to them, asking details about their career and family. All data refer to 2002, but the survey was carried out between April and December 2003.

1.1 Measures of Educational Attainment

In 2002, both questionnaires asked questions about the employee's educational attainment[2]. The official motivation for this apparent redundancy is not measurement error *per se*, but the high expected non-response rate. Highest completed level of education, while being a mandatory variable requested by EUROSTAT, is considered by the French survey administrators to be unknown to firms in many cases[3]. The fact that non-response rather than measurement error is the reason for asking the question twice is indirectly confirmed by evidence that, in preparing the data for EUROSTAT, the survey administrator merged the two responses: the worker's response was used only when the response to this item by the firm was missing.

The question asked to employers and employees about educational attainment is almost the same. A minor difference is the fact that employers are asked the highest (known) diploma of the worker, whilst the question for workers is more ambiguous, asking generically for obtained diplomas. Nevertheless, both are multiple choice questions, with the same 8 possible answers given (see table 1), and a note specifies that the provided list is a ranking, leaving no possible ambiguity about what the "highest" diploma is; consequently, it is possible to attribute discordant reports within an employer-employee pair to misreporting errors.

To eliminate any residual ambiguity in this paper I will make use of a 4 level scale for educational attainment instead of the original 8 level scale present in the survey. In the 4 level scale, the first level corresponds to workers who did not complete high school and get the "baccalauréat" certificate (levels 1,2,3,4; ISCED 0,1,2 and 3C); the second level corresponds to high school graduates, with no distinction between technical and general education (levels 5 and 6; ISCED 3A and 3B). "College" and "Graduate" levels are left unchanged.

Besides the two questions about diplomas, an additional source of information about educational attainment is age at school exit. This is asked to the worker before the question about earned diplomas. Workers are asked to report the age at which they stopped attending regularly school or university.

[2]A second information which is available from both questionnaires is the employee's citizenship.

[3]See on this the Decision regarding the 2006 wave of the survey by the Quality Label Committee of the French National Council for Statistical Information (CNIS): `http://www.cnis.fr/cnis/arretes/Avis-conformite/2005/ecmoss.pdf` (accessed in June 2007).

Table 1: Levels of Achievement

level	Definition (french)	ISCED	label
1	Aucun diplôme	0	less than HS
2	CEP	1	less than HS
3	Brevet des collèges, BEPC, brevet élémentaire	2	less than HS
4	CAP, BEP ou autre diplôme de ce niveau y compris diplômes sociaux et de la santé (aide-soignante, auxiliaire de vie...)	3C	less than HS
5	Baccalauréat technologique ou professionnel, brevet professionnel ou de technicien, BEA, BEC, BEI, BEH, BSEC	3B	High-School
6	Baccalauréat général, brevet supérieur, capacité en droit, DAEU	3A	High-School
7	Diplôme de 1er cycle universitaire, BTS, DUT, diplôme des professions sociales ou de la santé *(infirmier(ère))*	4,5A,5B	College
8	Diplôme de 2e ou 3e cycle universitaire y compris médecine, pharmacie, dentaire, doctorat, diplôme d'ingénieur, d'une grande école	6	Graduate

1.2 Wages

The Structure of Earnings Survey contains a rich and accurate set of measures of employee compensation. Throughout this study we will use hourly wages as defined by the ratio of total compensation (including variable components, but excluding all sorts of severance payments) divided by total hours worked.

1.3 Data selection and weighting

This study focuses on the informational content of the two diploma reports. For all analyses, the sample is logically restricted to observations for which both questionnaires are available (about 50'000 observations). In addition to this requirement, some loose consistency and relevance requirements are imposed on data.

- Trim observations lower than the first or higher than the last percentile of the hourly wage distribution (the 1st percentile is about 5 €; the last about 95 €; about 49'000 obs. are available after this step).
- Keep only full-time and full-year employees of the establishment; restrict the sample to prime-age workers only (30 to 59) (about 33'700 obs.)
- Drop employer questionnaires with more than one diploma ticked as 'highest'. Drop worker questionnaires if 'no diploma' and any other answer are ticked at the same time. Drop if the worker is still engaged in initial ed-

Table 2: Distribution of diploma reports

Firm			Worker		
	Frequency	Percent		Frequency	Percent
less th. HS	11066	33.54	less th. HS	15265	46.26
High-School	4174	12.65	High-School	4980	15.09
College	4360	13.21	College	5215	15.80
Graduate	5763	17.46	Graduate	7416	22.47
missing	7635	23.14	missing	122	0.37

Note: Total number of observations used is 32998 in both panels.

ucation; drop worker questionnaires if age at school exit is above 35 or missing (32'998 obs).

Most analyses in addition require that both reports are available on diplomas; total sample size is 25'275 in this case.

The SES Data include both pure sampling weights and adjustments for non-response within survey strata to make inference on the survey's target population. Within each firm, executive employees are sampled with a higher probability; this results in a higher proportion of observations with a high level of education relative to the true proportion in the survey's target population. Because the focus here is not on any population in particular, I do not use weights in any analysis.

1.4 Descriptive Statistics

A simple indicator of how employers and employees compare at reporting accurately the earned degrees of the worker is given by item non-response to this question. About 23% of completed firm questionnaires have missing answers to the question about the highest diploma; the corresponding rate for workers' questionnaires is below .5% (see table 2). However, this figure does not solve the question about how employers and employees compare when they both answer.

When both reports for diplomas are available, the joint distribution of reports reveals a high correlation, but also a non-negligible number of inconsistent reports: the agreement rate is below 80%, and this rate already excludes minor disagreements by aggregating answers up to a 4-level scale. This figure can be interpreted as indicating that for *at least* 20.4% of workers, either the employer or the workers themselves report education incorrectly.

More in detail, the proportions of answers falling in the various cells exhibit a slight imbalance revealing that self-reports tend to be higher than employer reports. Whether this corresponds to an under-reporting bias by the firm or to an over-reporting bias by the worker is still open at this stage, but the estimation of a statistical model that puts structure on the data will provide some evidence on that point.

2 Errors: Origins and Magnitude

The agreement rate of two partially independent reports of an ordinal variable is a crude indicator for the joint precision of the two reports. Mellow & Sider [1983,

Table 3: Joint distribution of diploma reports (4 levels)

firm	**worker**				
	less th. HS	High-School	College	Graduate	Total
less than HS	**38.64**	3.75	0.82	0.28	43.50
High-School	3.67	**9.00**	2.84	0.97	16.49
College	0.87	1.90	**11.57**	2.89	17.24
Graduate	0.29	0.51	1.57	**20.40**	22.78
Total	43.48	15.17	16.80	24.55	100.00

Agreement statistics:
Agreement rate: 79.6 (expected agreement rate, given the marginal distributions, if unrelated: 29.9); *Discordance rate*, by case: worker > employer: 11.6, worker < employer: 8.8; *Simple kappa*: 0.709; *Kendall's tau b*: 0.830.

Note: The table reports cell frequencies of the two-way tabulation of worker reports and firm reports, together with the two marginal distributions. All rates are expressed in percentage terms. Sample size is 25275.

p. 335] use the simple agreement rate to explore the determinants of errors, by regressing a dummy indicating agreement for various survey questions on a series of dummy variables for worker and firm characteristics, plus hourly wages.

However, agreement rates for diplomas do not only reflect accuracy on the part of respondents; they also reflect the underlying true distribution of qualifications. If the true distribution of qualifications has a clear mode, then two complete guesses will agree with high probability. For this reason, I use Cohen's kappa coefficients rather than the simple agreement rate to explore the determinants of discordant reports.

Cohen's kappa coefficient adjusts the simple, observed, agreement rate to take into account the possibility that agreement occurred by chance. Chance agreement, or expected agreement, is the level of agreement observed among two independent random draws from the observed marginal distributions of reports.

I use this indicator to explore whether errors in reporting the highest completed level of education are related (a) to the relevance of this information, (b) to discriminatory beliefs or (c) originate in memory flaws. I do so by dividing the sample along different observable dimensions and by studying how agreement levels vary across groups. I interpret a higher agreement as indicating lower error rates; this is granted, as a first approximation, even if measurement errors are correlated, as long as their correlation is not related to worker or firm characteristics.

Table 4, where kappa (κ) coefficients are presented for different sub-groups, reveals some regular patterns. First, the similar figures on agreement rates for male and female workers tend to exclude that firms under- or overreport education with different probabilities for men and women. Next, agreement rates rise with establishment size and with occupational level[4]. Both figures can

[4]When simple agreement rates are used, the inverse pattern is observed with respect to occupation. The fact that most unskilled blue-collar workers fall in the "less than High-School" category leads to very high agreement rates for them; the adjustment operated by the kappa coefficients reverts the comparison.

be justified with the varying importance of diplomas according to occupation and size of the production unit. It is probably more important, for firms, to have an accurate description of the education of their managerial staff, but much less so for unskilled workers. Similarly, the fact that small establishments have the lowest agreement rates, despite the proximity between worker and employer in these units, can be explained if this same proximity makes the information about diplomas less important, given the large number of personal characteristics that the employer can observe every day. The need for keeping registries of standardized information such as earned degrees on all employees is much more prevalent in large, impersonal, production units.

Finally, as shown in table 4, agreement statistics decline with age, but much less so with seniority. Agreement seems even to increase with seniority, up to a certain level. This could be deemed paradoxical, given the high correlation of age with seniority.

Inspection of the graph in figure 1 solves this paradox and shows the main factors driving this correlation with age and seniority. This graph shows simple agreement rates, rather than kappa coefficients. Each line in the graph displays almost no age or seniority gradient. Yet, the lines do not start at the same height: late entrants – workers hired when they are already 35 or older, which are likely to have completed education long before entering the firm – tend to have lower agreement rate short after hiring than their equally old co-workers, and to keep these lower agreement rates.

Overall, these pattern suggest that the main factor behind discordant reports is the irrelevance of this information at the moment of hiring. There is also some indication, however, of memory flaws on the workers' side, with declining agreement rates for the oldest workers.

This evidence corroborates the idea that employer's reports are not subject to changes over time, from the moment of hiring, and is thus consistent with the intuition that employer answers reflect codified employee records[5], while worker answers are produced out of memory.

2.1 Related Assessments of Measurement Error in Education

The difficulty of comparing agreement measures for different populations is even amplified when the comparison occurs across different data sources. Agreement measures are sensitive to the wording of questions, to the type of questions (open-ended vs multiple choice), to the heterogeneity of the population of interest, in addition to reflecting the knowledge and reliability of the respondent.

From comparisons of our agreement statistics with agreement measures previously computed on other similar data sets which contain a double information on educational attainment, we can nevertheless gain a reference point to assess the quality of the French employer-employee data used here.

A first reference point is given by reliability measures for a years of schooling variable. In summarizing previous research, Card [1999, p. 1816] finds that the reliability (or the signal-to-total-variance ratio) of self-reported years

[5]The survey administrators asked explicitly, in a note, employers to report the diploma declared by the worker at the time of her hire, or any diploma earned afterwards of which they have been informed.

Table 4: Agreement between reports of educational attainment

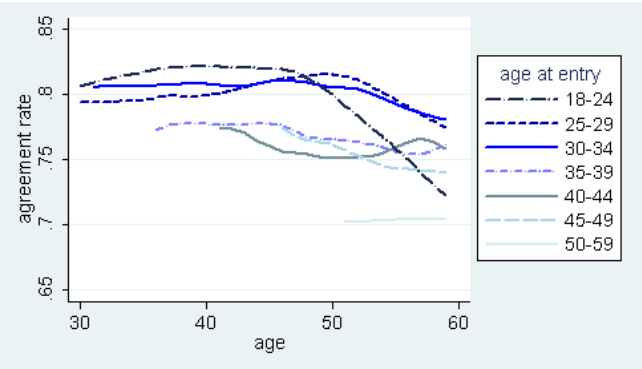
Cohen's kappa statistic for reports of educational attainment across subgroups

establishment characteristics	N	κ
industry		
C: mining and quarrying	*190*	0.75
D: manufacturing	*8989*	0.75
E: electricity, gas, water supply	*902*	0.83
F: construction	*1134*	0.72
G: wholesale and retail trade	*3406*	0.65
H: hotels and restaurants	*216*	0.51
I: transport, storage	*1864*	0.68
J: financial intermediation	*3371*	0.63
K: real estate, business activities	*5203*	0.68
establishment size		
size 0-9	*365*	0.64
size 10-49	*5939*	0.69
size 50-199	*7244*	0.70
size 200-499	*5190*	0.69
size 500-1999	*5954*	0.75
size 2000-	*583*	0.75
firm size		
size 10-49	*4422*	0.70
size 50-199	*4888*	0.69
size 200-499	*3656*	0.70
size 500-1999	*5455*	0.70
size 2000-	*6854*	0.73

worker characteristics	N	κ
sex		
men	*17597*	0.71
women	*7678*	0.70
age		
age 30-34	*4180*	0.72
age 35-39	*4633*	0.74
age 40-44	*4696*	0.73
age 45-49	*4439*	0.71
age 50-54	*4665*	0.66
age 55-59	*2662*	0.61
nationality		
french, born french	*22484*	0.71
french, born foreign	*1086*	0.66
foreign	*612*	0.69
worker-firm match characteristics		
occupation		
management workers	*10729*	0.64
clerical workers	*6391*	0.63
skilled blue-collar workers	*2614*	0.63
unskilled blue-collar workers	*5187*	0.57
seniority in firm		
0-4 years	*5310*	0.69
5-9 years	*3998*	0.71
10-14 years	*4699*	0.72
15-19 years	*2825*	0.72
20-24 years	*2972*	0.73
25-29 years	*2592*	0.70
30-34 years	*2025*	0.63

Source: SES2002, Full-time full-year employees (observations trimmed at the 1st and 99th percentile of the wage distribution), excluding observations with missing educational attainment.
Note: Sample size is 25275. The number of observations in each category does not necessarily sum up to the total sample size because of omitted categories and/or item non-response to the question defining the categories.
Cohen's kappa statistic is defined as $\kappa = \frac{P_O - P_E}{1 - P_E}$, where P_O is the agreement rate (observed proportion of agreement) and P_E is the proportion of agreement that can be expected if the two reports were random guesses with the observed marginal distribution.

Figure 1: Agreement rate by age and seniority, within cohorts defined by age at entry in the firm.



Note: The lines are kernel estimates of local averages of agreement rates. Given age at entry, age and seniority are perfectly collinear: due to grouping, however, each point represents one age only, but 5 possible seniorities.

of schooling is about 90%, with a similar reliability for administrative measures of schooling. Both self-reported and administrative sources are known to contain some measurement error, as repeated measures sometimes contain inconsistencies.

Reliability measures for categorical measures of educational attainment are less easily defined.

There are at least two sources which allow to compare educational self-reports on degrees with administrative transcript data. The National Longitudinal Study of the High School Class of 1972 has been used by Kane *et al.* [1999], and distinguishes three levels (no college, some college, BA+). This population is more homogeneous than the one examined in the present study, as all observations correspond to High School Graduates from the same cohort of graduates and all measures are taken at the same age. It is not surprising, then, that the two reports of educational achievement used in Kane *et al.* [1999], as reported in Bound *et al.* [2001, table 10], have a slightly higher agreement than our measures (the simple agreement rate .89 and Cohen's kappa is .83).

Data reported in Battistin & Sianesi [2006a, p. 17] from the British National Child Development Survey for a self-reported binary measure of educational achievement (no academic qualification vs. any academic qualification, recorded at age 23) and administrative transcripts for this same information also display high agreement rates (0.90) and kappa (0.79). But the same reasons can be cited for this, as the population is more uniform (same age, and same cohort) and only a binary classification of educational attainment is uses.

The two cited studies consider more homogeneous populations (one cohort only, with all individuals having at least some high school for Kane *et al.* [1999]) and distinguish fewer levels of education, both features having a positive impact on agreement rates.

To my knowledge, this study is the first considering errors in employer reports of educational attainment. Previously, employer and employee responses have been compared on other items: Mellow & Sider [1983], in particular, com-

puted agreement rates for survey questions on industry, occupation, union coverage, hours worked and wages. Mellow & Sider [1983], using CPS Data, find low agreement rates for employer and employee reports of occupation (0.576 at the 3-digits level, 0.810 at the 1-digit level), which Mathiowetz [1992] attributes mostly to the difficulty of coding survey answers.

This comparison yields two insights for interpreting SES data on diplomas. First, agreement statistics are in the lower range with respect to the cited studies, suggesting that independence of employer and employee reports of educational attainment in the SES 2002 is a plausible hypothesis. Second, if errors arise – as for occupation – within the process of answering the survey and had no link or impact with the reality of the workplace, then errors are essentially random and in particular non-differential with respect to covariates of (true) educational attainment. In the next section, I derive a model for the data-generating process which assumes independent and non-differential reports, which allows to identify the probability of every type of error in reporting.

3 A Model for the Error Generating Process

This section assumes that reporting errors occur independently across sources. It also assumes existence of a variable $\hat{z}_i$ which covaries with true education, but not with errors in reporting education. It will be shown that these hypotheses define a model for the data-generating process which allows identification and estimation of the probabilities of reporting wrongly any given diploma.

The model builds on the method developed by Kane *et al.* [1999] to correct for misclassification in estimating returns to education. Related arguments for identification of structural parameters from error-ridden data, when two independent non-differential measurements exist, are found in Mahajan [2006]; Lewbel [2007]; Battistin & Sianesi [2006b].

Let s_i^* be worker i's true educational attainment; let s_i^1 be the firm's belief about her educational attainment, and s_i^2 worker i's self-reported educational attainment.

Suppose that the worker reports level k ($s_i^2 = k$), while the firm reports level j ($s_i^1 = j$): we therefore observe the combination (j, k) for reports (s_i^1, s_i^2). Note that for any combination (j, k), the proportions reported in the cells of table 3 are an estimator for the ex-ante probability of observing it.

The hypothesis of independent errors makes it possible to write this probability in terms of the models parameters, i.e. the distribution of s_i^* and the error probabilities, conditional on each level of s_i^*. Indeed, if reporting errors occur independently across sources, the probability of observing a (j, k) pair of reports is equal to

$$\begin{aligned} \mathrm{P}(s_i^1, s_i^2 = j, k) &= \sum_{m=1}^{M} \mathrm{P}(s_i^1, s_i^2 = j, k | s_i^* = m) \cdot \mathrm{P}(s_i^* = m) \qquad (1) \\ &= \sum_{m=1}^{M} \mathrm{P}(s_i^1 = j | s_i^* = m) \cdot \mathrm{P}(s_i^2 = k | s_i^* = m) \cdot \mathrm{P}(s_i^* = m) \end{aligned}$$

where M is the cardinality of the set of degrees, i.e. the number of levels for s_i^*.

Within this model, the cell proportions in table 3 – the empirical counterpart to this probability – are thus informative about the probabilities on the right

hand side of the given expression. This constitutes the first source for identifying the parameters of the measurement model.

To derive this expression for $\mathrm{P}(s_i^1 = j, s_i^2 = k)$ we only used the hypothesis of independent reporting errors, together with the law of total probability.

The second source of information about the model's parameters is the mean of $\hat{z}_i$ for each combination of reports (j, k). By assumption, variable $\hat{z}_i$ verifies two conditions:

$$E(\hat{z}_i|s_i^* = j) \neq E(\hat{z}_i|s_i^* = k) \qquad \text{if } j \neq k \tag{2}$$

$$E(\hat{z}_i|s_i^*, s_i^1, s_i^2) = E(\hat{z}_i|s_i^*) \tag{3}$$

The first condition is similar to a rank condition: $\hat{z}_i$ must covary with true education. Condition (3) defines an exclusion restriction: $\hat{z}_i$ must have zero covariance with errors in reporting education ($s_i^1 - s_i^*$ and $s_i^2 - s^*i$). Variable $\hat{z}_i$ acts as an instrument to identify structural parameters from the data-generating process of s_i^1 and s_i^2. The conditions that it does verify are very similar to those of usual instrumental variables. I therefore call $\hat{z}_i$ an instrument-like variable.

Under conditions (2) and (3), the expected value for $\hat{z}_i$ given reports $(s_i^1, s_i^2) = (j, k)$ can be written as a function of model parameters and observed quantities:

$$\begin{aligned} &E(\hat{z}_i|s_i^1, s_i^2 = j, k) = \\ &\quad = \sum_{m=1}^{M} E(\hat{z}_i|s_i^*, s_i^1, s_i^2 = m, j, k) \times \mathrm{P}(s_i^* = m|s_i^1, s_i^2 = j, k) \\ &\quad = \sum_{m=1}^{M} E(\hat{z}_i|s_i^* = m) \times \frac{\mathrm{P}(s_i^1, s_i^2 = j, k|s_i^* = m) \cdot \mathrm{P}(s_i^* = m)}{\mathrm{P}(s_i^1, s_i^2 = j, k)} \end{aligned} \tag{4}$$

where the fraction on the right hand side of eq. (4) has the same expression as in (1) at the numerator, and an observed quantity as the denominator.

To develop intuition, $\hat{z}_i$ can be seen as a proxy variable for educational attainment, which provides a metrics for the levels of s_i^*. Assume that levels of s_i^* are ranked such that the mean of $\hat{z}_i$ is increasing in levels of s_i^*: $E(\hat{z}_i|s_i^* = j) > E(\hat{z}_i|s_i^2 = k)$ if and only if $j > k$.

Because by assumption the covariance of $\hat{z}_i$ with the two error-ridden reports (s_i^1 and s_i^2) only relies on the dependence of signals s_i^1 and s_i^2 from s_i^*, if the average of $\hat{z}_i$ for some combination (j, k) of signals is higher than for a second combination (l, m), then the underlying average level of s_i^* is also higher.

The formal argument for identification is given below, using matrix notations.

3.1 A measurement model with independent errors

Non-differential measurement error with respect to $\hat{z}_i$ corresponds formally to the following definition. Let s_i^* be worker i's true educational attainment; let s_i^1 be the firm's belief about her educational attainment, and s_i^2 worker i's self-reported educational attainment; x_i denotes a vector of observed covariates. Under non-differential measurement errors, the error-ridden measurements do not provide information about $\hat{z}_i$ beyond the information contained in the true value:

$$f_{\hat{z}_i|s_i^*, s_i^1, s_i^2, x_i} = f_{\hat{z}_i|s_i^*, x_i} \tag{5}$$

($f_{a|b}$ denotes the conditional distribution of a given b).

Suppose there exists a variable $\hat{z}_i$ in our data set with respect to which measurement errors are non-differential. Assuming $E(\hat{z}_i|s_i^*)$ exists, I define a model for the conditional mean as[6]:

$$E(\hat{z}_i|\mathbf{s}_i^*, \mathbf{s}_i^1, \mathbf{s}_i^2) = E(\hat{z}_i|\mathbf{s}_i^*) = \mathbf{s}_i^{*\prime}\tilde{\zeta} \tag{6}$$

Let $\mathbf{d}_i$ be a function of $\mathbf{s}_i^2$ and $\mathbf{s}_i^1$ only, defined as $\mathbf{d}_i = \text{vec}(\mathbf{s}_i^2\mathbf{s}_i^{1\prime})$: $\mathbf{d}_i$ is a vector that stacks the columns of matrix $\mathbf{s}_i^2\mathbf{s}_i^{1\prime}$, whose cells correspond to all possible combination of measurements.

Using this notation and from equation 6, we can write

$$E(\mathbf{d}_i\hat{z}_i|\mathbf{s}_i^*, \mathbf{s}_i^1, \mathbf{s}_i^2) = \mathbf{d}_i E(\hat{z}_i|\mathbf{s}_i^*) = \mathbf{d}_i\mathbf{s}_i^{*\prime}\tilde{\zeta}$$

and, thus, the following $M \times M$ unconditional moment restrictions[7]:

$$E(\mathbf{d}_i\hat{z}_i) = E(\mathbf{d}_i\mathbf{s}_i^{*\prime})\tilde{\zeta}$$

If survey reports of educational attainment are conditionally independent it is possible to identify $E(\mathbf{d}_i\mathbf{s}_i^{*\prime})$: with independent measurements, measurements verify the following system of equations.

$$\begin{cases} \mathbf{s}_i^{1\prime} = \mathbf{s}_i^{*\prime}\mathbf{\Pi}_1 + \mathbf{e}_i^{1\prime} & E(\mathbf{s}_i^*\mathbf{e}_i^{1\prime}) = 0 \\ \mathbf{s}_i^{2\prime} = \mathbf{s}_i^{*\prime}\mathbf{\Pi}_2 + \mathbf{e}_i^{2\prime} & E(\mathbf{s}_i^*\mathbf{e}_i^{2\prime}) = 0 \\ E(\mathbf{e}_i^2\mathbf{e}_i^{1\prime}) = 0 \end{cases} \tag{7}$$

This measurement model says, through matrix $\mathbf{\Pi}$, with which probability each level of completed education results in a reported level of completed education (each line in matrices $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ sums to 1).

Independence ensures that matrix $E(\mathbf{d}_i\mathbf{s}_i^{*\prime})$ depends on the distribution of true education in the population ($E(\mathbf{s}_i^*) = \delta^*$) and on the conditional probabilities in $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$, in a simple way. By appropriately combining information in $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ in a single matrix $\mathbf{T}$[8], $E(\mathbf{d}_i\mathbf{s}_i^{*\prime})$ can be given the following representation:

$$E(\mathbf{d}_i\mathbf{s}_i^{*\prime}) = [\mathbf{T}'diag(\delta^*)] \tag{8}$$

The above hypotheses (6 and 7) are then sufficient to identify $\mathbf{T}$ and δ^*, along with $\tilde{\zeta}$ through the following just identified system of $2J^2 - 1$ independent moment restrictions[9]:

[6] I use bold letters to denote vectors: in this section, educational attainment is not limited to be a binary variable. $\mathbf{s}_i^*, \mathbf{s}_i^1, \mathbf{s}_i^2$ are complete sets of mutually exclusive dummy variables, each representing a diploma or a report.

[7] M represents the dimension of vector $\mathbf{s}_i^*$.

[8] Each element of $\mathbf{T}$ is the product of one element in $\mathbf{\Pi}_1$ and one element in $\mathbf{\Pi}_2$; $\mathbf{T}$ contains therefore $2(J-1)J$ independent parameters. If π_n^1, π_n^2 and $\mathbf{t}_n$ are the n-th rows of $\mathbf{\Pi}_1$, $\mathbf{\Pi}_2$, $\mathbf{T}$, then $\mathbf{t}_n = [vec(\pi_n^{2\prime}\pi_n^1)]'$.

[9] I use ζ to denote the element-wise product of δ^* and $\tilde{\zeta}$.

$$E(\mathbf{d}_i) = \mathbf{T}'\delta^* \tag{9}$$

$$E(\mathbf{d}_i \hat{z}_i) = \mathbf{T}'\zeta \tag{10}$$

Moment conditions in (10) identify the system only if they add independent information beyond that contained on moment conditions in (9): to ensure this, variable $\hat{z}_i$ has to correlate with s_i^*.

3.2 Estimation

The model defined by (non-linear) equations (9) and (10) can be estimated if a variable which verifies the assumptions for $\hat{z}_i$ exists. With 4 distinct levels of educational attainment, the system defined by moment conditions (9) and (10) involves 31 equations and 31 parameters.

To ease computational burden, I perform estimation of the parameters and estimation of standard errors of the parameters in two separate steps.

A preliminary step consists of estimating the empirical counterparts to the left-hand-side of (9) and (10); the empirical counterpart of matrix $E(\mathbf{d}_i)$ stacks the columns of table 3; the empirical counterpart of matrix $E(\mathbf{d}_i \hat{z}_i)$ lists, in the same order, the elementwise products of entries in table 3 and average levels of $\hat{z}_i$ for each combination of reports.

These means are the only moments used for estimation of the parameters; the variance-covariance matrix of these means enters the computation for the standard errors of these parameters.

Estimation of parameter values is then equivalent to the search for a solution to the system defined by moment conditions (9) and (10). Because the model is just identified, the solution space is not empty. I solve this system and verify the solution[10].

By definition, the estimated parameter values verify the estimating equations perfectly; standard errors for parameters can be computed in a second step. In particular, a "sandwich" formula allows to compute parameter standard errors from the standard errors for the empirical means used in the first step [Benichou & Gail, 1989]. This transformation rests on the implicit functions theorem combined with the so-called delta method. The crucial assumption is that the Jacobian of the estimating equations with respect to parameters is of full rank – and thus verifies the assumptions for the implicit functions theorem. In particular, let θ indicate the vector of parameters; if $\mathbf{J}$ is the Jacobian (with respect to θ) of the $2J^2 - 1$ independent equations on the right-hand side of (9) and (10), and $\mathbf{\Sigma}$ is the variance-covariance matrix of moments on the left-hand side, the following result holds:

$$V^{as}(\theta) = \mathbf{J^{-1}}'\mathbf{\Sigma}\mathbf{J^{-1}} \tag{11}$$

[10] To solve the system I use the `nlsur` routine in Stata 10. I provide plausible initial values, whereby all error rates are set close to zero, and $E(\mathbf{s}_i^*)$ and the coefficient from the linear projection of $\hat{z}_i$ on $\mathbf{s}_i^*$ are initialised as their counterparts for the worker report. In a first estimation step, I apply the logit transformation $F : p \rightarrow \exp(p)/(1+\exp(p))$ to all parameters representing probabilities in moment conditions. I use the implied values from this first step as initial values for a second, unrestricted, estimation step.

Replacing the estimated Jacobian for $\mathbf{J}$ and the estimated variance-covariance matrix for $\mathbf{\Sigma}$ yields a consistent estimate for the asymptotic variance of θ.

In the present case the estimating equations only consist of sums and products of parameters: the Jacobian is therefore easily derived, and this two step procedure proves convenient.

To make inference robust to the clustered design of the survey, it is sufficient to use the robust estimator for the matrix-covariance matrix of reduced form parameters $\mathbf{\Sigma}$.

4 Assessing the Precision and Biases in Each Report: Results

4.1 Instrument-like variables

The model for the reporting behavior of firms and workers is identified if a valid instrument-like variable is available. As is often the case, instruments are not easy to find; while the rank condition can be empirically defended, the validity of the exclusion restriction rests on theoretical arguments.

This model is no exception to this pattern. However, in contrast to the usual application of instrumental variables, the required exclusion restriction is not with respect to some (real) outcome measure, but only with respect to reporting behavior. Thus, theoretically, good instruments are not difficult to construct. Any independent proxy for workers' qualifications makes a good instrument.

Theoretical examples are of little use when the choice is restricted to variables measured within the same survey.

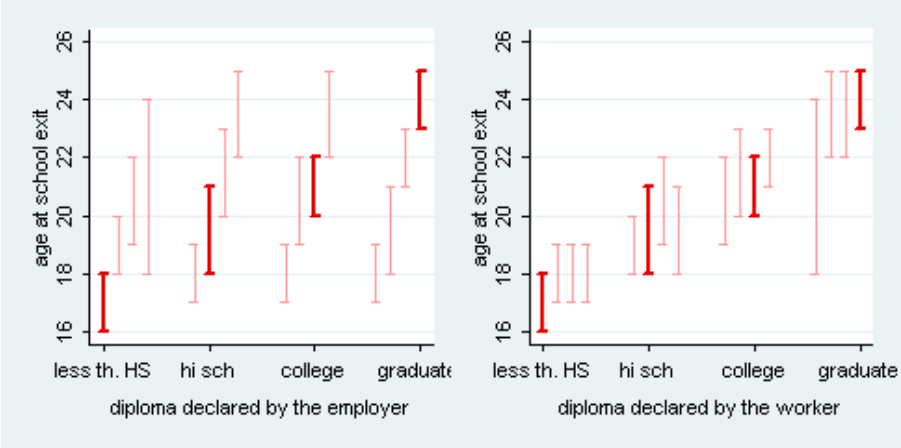
In the context of the Structure of Earnings Survey, the natural choice for an instrument for educational attainment is the age at which workers declare having finished attending school or university on a regular base. This information, collected from workers' declarations, is obviously strongly related to the highest earned diploma. Using this information as an instrument-like variable $\hat{z}_i$ supposes that age at school exit does not influence the probability of reporting educational attainment incorrectly; it also supposes that the eventual error the worker commits in reporting this information is unrelated to the error he commits in reporting his highest diploma.

4.2 Preliminary Evidence

Figure 2 illustrates graphically how reports relate to the third indirect indicator of educational attainment, age at school exit.

Age at school exit seems to be more consistent with the worker's declaration on achievement than with the employer's declaration. The left graph shows that, among workers for which their employers declared the same degree level, age at school exit strongly relates to what the worker declares. In contrast, the right graph reveals that workers' age at school exit is almost unrelated to what employer believe their education to be: groups defined by workers' reports share similar distributions of age at school-exit. Age at school exit thus rises systematically and in a predictible way with worker reports, but not, or much less, with employer answers (see figure 2)

Figure 2: Interquartile range of age at school exit by joint diploma report



Note: Each bar corresponds to a distinct combination of worker and employer report on diploma. For each possible combination, the corresponding bar connects the upper and lower quartile of the distribution of age at school exit. The dark bars are for concordant reports of diplomas. In the left graph, bars are grouped over employer reports: for each employer report, bars on the left of the dark bar represent observations with lower worker reports, bars on the right represent observations with higher worker reports. The right graph has the same bars grouped over worker reports.

If age at school exit can be considered an independent indicator of the true educational achievement, and its measure in our survey is not related to errors in reporting or assessing education, the pattern in figure 2 indicates higher precision for workers in reporting their diploma.

The pattern in figure 2 is confirmed by rank correlation coefficients. Spearman's rank correlation coefficient for age at school exit with worker reports on diplomas is 0.83, while the corresponding figure for age at school exit with firm reports is slightly lower, at 0.78.

4.3 Results

Table 5 displays the estimated results, and the implied error probabilities from the estimation of the measurement model using age at school exit as the instrument-like variable.

According to these estimates, errors in reporting diplomas tend to fall near the truth. Conditional on the true level being a high-school degree, the most frequent error corresponds to a report of less than a high-school degree – an error which 13.4% of workers with a high-school degree make, as well as 21.5% of employers of high-school graduates.

Second, the table also reveals that under-reports occur with higher probability than over-reports. This can be seen from the conditional probabilities in the top panel of table 5; to summarize this evidence, we also computed the unconditional error probabilities implied by the estimated parameters.

More in detail, we compute the unconditional probability of reporting the right diploma ("precision"); the unconditional probability of reporting a lower diploma than the true one ("under"); and the unconditional probability of reporting a higher diploma than the true one.These objects are formally defined

Table 5: Parameter estimates and implied unconditional error probabilities

	worker				employer			
truth	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
(1)	**0.972****	0.024**	0.003**	0.001*	**0.927****	0.054**	0.014**	0.005**
	(0.003)	*(0.003)*	*(0.001)*	*(0.001)*	*(0.003)*	*(0.003)*	*(0.001)*	*(0.001)*
(2)	0.134**	**0.834****	0.017	0.015**	0.215**	**0.680****	0.074**	0.030**
	(0.009)	*(0.015)*	*(0.013)*	*(0.005)*	*(0.010)*	*(0.011)*	*(0.008)*	*(0.004)*
(3)	0.011**	0.070**	**0.876****	0.043**	0.040**	0.165**	**0.720****	0.074**
	(0.003)	*(0.008)*	*(0.012)*	*(0.009)*	*(0.004)*	*(0.009)*	*(0.011)*	*(0.006)*
(4)	0.001†	0.001	0.018**	**0.979****	0.006**	0.029**	0.098**	**0.866****
	(0.001)	*(0.001)*	*(0.003)*	*(0.004)*	*(0.001)*	*(0.003)*	*(0.006)*	*(0.007)*

	(1)	(2)	(3)	(4)
$\widehat{\mathrm{P}(s_i^* = m)}$	0.424**	0.154**	0.183**	**0.240****
std. err.	*(0.004)*	*(0.004)*	*(0.004)*	*(0.004)*
$\widehat{\mathrm{E}(\hat{z}_i \mid s_i^* = m)}$	16.84**	19.58**	21.49**	24.09**
std. err.	*(0.02)*	*(0.04)*	*(0.04)*	*(0.03)*

Unconditional Error Probabilities

	worker		firm	
	est.	std.err.	est.	std.err.
precision	**0.935****	*(0.003)*	**0.837****	*(0.003)*
over-rep	**0.025****	*(0.003)*	**0.060****	*(0.002)*
under-rep	**0.041****	*(0.002)*	**0.103****	*(0.003)*

Notes: Robust standard errors allowing for correlation of residuals within establishments are shown in parentheses. Bold numbers signal that the parameter is defined as a deterministic function of other parameters and was not estimated within the system of estimating equations. Standard errors for these parameters are computed with the delta-method.
†: Significant at the 10% level. *: significant at the 5% level. **: significant at the 5% level. For diagonal terms in the matrix of conditional error probabilities, the test is with respect to $H_0: \ \hat{\theta} = 1$. For all other parameters the usual null of a parameter estimate equal to zero is used.

as follows:

$$
\begin{aligned}
\text{precision} &= \sum_{m=1}^{M} \Pr(s_i^* = m) \cdot \Pr(s_i^n = m | s_i^* = m) \\
\text{over} &= \sum_{m=1}^{M-1} \sum_{j=m+1}^{M} \Pr(s_i^* = m) \cdot \Pr(s_i^n = j | s_i^* = m) \\
\text{under} &= \sum_{m=2}^{M} \sum_{j=1}^{m-1} \Pr(s_i^* = m) \cdot \Pr(s_i^n = j | s_i^* = m) \qquad \text{for } n = 1, 2
\end{aligned}
$$

An estimator for these object is obtained by substituting probabilities with estimated probabilities, and standard errors are derived through delta method.

As these unconditional probabilities show, I estimate that 93.5% of workers report the correct diploma, while only 83.7% of firms which report their workers educational attainment classify workers correctly. Both for workers and for firms, under-reports of educational attainment seem to be more frequent than over-reports.

Table 5 also reports the the estimated proportion of observations within each diploma category, and the standard error of this estimate, as well as the estimate for the conditional expectation of age at school exit by level of educational attainment.

These estimates correspond to plausible values for these quantities.

4.4 Robustness Checks

As robustness checks, I perform the estimation of the model for reporting behavior on sub-samples, defined along dimensions which I expect to influence either error-reporting probabilities, or the distribution of diplomas. Full results are reported in appendix B.

The first comparison is between the model estimates for male and female workers. Consistent with previous evidence from table 4, there is hardly any difference in overall error rates for men and women: worker reports are correct for approximately 93% of observations, and firm reports for 83% of cases, for both female and male workers. Errors are, however, qualitatively somewhat different. Over-reports are more prevalent for male workers: their proportion is above 40% of errors both in self-reports and in firm-reports, whereas over-reports represent less than 30% of errors for female workers.

The next comparison is between small and big establishments. Unconditional error probabilities again confirm roughly the proportions of 17% of misclassified firm reports and 7% of misclassified worker reports. The distribution of degrees however is not the same in the two types of establishments: unqualified workers work much more frequently for small establishments. Given that the probability of correctly reporting degrees for someone who does not have any high-school qualification is very high, this contributes to a much greater extent to reduce the unconditional error rates in small establishments than in large establishments. Indeed, a look at conditional error rates reveals that small es-

tablishments correctly report higher degrees with a lower probability than big establishments.

Finally, the comparison of estimates on young vs old workers confirms the intuition that young workers - despite having higher qualifications - are somewhat more precise in reporting their degrees. The unconditional error rate for young workers is 5%, and the corresponding rate for old workers is 8%. Moreover, both among firms' and workers' errors, the proportion of over-reports of education tends to increase with age.

These estimates on sub-samples tend to indirectly validate the ability of the model to uncover the reporting behavior; indeed, many of their quantitative and qualitative predictions join our intuitive priors about what drives error rates.

5 Do Errors Matter? The Correlation of Reporting Error with Wages

5.1 Wages as Instruments

If measurement errors arise from random mistakes in reporting the truth, then any variable which has a real-world correlation with educational attainment should be unrelated to these errors. In a world of random measurement error, thus, wages satisfy to the conditions for acting as an instrument for uncovering the reporting behavior of firms and workers.

Non-differential measurement error is an assumption which is often made in the context of misclassification errors to analyze the biases that such errors produce in estimates. In this respect, non-differential measurement error is similar to the hypothesis of "classical" measurement error for continuous variables.

In the context of diploma reports, assuming that misclassification errors are non-differential amounts to as much as assuming that – conditional on true educational attainment – the propensity to lie to the survey, or to make a mistake, is uncorrelated with wages.

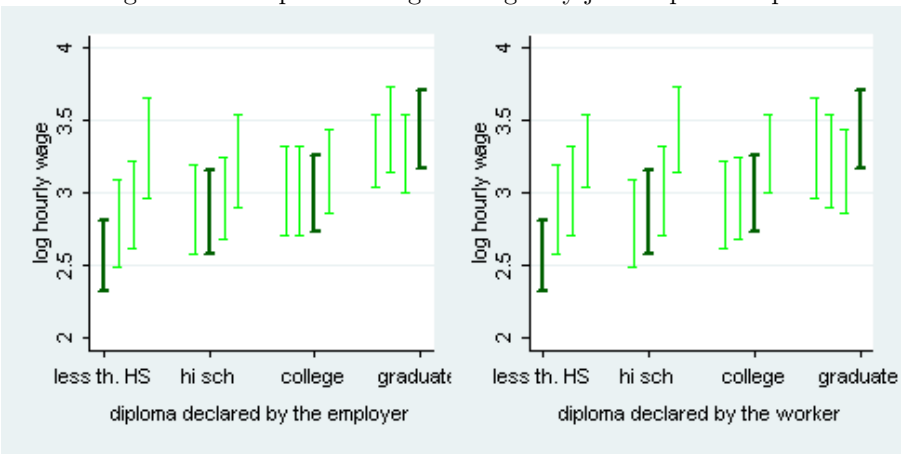
For workers, this means that well qualified high wage workers, for instance, do not report their qualifications differently than equally qualified workers who happen to be less paid. It also implies that – if lying to the survey is a proxy for lying to your employer – there is no return attached to lies about your qualifications.

For firms, assuming that wages are uncorrelated to reporting errors similarly implies that – if firm mistakes originate in the hiring process – there is no return attached to wrong assessments of workers' qualifications. Taking the inverse causality, it also implies that wages do not have an influence on the reporting behavior of firms.

To sum up, the uncorrelated nature of errors and wages excludes (a) any causal link from errors to wages, (b) any inverse causality from wages to errors, and (c) any simultaneity channel by which some unobserved third variable – some special form of relational human capital, for instance – influences both wages and the reporting behavior to the survey.

In the basic model presented so far, the assumptions for instrument-like variables are not testable. We can only compare the model's prediction to our priors about the reporting behavior of firms and workers. However, as we will see

Figure 3: Interquartile range of wages by joint diploma report



Note: Each bar corresponds to a combination of worker and employer report on diploma. For each possible combination, the corresponding bar connects the upper and lower quartile of log hourly wages. The dark bars are for concordant reports of diplomas. In the upper graph, bars are grouped over employer reports: for each employer report, bars on the left of the dark bar represent observations with lower worker reports, bars on the right represent observations with higher worker reports. The lower graph has the same bars grouped over worker reports.

later, when two or more instrument-like variables are available, we can test their joint compatibility because the estimating equations become over-identified.

In the same way as we did with age at school exit, we can gain useful insights on the predictions of a simple model with non-differential measurement error from figure 3. This figure shows how wages vary when worker reports vary, holding employer reports constant (on the left); and how wages vary with employer reports, holding worker reports constant (on the right).

Differently than in figure 2 (page 16), wages increase with both employer and employee reports: holding one report constant, the second generally correlates positively with wages. Moreover, wages seem to increase more steeply with employer reports than with workers' reports, at least for the three lower levels of qualification. If measurement errors were not correlated with wages (directly, or through their correlation with unobservable productive assets), this would indicate that a high employer report is a better indicator for high educational attainment than a high worker report.

Table 6 presents the results from estimation of our model for the reporting behavior of firms and workers, when wages are assumed to be uncorrelated with reporting errors.

These results confirm the insight from figure 3. When wages are used as instruments, the prediction of the model is that firms are more precise in reporting degrees (with an unconditional error rate just above 9%) than workers, whose unconditional error rates is estimated above 15%. This appears contrary to intuition. Moreover, closer inspection also reveals that firms never over-declare their workers' degrees. Given that this conclusion stems from the fact that errors are assumed to be uncorrelated with wages, this could indirectly be explained if, when firms over-declare their workers' qualifications, they also pay them a higher wage.

Table 6: Parameter estimates and implied unconditional error probabilities when wages are used as instrument-like variables

	worker				**employer**			
truth	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
(1)	**0.967****	0.029**	0.004†	0.000	**1.008**	-0.010	0.000	0.001*
	(0.006)	*(0.005)*	*(0.002)*	*(0.001)*	*(0.007)*	*(0.006)*	*(0.002)*	*(0.001)*
(2)	0.285**	**0.637****	0.067*	0.010†	0.232**	**0.784****	-0.012	-0.004
	(0.018)	*(0.025)*	*(0.028)*	*(0.006)*	*(0.019)*	*(0.032)*	*(0.033)*	*(0.004)*
(3)	0.056**	0.123**	**0.705****	0.117**	0.028**	0.132**	**0.804****	0.036**
	(0.008)	*(0.019)*	*(0.022)*	*(0.008)*	*(0.008)*	*(0.026)*	*(0.030)*	*(0.006)*
(4)	0.010**	0.021**	0.048**	**0.921****	0.008**	0.024**	0.045**	**0.924****
	(0.002)	*(0.002)*	*(0.004)*	*(0.005)*	*(0.002)*	*(0.003)*	*(0.005)*	*(0.007)*

	(1)	(2)	(3)	(4)
$\mathrm{P}(\widehat{s_i^* = m})$	0.384**	0.174**	0.204**	**0.239****
std. err.	*(0.007)*	*(0.012)*	*(0.011)*	*(0.004)*
$\mathrm{E}(\widehat{\hat{z}_i \mid s_i^* = m})$	2.59**	2.88**	3.00**	3.45**
std. err.	*(0.00)*	*(0.01)*	*(0.01)*	*(0.01)*

Unconditional Error Probabilities

	worker		**firm**	
	est.	std.err.	est.	std.err.
precision	**0.846****	*(0.006)*	**0.908****	*(0.008)*
over-rep	**0.050****	*(0.006)*	**0.001**	*(0.006)*
under-rep	**0.105****	*(0.006)*	**0.091****	*(0.007)*

Notes: see table 5.

A final observation which casts doubts on the validity of wages as instruments is the implied distribution for educational attainment among survey respondents, $P(s_i^* = m)$. This distribution is quite different from the marginal distributions of both reports ($P(s_i^1 = m)$ and $P(s_i^2 = m)$) as estimated from table 3.

5.2 Over-identification Tests

In the presence of more than one variable with respect to which measurement error is non-differential, an over-identification test performed after joint estimation of the resulting estimating equations can be implemented.

Suppose that there are two possible instrument-like variable; the system defined by estimating equations (9) and (10) is identified with just one of them. The second instrument-like variable adds a second set of $M \times M$ equations in the form of (10) to estimation, but only increases the number of parameters to be estimated by M (those in ζ): all parameters in $\mathbf{T}$ are over-identified.

Over-identification corresponds to the situation in which parameter estimates, derived separately from the previous just-identified models using instrument-like variables independently, are consistent. We have previously seen that this is unlikely to be the case: the over-identification test will formalize the argument.

An optimal minimum distance estimator is applied to the over-identified system including two sets of moment conditions involving instrument-like variables. Minimum distance estimation of this system involves $3 \times M^2 - 1$ independent reduced form parameters (the means on the left-hand side), which are defined as a function of only $2 \times M^2 + M - 1$ structural parameters. As an optimal weighting matrix the inverse of the variance-covariance matrix of reduced form parameters is used.

Denoting the vector of reduced-form parameters (the empirical counterparts to the left-hand-side of (9) and (10)) by $\mathbf{p}$ and the vector of structural parameters in $(\mathbf{T}, \delta^*, \zeta_1, \zeta_2)$ by θ, the estimator is the solution to minimization problem

$$\min_{\theta} Q(\theta) = (\hat{\mathbf{p}} - g(\theta))' \hat{\mathbf{V}}(\hat{\mathbf{p}})^{-1} (\hat{\mathbf{p}} - g(\theta)) \tag{12}$$

where vector function g stacks restrictions imposed in (9) and (10)), such that under the null hypothesis that the moment conditions are verified by the data generating process, we have

$$(\hat{\mathbf{p}} - g(\theta)) = 0$$

Moreover, under the null

$$H_0 \quad : \quad g(\theta_0) = \mathbf{p}_0$$

(where θ_0 and $\mathbf{p}_0$ denote true values of the parameters), the objective function is asymptotically distributed as $\chi^2(M^2 - M)$. This provides an over-identification test for the joint validity of instrument-like variables.

I perform the minimization of programme (12) for the system where two sets of restrictions in the form of equation (10), one for wages, one for age at school exit, are included.

The test statistic $Q(\hat{\theta})$, which, under the null, has a χ^2 distribution with 12 degrees of freedom, equals 380.9; the p-value for the null of jointly valid

instruments is 0.000. As could be expected from simple inspection of results in tables 5 and 6, age at school exit and wages cannot both be valid instruments.

Comparing the estimates from the just-identified models with priors, I argue that wages are unlikely to fulfill conditions for acting as an instrument-like variable. In particular, I suspect that employers' errors in reporting education correlate with wages.

5.3 A Fully Identified Model with Differential Measurement Error

Wages intuitively seem to be correlated with employers' errors in reporting degrees. But how important is this correlation, and how does it compare to the correlation of true education with wages?

To answer these questions I adapt the model from the previous sections to allow for measurement error to be differential with respect to wages. In particular, I assume

$$E(w_i|s_i^*, s_i^1, s_i^2) = E(w_i|s_i^*, s_i^1) \tag{13}$$

This allows me to add a new set of estimating equations – relating reduced form parameters to structural parameters – to the system formed by equations (9) and (10): for each combination $s_i^1, s_i^2 = i, j$

$$\begin{aligned} &E(w_i|s_i^1, s_i^2 = j, k) = \\ &\quad = \sum_{m=1}^{M} E(w_i|s_i^*, s_i^1, s_i^2 = m, j, k) \cdot \mathrm{P}(s_i^* = m|s_i^1, s_i^2 = j, k) \\ &\quad = \sum_{m=1}^{M} E(w_i|s_i^*, s_i^1 = m, j) \cdot \frac{\mathrm{P}(s_i^1, s_i^2 = j, k|s_i^* = m) \cdot \mathrm{P}(s_i^* = m)}{\mathrm{P}(s_i^1, s_i^2 = j, k)} \end{aligned} \tag{14}$$

The right-hand side of equation 13 corresponds to $M \times M$ parameters to estimate; the remaining parameters in equation (14) are already identified if misclassification errors are non-differential with respect to some variable $\hat{z}_i$.

Table 7 presents the estimates for conditional expectations of wages, and the implied estimates for wage gains.

Results indicate that workers for which the employer report is positively biased have significant wage gains. On the other hand, workers for which the employer report is negatively biased do not, in general, have lower wages. The only significant elements to the left of the diagonal are for graduate workers, but the pattern is less clear than for wage gains, and, in any case, even when penalized by a negative perception graduate workers still earn more than non-graduate workers do.

5.4 Interpretation

My findings are consistent with the following story:

Errors are of two types – survey errors, which arise in the process of answering the survey and do not have consequences in the real worlds, and lies.

Table 7: Conditional Expected Wages, Wage Premia and Wage Penalties

	Cond. Expected Wages				**Wage Premia and Wage Penalties**			
	employer				**employer**			
truth	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
(1)	**2.596**	2.929	3.007	3.212	.	0.333**	0.410**	0.615**
	(0.005)	*(0.026)*	*(0.051)*	*(0.084)*		*(0.027)*	*(0.027)*	*(0.027)*
(2)	2.858	2.879	3.022	3.451	-0.021		0.142**	0.571**
	(0.024)	*(0.010)*	*(0.042)*	*(0.055)*	*(0.026)*		*(0.043)*	*(0.056)*
(3)	2.946	2.966	2.998	3.254	-0.053	-0.032	.	0.256**
	(0.047)	*(0.019)*	*(0.008)*	*(0.031)*	*(0.047)*	*(0.020)*		*(0.032)*
(4)	3.665	3.346	3.188	3.445	0.220	-0.098†	-0.257**	.
	(0.212)	*(0.057)*	*(0.025)*	*(0.007)*	*(0.212)*	*(0.057)*	*(0.026)*	

Notes: Estimates for conditional expectations of wages, together with standard errors, are reported in the left panel (no significance tests are computed for these parameters). The right panel presents the implied wage gains and wage penalties, together with delta-method standard errors. These are defined as the difference with the diagonal element on the same row. Stars indicate the significance level of these wage gains and penalties.
†: Significant at the 10% level. *: significant at the 5% level. **: significant at the 5% level.

When a new employer-employee match is formed, the employer has some bargaining power; as a consequence, the subjective perception that he has of the employer reflects in the wage that the employee is offered. Positive biases are to the worker's advantage, and such offers are obviously accepted. Negative biases are to the worker's disadvantage, but some of these offers are nevertheless accepted.

However, the market, and in particular on-the-job search, prevents the matches with a negatively biased wage offer from lasting. On the other hand, positively biased wage offers translate into long-term employment relations. Therefore, in the survey, errors are either random, or positively biased.

Our interpretation has some similarity with a "winners curse" story; the employer which has a high subjective valuation for a worker is successful at hiring him, but is doomed to pay him higher than competitive wages.

6 Conclusive remarks

- possible applications of the empirical strategy: measuring discrimination (nationality). Union status (Jakubson).
- what are the implications of errors for ususal estimates of returns to education?

References

Aeberhardt, Romain, & Pouget, Julien. 2007 (May). *National Origin Wage Differentials in France: Evidence from Matched Employer-Employee Data.* IZA DP no 2779.

Battistin, Erich, & Sianesi, Barbara. 2006a. *Misreported Schooling and Returns to Education: Evidence from the UK.* Institute for Fiscal Studies CEMMAP working paper CWP07/06.

Battistin, Erich, & Sianesi, Barbara. 2006b. *Misreported Schooling, Multiple Measures and Returns to Educational Qualifications.* Unpublished Manuscript.

Benichou, Jacques, & Gail, Mitchell H. 1989. A Delta Method for Implicitly Defined Random Variables. *The American Statistician*, **43**(1), 41–44.

Bound, John, Brown, Charles, & Mathiowetz, Nancy. 2001. Measurement Error in Survey Data. *Pages 3705–3843 of: Handbook of Econometrics*, vol. 5. North-Holland.

Card, David. 1999. The Causal Effect of Education on Earnings. *Chap. 30, pages 1801–1863 of:* Ashenfelter, Orley, & Card, David (eds), *Handbook of Labor Economics*, vol. 3. Elsevier Science B.V.

Kane, Thomas J., Rouse, Cecilia Elena, & Staiger, Douglas. 1999. *Estimating Returns to Schooling when Schooling is Misreported.* Princeton University Industrial Relation Sections WP 419.

Lewbel, Arthur. 2007. Estimation of Average Treatment Effects with Misclassification. *Econometrica*, **75**(2), 537–551.

Mahajan, Aprajit. 2006. Identification and Estimation of Regression Models with Misclassification. *Econometrica*, **74**(3), 631–665.

Mathiowetz, Nancy A. 1992. Errors in Reports of Occupation. *Public Opinion Quarterly*, **56**, 352–355.

Mellow, Wesley, & Sider, Hal. 1983. Accuracy of Response in Labor Market Surveys: Evidence and Implications. *Journal of Labor Economics*, **1**, 331–344.

A Data Appendix

A.1 The 2002 Structure of Earnings Survey

A Structure of Earnings Survey (SES; french: *Enquête Structure des Salaires*) is conducted since 1972 at irregular intervals by the French National Institute for Statistics and Economic Studies (INSEE) as part of a programme initiated in 1966 by the European Statistical Office (EUROSTAT) which aims at producing harmonised labour cost statistics for all EU countries. Recent years of survey were 1992, 1994, 2002 and 2005-2006.

At the end of the Nineties, the SES underwent a process of reform at the European and national level. The most recent waves obey to Council Regulation 530/1999 and Commission Regulation 1916/2000, which specify the items to be collected and the field to be covered, but not the method. Countries are free to use tailor-made questionnaires, questionnaires of existing surveys, administrative records or a combination of these as their source for providing data to EUROSTAT, as long as the information is of "acceptable quality".

The French SES 2002 uses taylor-made questionnaires and covers firms with at least 10 employees inside NACE sections C to K[11] in mainland France. Sampling occurs at two levels: first, about 20'000 production units (establishments) belonging to firms with at least 10 employees are sampled according to their size, sector and geographical location. In a second stage, individuals employed at these units are sampled (10 on average and 26 at most in each unit), according to their position (executive or not). Executives are over-sampled relative to non-executives; the universe is known from the *Déclarations Annuelles de Données Sociales* (DADS) 2001, an administrative register covering all employees of the private sector which lists all ongoing contracts as of the last Friday in December 2001.

The French SES 2002 is both a business and a household survey. Two questionnaires are sent separately by post to both the establishment and the employee[12]. For firms, this is a mandatory survey, and after some time non-responding firms are phoned, and eventually visited, by an INSEE surveyor. The firm questionnaire is composed by a two-side paper-sheet about the establishment, and an additional two-side paper sheet for each sampled employee (the survey administrators estimate that it takes 10 minutes for a firm to fill out the questions for one employee). The questionnaire submitted to the worker is a two-side paper sheet with 16 multiple-choice or very short open-ended questions.[13]

[11] NACE codes: C - Mining and quarrying, D - Manufacturing, E - Electricity, gas, water supply, F- Construction, G - Trade, H - Hotels and Restaurants, I- Transport and Communication, J - Financial intermediation, K - Real estate, renting, business activities.

[12] In fact, not all sampled employees receive the questionnaire: those for whom the address is missing in the DADS data base, those on whom the employer did not fill out a questionnaire, and those who changed address between the end of 2001 and fall 2003 are excluded from the sample. The first deadline for establishments is in June; the extended implicit deadline after recall is in November. Worker questionnaires are only sent if the employing establishment responded to the survey, and their collection lasts until December.

[13] For additional information on the survey, see the description in Aeberhardt & Pouget [2007].

B Tables & Results

B.1 Women

truth	worker (1)	worker (2)	worker (3)	worker (4)	employer (1)	employer (2)	employer (3)	employer (4)
(1)	**0.977****	0.023**	-0.001	0.001	**0.925****	0.054**	0.017**	0.004**
	(0.006)	*(0.006)*	*(0.002)*	*(0.001)*	*(0.006)*	*(0.005)*	*(0.003)*	*(0.001)*
(2)	0.115**	**0.848****	0.020	0.017*	0.211**	**0.735****	0.047**	0.008†
	(0.014)	*(0.021)*	*(0.017)*	*(0.007)*	*(0.017)*	*(0.019)*	*(0.011)*	*(0.004)*
(3)	0.016**	0.095**	**0.869****	0.020	0.040**	0.181**	**0.734****	0.045**
	(0.004)	*(0.013)*	*(0.020)*	*(0.015)*	*(0.007)*	*(0.015)*	*(0.017)*	*(0.007)*
(4)	-0.000	0.005	0.020**	**0.975****	0.005*	0.040**	0.134**	**0.821****
	(0.000)	*(0.004)*	*(0.006)*	*(0.007)*	*(0.002)*	*(0.006)*	*(0.012)*	*(0.014)*

	(1)	(2)	(3)	(4)
$\mathrm{P}(\widehat{s_i^* = m})$	0.413**	0.180**	0.197**	**0.211****
std. err.	*(0.007)*	*(0.007)*	*(0.007)*	*(0.006)*
$\mathrm{E}(\widehat{\hat{z}_i \mid s_i^* = m})$	17.04**	19.27**	21.32**	23.83**
std. err.	*(0.03)*	*(0.06)*	*(0.06)*	*(0.06)*

Unconditional Error Probabilities

	worker		firm	
	est.	std.err.	est.	std.err.
precision	**0.932****	*(0.006)*	**0.831****	*(0.006)*
over-rep	**0.020****	*(0.005)*	**0.049****	*(0.004)*
under-rep	**0.048****	*(0.004)*	**0.119****	*(0.006)*

B.2 Men

truth	worker (1)	worker (2)	worker (3)	worker (4)	employer (1)	employer (2)	employer (3)	employer (4)
(1)	**0.969****	0.025**	0.005**	0.001†	**0.928****	0.055**	0.013**	0.005**
	(0.003)	*(0.003)*	*(0.001)*	*(0.001)*	*(0.004)*	*(0.003)*	*(0.002)*	*(0.001)*
(2)	0.145**	**0.829****	0.012	0.014*	0.215**	**0.646****	0.093**	0.045**
	(0.012)	*(0.021)*	*(0.019)*	*(0.007)*	*(0.012)*	*(0.014)*	*(0.012)*	*(0.005)*
(3)	0.007†	0.055**	**0.886****	0.052**	0.041**	0.159**	**0.711****	0.090**
	(0.004)	*(0.010)*	*(0.016)*	*(0.012)*	*(0.005)*	*(0.011)*	*(0.014)*	*(0.007)*
(4)	0.002†	-0.000	0.017**	**0.982****	0.007**	0.025**	0.087**	**0.881****
	(0.001)	*(0.001)*	*(0.004)*	*(0.004)*	*(0.002)*	*(0.003)*	*(0.007)*	*(0.008)*

	(1)	(2)	(3)	(4)
$\mathrm{P}(\widehat{s_i^* = m})$	0.429**	0.143**	0.176**	**0.252****
std. err.	*(0.005)*	*(0.005)*	*(0.005)*	*(0.005)*
$\mathrm{E}(\widehat{\hat{z}_i \mid s_i^* = m})$	16.76**	19.78**	21.57**	24.18**
std. err.	*(0.03)*	*(0.06)*	*(0.04)*	*(0.04)*

Unconditional Error Probabilities

	worker est.	worker std.err.	firm est.	firm std.err.
precision	**0.938****	*(0.004)*	**0.838****	*(0.004)*
over-rep	**0.026****	*(0.004)*	**0.066****	*(0.003)*
under-rep	**0.036****	*(0.003)*	**0.096****	*(0.004)*

B.3 Small

truth	worker (1)	worker (2)	worker (3)	worker (4)	employer (1)	employer (2)	employer (3)	employer (4)
(1)	**0.982****	0.015**	0.003*	0.000	**0.927****	0.054**	0.013**	0.005**
	(0.004)	*(0.004)*	*(0.001)*	*(0.001)*	*(0.004)*	*(0.004)*	*(0.002)*	*(0.001)*
(2)	0.147**	**0.831****	0.006	0.016*	0.227**	**0.674****	0.079**	0.020**
	(0.013)	*(0.020)*	*(0.017)*	*(0.006)*	*(0.013)*	*(0.015)*	*(0.011)*	*(0.004)*
(3)	0.015**	0.090**	**0.862****	0.033*	0.041**	0.180**	**0.710****	0.069**
	(0.004)	*(0.012)*	*(0.018)*	*(0.014)*	*(0.005)*	*(0.013)*	*(0.015)*	*(0.008)*
(4)	0.003*	0.002	0.030**	**0.964****	0.004*	0.029**	0.126**	**0.841****
	(0.001)	*(0.002)*	*(0.006)*	*(0.006)*	*(0.002)*	*(0.004)*	*(0.010)*	*(0.011)*

	(1)	(2)	(3)	(4)
$\mathrm{P}(\widehat{s_i^* = m})$	0.464**	0.161**	0.173**	**0.201****
std. err.	*(0.006)*	*(0.006)*	*(0.006)*	*(0.005)*
$\mathrm{E}(\widehat{\hat{z}_i \mid s_i^* = m})$	16.79**	19.54**	21.56**	24.12**
std. err.	*(0.03)*	*(0.05)*	*(0.05)*	*(0.05)*

Unconditional Error Probabilities

	worker		firm	
	est.	std.err.	est.	std.err.
precision	**0.933****	*(0.005)*	**0.831****	*(0.005)*
over-rep	**0.018****	*(0.004)*	**0.062****	*(0.003)*
under-rep	**0.049****	*(0.003)*	**0.107****	*(0.004)*

B.4 Big

truth	worker (1)	worker (2)	worker (3)	worker (4)	employer (1)	employer (2)	employer (3)	employer (4)
(1)	**0.958****	0.037**	0.003	0.003*	**0.928****	0.053**	0.015**	0.004**
	(0.005)	*(0.005)*	*(0.002)*	*(0.001)*	*(0.005)*	*(0.004)*	*(0.002)*	*(0.001)*
(2)	0.119**	**0.834****	0.032	0.015†	0.199**	**0.687****	0.070**	0.043**
	(0.012)	*(0.023)*	*(0.021)*	*(0.008)*	*(0.014)*	*(0.017)*	*(0.013)*	*(0.007)*
(3)	0.006	0.050**	**0.891****	0.053**	0.040**	0.150**	**0.732****	0.079**
	(0.004)	*(0.010)*	*(0.017)*	*(0.013)*	*(0.006)*	*(0.013)*	*(0.016)*	*(0.008)*
(4)	-0.000	0.000	0.009*	**0.991**†	0.008**	0.029**	0.077**	**0.887****
	(0.001)	*(0.002)*	*(0.004)*	*(0.005)*	*(0.002)*	*(0.004)*	*(0.007)*	*(0.009)*

	(1)	(2)	(3)	(4)
$\mathrm{P}(\widehat{s_i^* = m})$	0.377**	0.146**	0.192**	**0.284****
std. err.	*(0.006)*	*(0.006)*	*(0.006)*	*(0.006)*
$\mathrm{E}(\widehat{\hat{z}_i \mid s_i^* = m})$	16.93**	19.63**	21.42**	24.07**
std. err.	*(0.03)*	*(0.07)*	*(0.05)*	*(0.04)*

Unconditional Error Probabilities

	worker		firm	
	est.	std.err.	est.	std.err.
precision	**0.936****	*(0.005)*	**0.843****	*(0.005)*
over-rep	**0.033****	*(0.004)*	**0.059****	*(0.003)*
under-rep	**0.031****	*(0.003)*	**0.098****	*(0.005)*

B.5 Young

truth	worker (1)	worker (2)	worker (3)	worker (4)	employer (1)	employer (2)	employer (3)	employer (4)
(1)	**0.990**†	0.007†	0.001	0.002	**0.931****	0.055**	0.011**	0.003**
	(0.005)	*(0.004)*	*(0.002)*	*(0.001)*	*(0.005)*	*(0.004)*	*(0.002)*	*(0.001)*
(2)	0.108**	**0.883****	-0.000	0.009	0.204**	**0.703****	0.072**	0.021**
	(0.012)	*(0.023)*	*(0.020)*	*(0.007)*	*(0.014)*	*(0.015)*	*(0.011)*	*(0.004)*
(3)	0.009**	0.065**	**0.885****	0.042**	0.033**	0.151**	**0.760****	0.056**
	(0.003)	*(0.009)*	*(0.014)*	*(0.011)*	*(0.004)*	*(0.011)*	*(0.012)*	*(0.005)*
(4)	0.000	0.003†	0.020**	**0.977****	0.006**	0.026**	0.110**	**0.858****
	(0.001)	*(0.001)*	*(0.003)*	*(0.004)*	*(0.002)*	*(0.003)*	*(0.007)*	*(0.009)*

	(1)	(2)	(3)	(4)
$\mathrm{P}(\widehat{s_i^* = m})$	0.342**	0.139**	0.224**	**0.295****
std. err.	*(0.005)*	*(0.005)*	*(0.006)*	*(0.005)*
$\mathrm{E}(\widehat{\hat{z}_i \mid s_i^* = m})$	17.44**	19.72**	21.57**	24.16**
std. err.	*(0.03)*	*(0.05)*	*(0.04)*	*(0.04)*

Unconditional Error Probabilities

	worker est.	worker std.err.	firm est.	firm std.err.
precision	**0.948****	*(0.004)*	**0.839****	*(0.005)*
over-rep	**0.014****	*(0.004)*	**0.049****	*(0.003)*
under-rep	**0.038****	*(0.003)*	**0.112****	*(0.004)*

B.6 Old

truth	worker (1)	worker (2)	worker (3)	worker (4)	employer (1)	employer (2)	employer (3)	employer (4)
(1)	**0.963****	0.031**	0.004**	0.001	**0.926****	0.053**	0.015**	0.006**
	(0.004)	*(0.004)*	*(0.001)*	*(0.001)*	*(0.004)*	*(0.004)*	*(0.002)*	*(0.001)*
(2)	0.164**	**0.792****	0.024	0.020**	0.241**	**0.647****	0.074**	0.037**
	(0.014)	*(0.021)*	*(0.018)*	*(0.007)*	*(0.014)*	*(0.016)*	*(0.012)*	*(0.006)*
(3)	0.018*	0.086**	**0.848****	0.048**	0.059**	0.201**	**0.635****	0.105**
	(0.007)	*(0.017)*	*(0.024)*	*(0.017)*	*(0.009)*	*(0.017)*	*(0.020)*	*(0.012)*
(4)	0.003†	-0.001	0.019**	**0.979****	0.007**	0.035**	0.074**	**0.884****
	(0.002)	*(0.003)*	*(0.007)*	*(0.008)*	*(0.002)*	*(0.005)*	*(0.009)*	*(0.011)*

	(1)	(2)	(3)	(4)
$\widehat{\mathrm{P}(s_i^* = m)}$	0.513**	0.172**	0.138**	**0.177****
std. err.	*(0.006)*	*(0.007)*	*(0.006)*	*(0.005)*
$\widehat{\mathrm{E}(\hat{z}_i \mid s_i^* = m)}$	16.37**	19.43**	21.31**	23.97**
std. err.	*(0.03)*	*(0.07)*	*(0.07)*	*(0.06)*

Unconditional Error Probabilities

	worker		firm	
	est.	std.err.	est.	std.err.
precision	**0.921****	*(0.005)*	**0.830****	*(0.005)*
over-rep	**0.033****	*(0.004)*	**0.072****	*(0.004)*
under-rep	**0.046****	*(0.004)*	**0.098****	*(0.005)*