Diploma Free Riders. Errors in Assessing Employees' Education, and Their Consequences.

Francesco Avvisati*

This version: 6th June 2008

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

This paper documents large misclassification errors for employer reports of workers' educational attainment in a French employer-employee survey data set. I show that employers' errors have consequences on workers' wages. Results indicate that those who falsely appear as having a given diploma experience significant wage gains. Causal and non-causal interpretations of these estimates are discussed.
Keywords: Misclassification, Returns to education

Sessione/Sottosessione prescelta: sessione tematica 1 (Lavoro, produttività e crescita), oppure sottosessione libera 2 (Macro e microeconomia delle retribuzioni)

*PhD candidate, Paris School of Economics, 48 Boulevard Jourdan, 75014 Paris. francesco.avvisati@ens.fr

Contents

Educational reports obtained for the same person from two separate sources display a surprisingly high number of discrepancies. This observation has been documented from twin data (self-reported vs. twin-reported educational attainment), recall data at different dates, self-reported vs. administrative data. Using the French Structure of Earnings Survey 2002, we document this fact for self- vs. employer-reported educational attainment. Differences between the two responses for the same employer-employee pair represent measurement error on the part of at least one side.

In this paper I ask whether these errors, and in particular employers' errors in reporting educational attainment, affect wages. Before turning to answering this question, I introduce the data and provide some insights as to the determinants of errors in reporting education.

1 Data

1.1 The 2002 Structure of Earnings Survey

A Structure of Earnings Survey (SES; french: *Enquête Structure des Salaires*) is carried out 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[1] 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; its objective is to study the effect of worker and firm characteristics on 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

[1] 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.

description of the worker is completed by a questionnaire directly submitted to her, asking details about the career and the family. All data refer to 2002, but the survey was carried out between April and December 2003. The questionnaires are sent separately by post to both the establishment and the employee[2]. 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.[3]

In 2002 both the firm and the worker questionnaire asked questions about the employee's educational attainment and citizenship. 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[4]; in all cases where the response to this item by the firm was missing, the survey administrator used the worker's response in building the data for EUROSTAT. In order to do so, the question asked to employers and employees about educational attainment needs to be the same. There is actually a minor difference: employers are asked the highest (known) diploma of the worker, whilst the question for workers is more ambiguous, asking generically for earned diplomas. Nevertheless, both are multiple choice questions, with the same 8 possible answers given, and a note specifies that the list provided constitutes a ranking, leaving no possible ambiguity about what the "highest" diploma is; as a consequence, we can compare the highest diploma from the worker questionnaire with her employer's answer.

1.2 Data selection

Some loose consistency and relevance requirements are imposed on data.

- drop employer questionnaires with more than one diploma ticked as 'highest'. Keep only full-time and full-year employees of the establishment.
- 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 education.
- 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 €).

All empirical analyses are based on prime-age workers (aged between 30 and 59) only.

[2]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.

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

[4]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`

Table 1: Distribution of diploma reports

Firm

	Frequency	Percent	Proportion of women
1	3132	9.5	32.22
3	1515	4.59	38.81
4	6413	19.45	26.70
5	2583	7.83	33.41
6	1587	4.81	41.97
7	4358	13.22	33.13
8	5760	17.47	24.64
m	7628	23.13	31.46

Worker

	Frequency	Percent	Proportion of women	Proportion missing in firm
1	4163	12.62	30.39	32.79
3	1988	6.03	39.08	28.82
4	9103	27.6	27.46	25.62
5	3256	9.87	32.25	21.81
6	1722	5.22	43.32	25.26
7	5212	15.81	33.08	18.57
8	7411	22.47	27.00	16.34
m	121	0.37	31.40	38.10

Table 2: Distribution of joint diploma reports

worker / firm	1	3	4	5	6	7	8
1	6.67	1.02	3.59	0.44	0.17	0.26	0.09
3	0.90	2.17	1.84	0.55	0.30	0.12	0.08
4	2.89	1.15	18.40	1.91	0.38	0.44	0.11
5	0.26	0.55	1.73	4.62	1.12	1.54	0.38
6	0.17	0.44	0.51	1.25	2.01	1.30	0.59
7	0.12	0.21	0.54	0.99	0.91	11.58	2.89
8	0.06	0.06	0.18	0.32	0.20	1.57	20.40

agreement rate: 65.86
worker > employer: 19.13
worker < employer: 15.01

1.3 Descriptive Statistics

The number of missing answers to the question about the highest diploma is bigger for the firm questionnaire than for the worker's, where missing values are very rare.

When plotted against each other, the firm and worker answer reveal a high correlation, but also a non-negligible number of inconsistent reports: their proportion reaches 34%, excluding missing answers from our sample.

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 the empirical model will provide some evidence on that point.

Wages increase with both employer and employee reports (see table 3): holding one report constant, the second correlates positively with wages. Moreover, wages seem to increase more steeply with employer reports than with workers'

Table 3: mean wages by joint diploma reports

firm	worker	1	3	4	5	6	7	8
1	logwage	2.46	2.58	2.55	2.69	2.66	2.94	3.07
	se	0.01	0.03	0.01	0.04	0.06	0.06	0.14
3	logwage	2.49	2.76	2.69	2.88	2.90	3.05	3.37
	se	0.02	0.02	0.02	0.03	0.05	0.08	0.10
4	logwage	2.50	2.64	2.65	2.81	2.69	2.83	3.28
	se	0.01	0.02	0.01	0.02	0.04	0.04	0.09
5	logwage	2.92	2.97	2.84	2.84	2.88	2.96	3.20
	se	0.06	0.04	0.02	0.01	0.02	0.02	0.05
6	logwage	3.09	3.05	2.89	2.88	2.97	2.97	3.23
	se	0.08	0.05	0.04	0.02	0.02	0.02	0.04
7	logwage	3.06	3.10	2.96	2.97	3.05	3.00	3.15
	se	0.10	0.06	0.04	0.03	0.03	0.01	0.02
8	logwage	3.42	3.31	3.23	3.38	3.46	3.30	3.44
	se	0.11	0.10	0.07	0.05	0.06	0.02	0.01

Table 4: median age at school exit by joint diploma reports

firm \ worker	1	3	4	5	6	7	8
1	16	17	17	19	19	21	23
3	16	18	18	19	19	20	20
4	16	17	18	18	19	21	21
5	17	18	18	19	20	21	23
6	18	18	18	19	20	21	23
7	17	19	18	20	21	21	23
8	19	18	18	20	20	22	24

reports. If measurement errors do not themselves influence wages, this would indicate that a high employer report is a better indicator for high educational attainment than a high worker report. Contrary to this intuition, age at school exit[5] rises systematically and in a predictible way with worker's report, but not, or much less, with employers' answers.

This motivates our question whether workers with the same degree, but who are perceived as having different degrees by their employers, earn different wages. In particular, we will try to quantify by how much the wages of workers with no university degree who falsely appear as being university graduates, exceed wages of non university graduates who are recognised as such.

2 Errors: Origins and Magnitude

The strength of association of two partially independent reports of an ordinal variable such as the highest completed level of education can be considered as an indicator for the joint precision of the two reports.

We explore whether errors in reporting education are related to the relevance of this information, to discrimination or originate in memory flaws by dividing the sample along different observable dimensions and studying how measures

[5]Age at school exit is derived from a question on the worker's questionnaire, where people are asked to report the age at which they stopped attending regularly school or university.

Table 5: Sample Summary Statistics

global statistics

	N	
non-response rate (worker)	*32976*	.004
non-response rate (firm)	*32976*	.231
concordance rate	*25261*	.659
simple kappa statistic	*25261*	.582
Kendall's tau	*25261*	.770

of agreement vary between groups. I interpret a higher agreement as indicating lower error rates; this is granted, as a first approximation, as long as the two reports are independent, or, if measurement errors are correlated, if this correlation is not itself a function of other covariates.

Total sample agreement statistics are given in table 5.

Table 6, where agreement rates and κ coefficients are presented for different sub-groups, reveals some regular patterns. First, the similar figures on agreement rates by gender 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. Both figures can be motivated with the varying importance of diplomas according to occupation and size of the production unit. 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.

Finally, as shown in table 6 and, in further details, in table 7, agreement statistics decline with age, but not (or much less) with seniority, holding age constant. This might result from memory flaws on the worker's side; the remarkable stability as seniority increases tends to suggest that employer's reports are not subject to changes over time, from the moment of hiring. This is consistent with the intuition that employer answers reflect codified employee records[6], while worker answers are produced out of memory.

2.1 Related Assessments of Measurement Error in Education

The research cited by Card [1999, p. 1816] finds that the reliability, or the signal-to-total-variance ratio of self-reported years of schooling is about 90%, with a similar reliability for administrative measures of schooling. Consistent with this evidence, when categorical measures of educational achievement are used, misclassification errors have been shown to be widespread both in self-reports and in administrative data [Kane *et al.* , 1999; Bound *et al.* , 2001; Battistin & Sianesi, 2006].

[6]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 6: Agreement between reports of educational attainment

Measures of agreement of reports of educational attainment
Concordance rate and simple kappa statistic

establishment characteristics	N	κ	cc.
industry			
C: mining and quarrying	*190*		.52.63
D: manufacturing	*8985*		.63.70
E: electricity, gas, water supply	*901*		.75.81
F: construction	*1132*		.56.66
G: wholesale and retail trade	*3405*		.51.60
H: hotels and restaurants	*216*		.42.54
I: transport, storage	*1863*		.52.61
J: financial intermediation	*3369*		.49.59
K: real estate, business activities	*5200*		.58.67
establishment size			
size 0-9	*365*		.50.59
size 10-49	*5934*		.53.62
size 50-199	*7242*		.56.64
size 200-499	*5185*		.58.66
size 500-1999	*5952*		.64.71
size 2000-	*583*		.65.72
firm size			
size 10-49	*4418*		.55.64
size 50-199	*4886*		.57.65
size 200-499	*3653*		.59.67
size 500-1999	*5453*		.60.67
size 2000-	*6851*		.59.66
collective agreements			
no	*1467*		.61.68
yes	*23556*		.58.66

worker characteristics	N	κ	cc.
gender			
men	*17585*	.59	.67
women	*7676*	.57	.64
age			
age 30-34	*4179*	.62	.71
age 35-39	*4629*	.61	.69
age 40-44	*4694*	.59	.66
age 45-49	*4437*	.57	.65
age 50-54	*4662*	.54	.62
age 55-59	*2660*	.50	.59
nationality			
french, born french	*22486*	.58	.66
french, born foreign	*1089*	.54	.62
foreign	*610*	.54	.64
worker-firm match characteristics			
occupation			
executive employees	*10723*	.58	.71
clerical workers	*6388*	.51	.61
skilled blue-collar workers	*2614*	.47	.58
unskilled blue-collar workers	*5182*	.45	.66
seniority in firm			
0-4 years	*5309*	.57	.67
5-9 years	*3995*	.59	.68
10-14 years	*4696*	.60	.68
15-19 years	*2823*	.60	.67
20-24 years	*2971*	.59	.67
25-29 years	*2588*	.56	.64
30-34 years	*2025*	.48	.58

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 refers to number of workers in each category. In some cases figures do not sum up to the total sample size because of omitted categories and/or missing or incomplete reports.

Table 7: Agreement between reports of educational attainment

Sample size

	age					
seniority	30-34	35-39	40-44	45-49	50-54	55-59
0-4	1935	1287	860	603	440	184
5-9	1450	899	584	443	414	205
10-14	749	1551	940	619	524	313
15-19		727	951	486	395	219
20-24			1135	875	527	269
25-29				1063	966	336
30-34					1082	595
35-39						406

Concordance rates

	age					
seniority	30-34	35-39	40-44	45-49	50-54	55-59
0-4	0.71	0.69	0.67	0.64	0.59	0.53
5-9	0.73	0.68	0.67	0.60	0.60	0.58
10-14	0.67	0.72	0.67	0.67	0.64	0.60
15-19		0.67	0.68	0.67	0.64	0.66
20-24			0.66	0.70	0.70	0.61
25-29				0.63	0.66	0.65
30-34					0.58	0.59
35-39						0.53

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 refers to number of workers in each category. In some cases figures do not sum up to the total sample size because of omitted categories and/or missing or incomplete reports.

The two reports of educational achievement used in Kane *et al.* [1999], as reported in Bound *et al.* [2001, table 10], display an agreement rate of .891 and a kappa of .830. Their data come from the National Longitudinal Study of the High School Class of 1972 and distinguish three levels (no college, some college, BA+).

Data reported in Battistin & Sianesi [2006, 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 (0.902) and kappa (0.792).

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 less 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] find low agreement rates for employer and employee reports of occupation (0.576 at the 3-digits level), which Mathiowetz [1992] attributes mostly to the difficulty of coding survey answers. Moving to broad occupational classifications or to direct comparison of descriptions, rather than codes, there is still a significant percent of disagreement (13% in Mathiowetz [1992], for direct comparison of descriptions, and 19% in Mellow & Sider [1983], for comparisons of the first digit only), which is of the same magnitude of that found in the present study on education.

The low agreement rates in comparison with the cited studies suggest employer and employee reports of educational attainment in the SES 2002 can be viewed as independent, or almost independent of each other.

3 Related Literature

Due to misclassification error, some workers are considered by their employers more educated than they in truth are. This paper asks whether these workers get higher wages on average than what their true educational attainment would predict.

This is a very similar question to that asked by Hu & Lewbel [2008], namely whether a person who falsely claims college experience has on average higher wages than a person who tells the truth about not having college. Our identification strategy is indeed based on their paper. Hu & Lewbel [2008] find positive wage differentials associated with overstatements of own education, and propose to interpret these differentials as a return to lying. We suggest that a lie can give a return only if it is effective, that is if the employer is convinced by the lie. If their results can be interpreted as returns to lying, we therefore expect returns associated with overstatements of education by the employer to be higher.

It is a meaningful question for two reasons. First, a positive answer would constitutes evidence in favour of the existence of sheepskin effects in education. By sheepskin effects we mean that credentials matter more than schooling per se: whether this credential is incorporated in a piece of paper or remains virtual is secondary to our analysis.

More generally our analysis is connected to the signaling literature. The possibility of free-riding behaviour on diplomas (profit from the positive signal of a diploma without having to pay the cost of schooling) opens the field for a pure signaling explanation of returns to schooling. It is however not a direct test of the signaling theory, because a positive answer would not tell anything as to the origin of the difference in productivity corresponding to different degrees.

4 Consequences: Identification

Do expected wages vary, for workers with similar education, with their employers' belief about their educational attainment? The general question examined by this paper restates, in the context of wage returns to education, the hypothesis of non-differential measurement error in employer reports of educational attainment: with non-differential measurement error, wages only depend on true education, not on reports of education.

Formally, let s_i^* be worker i's true educational attainment, and s_i^1 be the firm's belief about her educational attainment. Wages are indicated by w_i, and observed covariates by x_i. Under non-differential measurement errors, the error-ridden measurement does not provide information about wages beyond the information contained in the true value:

$$f_{w_i|s_i^*,s_i^1,x_i} = f_{w_i|s_i^*,x_i}$$

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

Assuming $E(w_i|s_i^*, s_i^1, \mathbf{x}_i)$ exists, I define a model for the conditional mean as

$$w_i = E(w_i|s_i^*, s_i^1, \mathbf{x}_i) + \eta_i = f(s_i^*, s_i^1, x_i) + \eta_i$$

Hu & Lewbel [2008] present a partial identification result for f when s_i^* and s_i^1 are binary indicators. When s_i^* and s_i^1 are binary, without any parametric assumption, the conditional mean can be separated into two components.

$$\begin{aligned} w_i &= E(w_i|s_i^* = 0, s_i^1, \mathbf{x}_i) + s_i^* E(w_i|s_i^* = 1, s_i^1, \mathbf{x}_i) + \eta_i \\ &= g(s_i^1, \mathbf{x}_i) + s_i^* h(s_i^1, \mathbf{x}_i) + \eta_i \end{aligned} \tag{1}$$

If s_i^* is an indicator for holding a degree, $g(s_i^1, \mathbf{x}_i)$ represents how wages vary with employer beliefs for workers who do not hold that degree: in particular, $g(1, \mathbf{x}_i) - g(0, \mathbf{x}_i)$ is the average difference in wages between those workers with characteristics $\mathbf{x}_i$ who are falsely considered degree holders and those who are correctly recognised as not holding the degree.

Under the hypothesis of non-differential measurement error, $g(1, \mathbf{x}_i) - g(0, \mathbf{x}_i)$ is equal to 0.

The function $h(s_i^1, \mathbf{x}_i)$ represents how wages vary with beliefs for degree holders ($s_i^* = 1$). The loss in mean wages when the firm falsely believes the worker does not possess the degree can be computed as $h(1, \mathbf{x}_i) - h(0, \mathbf{x}_i)$ and equals 0 when measurement error is non-differential.

Following Hu & Lewbel [2008], I present identification results for $g(1, \mathbf{x}_i) - g(0, \mathbf{x}_i)$.

4.1 Identification of g when the truth is known

When s_i^* is known, identification of g and h is straightforward. If workers correctly report their own education to the survey (self-reports do not contain measurement error), then s_i^* is equal to self-reported education. Given this hypothesis, if measurement error by the firm were non-differential, wages should not differ according to the firm's report for workers who self-report the same level of education.

4.2 Nonparametric identification of g when the truth is unobserved

When the second available measurement of s^* is imperfect, as is plausible with self-reported education due to deliberate or involuntary errors in answering the survey, identification of g requires this second measurement to verify an exclusion restriction.

Assumption 1. *A second measurement of educational attainment s_i^2 exists and verifies*

$$E(\eta_i s_i^2 | s_i^1, \boldsymbol{x}_i) = 0 \tag{2}$$

$$E(s_i^2 | s_i^1 = 1, \boldsymbol{x}_i) \neq E(s_i^2 | \boldsymbol{x}_i) \tag{3}$$

In the context of this paper, I construct binary reports of educational attainment as indicators of an university degree (levels 7 and 8 in the survey classification). Assumption 1 requires two properties to be verified by selfreports

of educational attainment. Equation 2 is implied by its unconditional version, which states that s_i^2 (self reported status with respect to college graduation) is redundant in the model for the conditional mean: interpreting self-reports as an instrument, this is a typical exclusion restriction). It means that errors in self-reporting education do not contain information on wages, once the firm's belief has been controlled for. Finally, equation 3 states that in general, s_i^1 is related to s_i^2: this is a standard condition for s_i^2 acting as an instrument for s_i^1. Note that this statement is not conditional on s_i^*, so that it holds even with s_i^1 and s_i^2 being independent measurements of s_i^*.

4.2.1 No underreporting

A special case that verifies assumption 1 deserves a separate treatment: when under-reporting of degrees is ruled out by assumption. This might seem plausible, given social norms that consider education a value, meaning that workers never understate their education in surveys, but eventually overstate their achievement.

If $(s_i^* = 1)$ implies $(s_i^2 = 1)$, then, given the binary nature of these variables, $(s_i^2 = 0)$ implies $(s_i^* = 0)$. People who self-report not having a university degree truly do not have one. If as we assumed, s_i^2 does not influence wages (conditional on s_i^1, s_i^* and $\mathbf{x}_i$), then $g(s_i^1, \mathbf{x}_i)$ is directly identified:

$$\begin{aligned} E(w_i|s_i^2 = 0, s_i^1, \mathbf{x}_i) &= E(w_i|s_i^2 = 0, s_i^* = 0, s_i^1, \mathbf{x}_i) \\ &= E(w_i|s_i^* = 0, s_i^1, \mathbf{x}_i) = g(s_i^1, \mathbf{x}_i) \end{aligned}$$

and

$$g(1, \mathbf{x}_i) - g(0, \mathbf{x}_i) = E(w_i|s_i^2 = 0, s_i^1 = 1, \mathbf{x}_i) - E(w_i|s_i^2 = 0, s_i^1 = 0, \mathbf{x}_i)$$

.

4.2.2 Underreporting

More generally, given assumption 1, $g(s_i^1, \mathbf{x}_i)$ is a function of $E(w_i|s_i^2 = 0, s_i^1, \mathbf{x}_i)$, $E(w_i|s_i^2 = 1, s_i^1, \mathbf{x}_i)$, $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$ and $E(s_i^2|s_i^1, \mathbf{x}_i)$. The only term for which there is no observable equivalent is $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$, the (conditional) probability for a graduate worker i not under-reporting his graduation status.

Theorem 1.

$$\begin{aligned} g(s_i^1, \boldsymbol{x}_i) =& E(w_i|s_i^2 = 0, s_i^1, \boldsymbol{x}_i) - \dots \\ & \dots - \left[\frac{[1 - E(s_i^2|s_i^* = 1, s_i^1, \boldsymbol{x}_i)]E(s_i^2|s_i^1, \boldsymbol{x}_i)}{E(s_i^2|s_i^* = 1, s_i^1, \boldsymbol{x}_i) - E(s_i^2|s_i^1, \boldsymbol{x}_i)}\right] (E(w_i|s_i^2 = 1, s_i^1, \boldsymbol{x}_i) - E(w_i|s_i^2 = 0, s_i^1, \boldsymbol{x}_i)) \end{aligned}$$

Proof of Theorem 1. Observe that

$$\begin{aligned} E(w_i|s_i^2 = 0, s_i^1, \mathbf{x}_i) &= g(s_i^1, \mathbf{x}_i) + h(s_i^1, \mathbf{x}_i)E(s_i^*|s_i^2 = 0, s_i^1, \mathbf{x}_i) \\ E(w_i|s_i^2 = 1, s_i^1, \mathbf{x}_i) &= g(s_i^1, \mathbf{x}_i) + h(s_i^1, \mathbf{x}_i)E(s_i^*|s_i^2 = 1, s_i^1, \mathbf{x}_i) \end{aligned}$$

Apply Bayes' rule to $E(s_i^*|s_i^2 = 1, s_i^1, \mathbf{x}_i) = Pr(s_i^* = 1|s_i^2 = 1, s_i^1, \mathbf{x}_i)$.

$$Pr(s_i^* = 1|s_i^2 = 1, s_i^1, \mathbf{x}_i) = \frac{Pr(s_i^2 = 1|s_i^* = 1, s_i^1, \mathbf{x}_i)}{Pr(s_i^2 = 1|s_i^1, \mathbf{x}_i)} Pr(s_i^* = 1|s_i^1, \mathbf{x}_i)$$

Let $k(s_i^*, s_i^1, \mathbf{x}_i) = Pr(s_i^* = 1|s_i^1, \mathbf{x}_i)h(s_i^1, \mathbf{x}_i)$; The initial system can then be written as

$$E(w_i|s_i^2 = 0, s_i^1, \mathbf{x}_i) = g(s_i^1, \mathbf{x}_i) + \frac{1 - E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)}{1 - E(s_i^2|s_i^1, \mathbf{x}_i)} k(s_i^*, s_i^1, \mathbf{x}_i)$$

$$E(w_i|s_i^2 = 1, s_i^1, \mathbf{x}_i) = g(s_i^1, \mathbf{x}_i) + \frac{E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)}{E(s_i^2|s_i^1, \mathbf{x}_i)} k(s_i^*, s_i^1, \mathbf{x}_i)$$

Differencing the above system yields the following expression for $k(s_i^*, s_i^1, \mathbf{x}_i)$:

$$k(s_i^*, s_i^1, \mathbf{x}_i) = \left[\frac{(1 - E(s_i^2|s_i^1, \mathbf{x}_i))E(s_i^2|s_i^1, \mathbf{x}_i)}{E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i) - E(s_i^2|s_i^1, \mathbf{x}_i)} \right] [E(w_i|s_i^2 = 1, s_i^1, \mathbf{x}_i) - E(w_i|s_i^2 = 0, s_i^1, \mathbf{x}_i)]$$

and thus the expression in theorem 1 for $g(s_i^1, \mathbf{x}_i)$. □

Under Theorem 1, $g(s_i^1, \mathbf{x}_i)$ is directly identified, as long as $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$ (the proportion of correct self-reports for true diploma holders) is identified and $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i) \neq E(s_i^2|s_i^1, \mathbf{x}_i)$ to avoid division by zero. This identification result is essentially the same as theorem 1 in Hu & Lewbel [2008], except that I allow $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$ to depend on s_i^1.

Under theorem 1 estimation of $g(s_i^1, \mathbf{x}_i)$ can be performed as a two-step procedure, if a first-step estimator for $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$ is available.

In order to identify $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$ additional hypotheses are needed. These hypotheses can be of the following three forms.

First, we might suppose that under-reporting errors are essentially random and do not depend on the employer's declaration ($E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i) = E(s_i^2|s_i^1, \mathbf{x}_i)$), and have an intuition of the amount by which university graduates under-report their graduation status. Even an imperfect intuition might help to derive bounds.

Next, $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$ can be identified on validation data. By validation data, I intend a sample on which the reported graduation status is observed along with the true graduation status. With external validation data, hypotheses that limit the conditioning arguments are likely to become necessary: $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i) = E(s_i^2|s_i^* = 1, \mathbf{u}_i)$, where $\mathbf{u}_i$ is a sub-vector of $\mathbf{x}_i$. With internal validation data, $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$ can be identified on the subsample for which we observe $s^* = 1$ with certainty. Internal validation data that allow to identify under-reporting behaviour can also come in the form of an indicator implying $s^* = 1$ that is not itself related to self-reporting behaviour: $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i) = E(s_i^2|v_i = 1, s_i^1, \mathbf{x}_i)$ (v_i is an observed indicator for being in the validation sample and having $s_i^* = 1$, or for belonging to a particular population which implies $s_i^* = 1$: for instance, the proportion of degree holders who declare themselves not having the degree is identified on some profession which requires that degree).

Finally, $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i)$ can be identified from a measurement model. Under the two assumptions that measurement errors in the two reports are independent, and the auxiliary assumption that measurement errors are non-differential with respect to some observed variable which covaries with education, we can identify $E(s_i^2|s_i^*, \mathbf{x}_i)$ and $E(s_i^1|s_i^*, \mathbf{x}_i)$ using restrictions derived from a structural model of measurements inspired by Kane *et al.* [1999]. The independence assumption ensures that $E(s_i^2|s_i^* = 1, s_i^1, \mathbf{x}_i) = E(s_i^2|s_i^* = 1, \mathbf{x}_i)$.

4.2.3 A measurement model with independent measurement error

Suppose there exists a variable $\hat{z}_i$ in our data set with respect to which measurement errors are non-differential[7]:

$$E(\hat{z}_i|\mathbf{s}_i^*, \mathbf{s}_i^1, \mathbf{s}_i^2) = E(\hat{z}_i|\mathbf{s}_i^*) = \mathbf{s}_i^{*\prime}\tilde{\zeta} \tag{4}$$

Let $\mathbf{d}_i$ be a function of $\mathbf{s}_i^2$ and $\mathbf{s}_i^1$ only, defined as $\mathbf{d}_i = vec(\mathbf{s}_i^2\mathbf{s}_i^{1\prime})$: each cell in matrix $\mathbf{s}_i^2\mathbf{s}_i^{1\prime}$ corresponds to a possible combination of measurements.

Using this notation and from equation 4, 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 $J \times J$[8] unconditional moment restrictions:

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

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.

$$\begin{cases} \mathbf{s}_i^{1\prime} = \mathbf{s}_i^{*\prime}\mathbf{\Pi}_1 + \mathbf{e}_i^{1\prime} & E(\mathbf{s}_i^*\mathbf{e}_i^{1\prime}) = 0 \\ \mathbf{s}_i^{2\prime} = \mathbf{s}_i^{*\prime}\mathbf{\Pi}_2 + \mathbf{e}_i^{2\prime} & E(\mathbf{s}_i^*\mathbf{e}_i^{2\prime}) = 0 \\ E(\mathbf{e}_i^2\mathbf{e}_i^{1\prime}) = 0 \end{cases} \tag{5}$$

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 $E(\mathbf{d}_i\mathbf{s}_i^{*\prime})$ only depends on the prior probabilities for true education ($E(\mathbf{s}_i^*) = \delta^*$) and on the conditional probabilities in $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$, in a simple way. By appropriately resuming information in $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ in a single matrix $\mathbf{T}$[9], $E(\mathbf{d}_i\mathbf{s}_i^{*\prime})$ can be given the following representation:

$$E(\mathbf{d}_i\mathbf{s}_i^{*\prime}) = [\mathbf{T}'diag(\delta^*)] \tag{6}$$

The above hypotheses (4 and 5) 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[10]:

$$E(\mathbf{d}_i) = \mathbf{T}'\delta^* \tag{7}$$

$$E(\mathbf{d}_i\hat{z}_i) = \mathbf{T}'\zeta \tag{8}$$

[7] We use bold letters to denote vectors: in this section, educational attainment is not limited to be a binary variable. $\mathbf{s}_i^*$ is a complete set of mutually exclusive dummy variables, each representing a diploma.

[8] J represents the dimension of vector $\mathbf{s}_i^*$

[9] 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 $\mathbf{t}_n = [vec(\pi_n^{2\prime}\pi_n^1)]'$.

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

Table 8: Distribution of binary reports of academic qualifications, and of corresponding mean wages

worker / firm	0	1	Total
0	13912	1240	15152
	55.07	4.91	59.98
1	905	9204	10109
	3.58	36.44	40.02
Total	14817	10444	25261
	58.66	41.34	100

	worker	0	1
firm			
0	logwage	2.68	3.02
	se	0.00	0.01
1	logwage	3.09	3.27
	se	0.02	0.00

Table 9: Preliminary analysis: estimates of $g = E(g(1, \mathbf{x}_i) - g(0, \mathbf{x}_i))$ when selfreports represent the truth

	x	est.	se
g	no	0.411	0.015
g	linear	0.418	0.013
g	nonparam.	0.406	0.014

Note: All results are based on 25261 observations. Standard errors do not correct for survey design. In the linear specification, control variables include sex, age and age squared. The nonparametric estimator computes g separately for all sex/age combinations and averages over these estimates.

5 Empirical Results

5.1 Binary educational variables: Preliminary analysis

The binary version of educational attainment considered in the empirical application distinguishes "Any academic qualification" from "No academic qualification". Academic degrees correspond to levels '7' and '8' in the SES classification.

The distribution of reports of academic qualifications in the data set is given in table 8, along with the variation of wages with each combination of reports. The agreement rate for this binary version of educational attainment is 91.5%.

Under the hypothesis that self-reports reflect true educational attainment, the numbers in the right panel of table 8 are sufficient to compute the return attached to a false belief by the employer without controlling for covariates.

In table 9, the estimates corresponding to this hypothesis are displayed. Those who are falsely reported as being university graduates by their employers earn approximately 40% higher wages than correctly reported non graduates.

5.2 Nonparametric and semi-parametric estimation of g

When self-reports contain only one type of error, namely over-statements of educational attainment, then the estimates of g reported in table 9 hold valid, as shown in section 4.2.1 (p. 12).

With random underreporting, I estimate $E(s_i^2|s^*i = 1, \mathbf{x}_i) = \phi(\mathbf{x}_i)$ as the fraction of people which, within a cell defined by values of $\mathbf{x}_i$, reports having been regularly at school up to the age of 25. The identifying strategy is based on the hypothesis that any person who exits school at 25 or later has some

Table 10: Estimates of g with random underreporting

sex	agecat	w^{00}	w^{01}	$s^{1\|1}$	N	$\hat{\phi}$	U	$\hat{g}$	se
men	30-34	2.455	2.668	0.953	2736	0.986	422	0.213	
	35-39	2.581	2.973	0.939	3157	0.980	440	0.392	
	40-44	2.702	3.090	0.920	3290	0.970	335	0.389	
	45-49	2.769	3.180	0.896	3066	0.979	241	0.411	
	50-54	2.851	3.200	0.873	3348	0.957	279	0.349	
	55-59	2.919	3.315	0.846	1988	0.936	219	0.396	
women	30-34	2.353	1.815	0.943	1443	0.960	224	-0.538	
	35-39	2.449	2.428	0.917	1472	0.952	147	-0.021	
	40-44	2.534	2.665	0.912	1404	0.970	99	0.131	
	45-49	2.590	2.752	0.866	1371	0.957	70	0.163	
	50-54	2.666	2.961	0.823	1314	0.979	47	0.295	
	55-59	2.758	2.894	0.769	672	0.900	30	0.136	
Total					25261	0.966		0.253	

sex	agecat	w^{00}	w^{01}	$s^{1\|1}$	N	$\hat{\phi}$	$\hat{g}$	se
men	30-34	2.455	2.535	0.953	2736	0.977	0.080	
	35-39	2.581	2.913	0.939	3157	0.973	0.332	
	40-44	2.702	3.090	0.920	3290	0.970	0.388	
	45-49	2.769	3.163	0.896	3066	0.974	0.394	
	50-54	2.851	3.212	0.873	3348	0.960	0.361	
	55-59	2.918	3.298	0.846	1988	0.932	0.379	
women	30-34	2.354	2.331	0.943	1443	0.977	-0.024	
	35-39	2.450	2.643	0.917	1472	0.973	0.193	
	40-44	2.534	2.668	0.912	1404	0.970	0.134	
	45-49	2.590	2.824	0.866	1371	0.974	0.234	
	50-54	2.665	2.924	0.823	1314	0.960	0.259	
	55-59	2.759	2.976	0.769	672	0.932	0.217	
Total					25261	0.967	0.276	

Column headings: U: number of observations with reported age at school exit bigger than 25.

$$w^{00} = \widehat{E}(w_i|s_i^* = 0, s_i^1 = 0) \qquad w^{01} = \widehat{E}(w_i|s_i^* = 0, s_i^1 = 1)$$
$$s^{1|1} = \widehat{E}(s_i^2|s_i^1 = 1 \qquad \hat{\phi} = \widehat{E}(s_i^2|s_i^* = 1)$$

Note: Standard errors do not correct for survey design. The nonparametric estimator computes g separately for all sex/age category combinations and averages over these estimates. The bottom panel imposes ϕ to be the same for men and women belonging to the same age category.

academic qualification, and that reporting an age at school exit of 25 or more does not influence the reporting behaviour with respect to diplomas.

Table 10 displays estimates of g under these hypotheses. These results suggest that the underreporting rate increases with age, but is not different for men and women with similar age. I therefore impose this equality to estimate ϕ in the bottom panel. The estimates of ϕ are decreasing with age, as is consistent with memory flaws being more important for senior workers. The wage differential among non-university graduates attached to employer reports is, in contrast, increasing with age. Estimates also indicate that it is significantly lower for women than for men.

If reporting errors are independent of each other and nondifferential with respect to gender, then ϕ can be estimated from the data as a parameter in the model defined by equations (7) and (8).

I use gender as our $\hat{z}_i$ variable, to exploit the fact that the proportion of women in the labor force differs by educational level. The rationale is that I exclude measurement errors to follow different patterns depending on the worker's

Table 11: GMM estimates of ϕ, based on independence of measurement errors and nondifferential errors with respect to gender

age	pooled	30-34	35-39	40-44	45-49	50-54	55-59
no ac.	0.625	0.439	0.548	0.633	0.727	0.753	0.742
ac.	0.375	0.561	0.452	0.367	0.273	0.247	0.258
women/n.ac	0.318	0.301	0.319	0.316	0.335	0.312	0.297
women/ac.	0.281	0.372	0.317	0.268	0.238	0.187	0.127
firm							
precision	0.975	0.971	0.965	0.965	0.964	0.960	0.950
overreport	0.025	0.029	0.035	0.035	0.036	0.040	0.050
underreport	0.000	0.000	0.000	0.000	0.000	0.000	0.000
worker							
precision	0.937	0.960	0.946	0.956	0.955	0.947	0.940
overreport	0.051	0.040	0.054	0.044	0.045	0.053	0.060
underreport	0.013	0.000	0.000	0.000	0.000	0.000	0.000
ϕ	0.966	1.000	1.000	1.000	1.000	1.000	1.000
N	25261	4179	4629	4694	4437	4662	2660

Note: no ac. is the estimated proportion of the sample having no academic degree, ac. is the proportion with any academic degree. women/no ac. and women/ac. are the estimated proportions of women in both populations. Precision is the probability that the report equals the true value, overreport the probability that the report is academic degree when in fact the worker does not possess any. ϕ denotes the estimated probability that a worker having an academic degree reports so.

gender. Hereby I not only exclude that men lie more than women (or differently), but also that employers do not have discriminatory beliefs about the education of their male and female labour force.

Although the parameters are identified without further restrictions, many parameters represent probabilities, and are therefore restricted to take values between 0 and 1; to ease estimation and incorporate these restrictions, I apply the transformation $F : p \rightarrow ((\arctan(p)/\pi) + .5)$ to all parameters representing probabilities in moment conditions.

Results of this measurement model are displayed in table 11. When observation are pooled, I estimate a conditional probability of underreporting (ϕ) equal to .966, which is close to the average estimate in table 10. Because of the feminisation of higher education over the last 40 years, however, the nondifferential nature of measurement error with respect to gender might hold only conditional on age: because the gender distribution across diplomas covaries with age, gender by itself might covary with errors in reporting diplomas. To get rid of this problem I estimate the model separately on 6 agegroups.

Estimates of these models are consistent with major trends in the French labor force: the increase in schooling across cohorts, the feminisation of higher education coupled with high participation rates among skilled women. Informational parameters suggest that employer reports are more precise at any age category; they also suggest that errors are only in one direction, ruling out the presence of underreports in data.

Based on this evidence, we might therefore trust the estimates of g reported in table 9. In table 12 I display the value of g computed under the hypothesis that workers do not underreport their education by age group and gender combination.

Differences across gender and age categories are barely or not significant. Overall, a false university graduate significantly earns more than the average non graduate.

Table 12: Estimates of g with no underreporting, by gender and age category

sex	agecat	N	$\hat{g}$	se
men	30-34	2736	0.319	0.040
	35-39	3157	0.491	0.040
	40-44	3290	0.496	0.038
	45-49	3066	0.465	0.039
	50-54	3348	0.461	0.037
	55-59	1988	0.531	0.046
women	30-34	1443	0.187	0.054
	35-39	1472	0.314	0.049
	40-44	1404	0.291	0.055
	45-49	1371	0.306	0.056
	50-54	1314	0.326	0.059
	55-59	672	0.315	0.079

Note: Standard errors do not correct for survey design.

6 Relaxing independence assumptions

The identification argument in theorem 1 critically depends on the (partial) independence of measurement error in self-reports and employer reports of education. The measurement model estimated in table 11 assumes even complete independence of measurement errors.

While we cannot identify g without this assumption, we can identify the coefficient on the employer's report in the linear projection $L(w_i|s_i^*, s_i^1, \mathbf{x}_i)$ without having resort to this assumption, using common IV techniques.

If the model for the conditional mean is linear in $\mathbf{x}_i$, the coefficient on s_i^1 in the linear projection can be interpreted as a mixture of $[g(1) - g(0)]$ and $[h(1) - h(0)]$.

Indeed, when the effect of all control variables $\mathbf{x}_i$ is additive and separable, equation (1) can be rewritten as

$$w_i = s_i^* \beta_{01} + s_i^1 \beta_{10} + s_i^* s_i^1 \beta_{11} + \mathbf{x}_i' \theta + \eta_i$$

where $\beta_{10} = g(1) - g(0)$, $\beta_{10} + \beta_{11} = h(1) - h(0)$.

From the linear projection $L(s_i^* s_i^1 | s_i^*, s_i^1, \mathbf{x}_i) = \hat{\alpha} + \hat{\gamma}_0 s_i^* + \hat{\gamma}_1 s_i^1$, the coefficient β_1 in equation 9 equals $\beta_1 = (\beta_{10} + \hat{\gamma}_1 \beta_{11})$, with $0 < \hat{\gamma}_1 < 1$

$$w_i = \alpha + \beta_0 s_i^* + \beta_1 s_i^1 + \mathbf{x}_i' \theta + v_i \qquad E((s_i^*, s_i^1, \mathbf{x}_i')v_i) = 0 \tag{9}$$

In general we assume that s_i^* is unobserved, so that the parameters of the linear projection (9) can not be directly estimated on the data. Using additional independent indicators for educational attainment, instrumental variable techniques allow the identification of the parameter β_1 in the following linear projection:

Suppose there exists an indicator y_i for true education which verifies equation (11), and that the worker's selfreport verifies (12), along with (10):

$$L(w_i|s_i^*, s_i^1, s_i^2, y_i, \mathbf{x}_i) = L(w_i|s_i^*, s_i^1, \mathbf{x}_i) \tag{10}$$

$$L(y_i|s_i^*, s_i^1, s_i^2, \mathbf{x}_i) = L(y_i|s_i^*, \mathbf{x}_i) \tag{11}$$

$$L(s_i^2|s_i^*, s_i^1, y_i, \mathbf{x}_i) = L(s_i^2|s_i^*, s_i^1, \mathbf{x}_i) \tag{12}$$

The hypotheses in (11) and (12) allow to implement the "multiple indicator solution" to get a correct estimate of β_1 (not of other parameters in eq. 9, though), by plugging in y_i for s_i^* in equation (9) and instrumenting this indicator with s_i^1.

Note that the identifying hypotheses, and in particular equation (12), allow for correlated reports of educational attainment. The crucial hypotheses state that the indicator y_i must not predict measurement error by either the worker or the employer.

I use, as additional independent indicator for having a university degree, a dummy taking value 1 if the worker reports having been regularly at school beyond the age of 20. This variable needs to verify in particular the following identifying hypotheses:

1. selfreported graduation status must not be conditionally correlated with unobserved wage determinants (eq. 10);

2. a non-graduated who has been at school for longer than the average non degree holders must not be more likely to report being graduated (eq. 12)...

3. ...nor more likely to be reported as graduated by his employer (eq. 11).

Table 13 displays the results of this analysis. Note first that the reduced form is just a regression of wages on worker and employer reports of university graduation: employer reports appear systematically as having a higher coefficient than workers'. Next, Iv estimates indicate that the coefficient on employers' reports is positive and significant when controlling for true education, supporting the hypothesis that good opinions on education influence wages significantly. The coefficient is correctly identified if our third indicator, based on reported age at school exit, is not correlated with wages and reported education conditional on true education and covariates. We might suspect a positive correlation with selfreported education arising from the fact that the two are reported by the same person. If this was the case, there would be a downward bias in the IV estimate of β_1[11].

7 Interpretation

7.1 Differential misclassification as an effect of the binary nature of s^*?

The use of binary indicators for education in place of full sets of dummy variables or of a continuous indicator such as years of schooling, when there are reasons to believe that these are more appropriate in a model for the conditional mean of wages, might be itself the reason why a non-differential measurement becomes differential.

[11] We proceed by assuming that only one necessary orthogonality condition is violated to characterise the sign of the bias. Indeed, the estimated parameter vector $\hat{\beta}^{IV}$ equals $\beta + (Z'X)^{-1}(Z'\varepsilon)$. Where $E(Z'\varepsilon)$ is nonzero, giving a sign to the element which is responsible for this violation is sufficient to sign the bias in every coefficient included in $\hat{\beta}^{IV}$ ($(Z'X)^{-1}$ is observed), and in particular $\beta_1^{\hat{I}V}$.

Table 13: Estimates of β_1 using age at school exit as indicator for educational attainment

	reduced form		1st stage		2nd stage	
dep. var.	log(w)		ase21		log(w)	
	est.	*se*	est.	*se*	est.	*se*
Intercept	2.685	*0.004*	0.073	*0.003*	2.650	*0.004*
firm	0.324	*0.010*	0.238	*0.007*	**0.209**	*0.014*
ase21					0.480	*0.018*
worker	0.273	*0.010*	0.568	*0.007*		

	reduced form		1st stage		2nd stage	
dep. var.	log(w)		ase21		log(w)	
	est.	*se*	est.	*se*	est.	*se*
Intercept	0.941	*0.077*	1.267	*0.060*	0.199	*0.085*
firm	0.348	*0.009*	0.229	*0.007*	**0.213**	*0.013*
ase21					0.586	*0.017*
age	0.062	*0.004*	-0.051	*0.003*	0.092	*0.004*
agesq/100	-0.047	*0.004*	0.054	*0.003*	-0.078	*0.004*
femme	-0.194	*0.005*	-0.027	*0.004*	-0.178	*0.006*
worker	0.325	*0.009*	0.555	*0.007*		

Legend: ase21: dummy for age at school exit bigger than or equal to 21; firm: dummy for firm reporting any academic qualification; worker: dummy for worker reporting any academic qualification.
Note: Standard errors do not correct for survey design.

Suppose the true model for wages is in terms of a non-binary measure of education e_i; and s_i^* reflects a binary version of this variable. Let s_i^1 be an indicator for this binary version, subject to misclassification: it might itself be constructed by dichotomising a richer indicator for e_i (call this e^1).

Suppose $E(w|e, e^1) = E(w|e, e^1, s^*, s^1) = E(w|e)$ (measurement error in e^1 is nondifferential with respect to w). Still, under fairly general conditions, $E(w|s^*, s^1) \neq E(w|s^*)$: misclassification in s^1 is differential with respect to w, as first shown by Flegal *et al.* [1991].

The reason is that when misclassification probabilities are not constant for a given value of s^*, s^1 contains information on e beyond the information contained in s^*:

$$f_{e_i|s_i^*, s_i^1} \neq f_{e_i|s*_i}$$

Furthermore, if e is informative about wages beyond s^*, then s^1 will also be informative about wages, even conditioning on s^*.

In conclusion, reducing education to a binary variable might itself be the origin of non-differential misclassification. It is therefore important to keep all the information available in the survey about education in our tests for non-differential misclassification. In the context of estimation of g, if the population with $(s^* = 0, s^1 = 0)$ is on average less educated than the population with $(s^* = 0, s^1 = 1)$, because misclassification occurs mainly near the border, then we could interpret a positive g just as reflecting the fact that wages depend on richer accounts of education than s^*.

Indeed, estimates in table 14 - where the measurement model was estimated using the full vector of possible educational reports - suggest that erroneous reports tend to fall near the truth, and therefore the differential nature of binary measurements with respect to wages could originate from a statistical artifact.

A test for the non-differential nature of measurement error when all educational levels are taken into account can however be implemented: if, conditional

on $\mathbf{x}_i$, measurement error is non-differential with respect to wages, and if the model for the conditional mean of wages is linear in $\mathbf{x}_i$, we can add restrictions derived from

$$E(w_i|\mathbf{s}_i^*, \mathbf{s}_i^1, \mathbf{s}_i^2, \mathbf{x}_i) = E(w_i|\mathbf{s}_i^*, \mathbf{x}_i) = \mathbf{s}_i^{*\prime}\tilde{\beta} + \mathbf{x}_i'\theta \tag{13}$$

to the just-identified set of estimating equation defined by equations (7) and (8), and test for over-identifying restrictions. Parameters reflecting the informational structure of measurements are indeed over-identified if we add the following restrictions to estimation:

$$E(\mathbf{d}_i w_i) = \mathbf{T}'\beta + E(\mathbf{d}_i \mathbf{x}_i{}')\theta \tag{14}$$

The over-identification test strongly rejects the null hypothesis that the above system of identifying assumptions formed by equations (14), (7) and (8) is jointly valid on data from the ESS 2002 (Sargan overidentification statistic $\widehat{S}$=472.11, p-value (39 DF)=0.000). The rejection of the over-identification test provides evidence in favour of differential measurement error being a genuine feature of employer's knowledge about the employee's education.

This rejection could result from any of the identifying hypotheses failing. The independence assumption on measurements and the non-differential nature of measurement errors with respect to wages are plausible candidates to explain that failure.

However, in contrast to estimates in table 14 that look very plausible, when the moment restrictions derived from the hypothesis that measurement errors are also non-differential with respect to wages are added to estimation, the estimates do not converge to plausible values (see table 15), suggesting that this last assumption is responsible for the violation of the over-identification test.

7.2 Causal and further non-causal interpretations

The fact that those who falsely appear as having a given diploma have a significantly higher wages than non-degree holders recognised as such, if proven to be a genuine stylised fact, needs still to be interpreted.

Causal interpretation If the wage differential is to be interpreted as a genuine wage gain, this suggests that it is possible in France to free-ride on diplomas; there is a sort of placebo return[12] to schooling which is almost as big as the total return to schooling. The existence of such a return attached to the mere belief by the employer that the worker is educated is not sufficient to settle the debate on returns to schooling in favour of the pure signaling theory, because this study does not allow to say whether diplomas signal innate ability or knowledge acquired at school, but suggests that firms use diplomas as signals. If, as the survey suggests,

[12]Placebo effects are observed in randomised medical trials when patients assigned to a control group treated with pharmacologically inert preparations (placebos) experience improvements relative to a second control group, who does not receive treatment of any kind. The analogy is with an employer that hires a placebo-educated worker.

Table 14: GMM estimates, based on independence of measurement errors and nondifferential errors with respect to gender

Implied distribution of educational attainment

	Percent	Proportion of women
1	7.03	33.65
3	8.50	43.67
4	28.93	24.86
5	9.44	28.71
6	7.61	46.31
7	14.87	33.45
8	23.63	24.60

Conditional probabilities for firm reports

firm truth	1	3	4	5	6	7	8
1	0.902	0.085	0.000	0.000	0.013	0.000	0.000
3	0.215	0.473	0.175	0.034	0.081	0.022	0.000
4	0.139	0.034	0.810	0.003	0.011	0.003	0.000
5	0.000	0.026	0.000	0.792	0.000	0.181	0.000
6	0.003	0.000	0.057	0.307	0.569	0.063	0.000
7	0.000	0.000	0.000	0.000	0.037	0.963	0.000
8	0.003	0.003	0.000	0.000	0.010	0.021	0.963

Conditional probabilities for worker reports

worker truth	1	3	4	5	6	7	8
1	0.972	0.000	0.000	0.000	0.000	0.028	0.000
3	0.047	0.535	0.262	0.085	0.071	0.000	0.000
4	0.120	0.015	0.774	0.071	0.000	0.015	0.004
5	0.034	0.054	0.211	0.513	0.000	0.156	0.032
6	0.000	0.000	0.022	0.278	0.459	0.178	0.063
7	0.000	0.000	0.000	0.000	0.045	0.789	0.166
8	0.002	0.001	0.008	0.009	0.009	0.067	0.905

Implied precision of educational reports

	firm	worker
$Pr(\mathbf{s}_i = \mathbf{s}_i^*)$	0.827	0.752
$Pr(\mathbf{s}_i > \mathbf{s}_i^*)$	0.060	0.124
$Pr(\mathbf{s}_i < \mathbf{s}_i^*)$	0.113	0.123
ϕ		0.966

Note: Each entry in the two matrices reads as the probability of a report conditional on the truth (example: the probability of the firm reporting a "baccalauréat technologique ou professionnel" (5) when the true attainment is a "baccalauréat général" (6) is 0.307). ϕ denotes the estimated probability that a worker having an academic degree reports having any academic degree.

Table 15: GMM estimates, based on independence of measurement errors and nondifferential errors with respect to gender and wage

Conditional probabilities for firm reports

firm truth	1	3	4	5	6	7	8
1	0.881	0.113	0.000	0.000	0.006	0.000	0.000
3	0.167	0.050	0.782	0.000	0.001	0.000	0.000
4	0.013	0.000	-0.013	0.000	0.000	1.000	0.000
5	0.000	0.855	0.000	0.145	0.000	0.000	0.000
6	0.000	0.000	0.201	0.651	0.148	0.000	0.000
7	0.000	0.000	0.000	0.000	0.680	0.320	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	1.000

Conditional probabilities for worker reports

worker truth	1	3	4	5	6	7	8
1	1.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.053	0.926	0.021	0.000	0.000	0.000
4	0.004	0.007	0.030	0.065	0.000	0.765	0.129
5	0.000	0.653	0.216	0.120	0.000	0.011	0.000
6	0.000	0.000	0.192	0.526	0.092	0.170	0.020
7	0.000	0.000	0.000	0.000	0.593	0.306	0.100
8	0.001	0.000	0.007	0.008	0.007	0.060	0.917

Note: Each entry in the two matrices reads as the probability of a report conditional on the truth (example: the probability of the firm reporting a "baccalauréat technologique ou professionnel" (5) when the true attainment is a "baccalauréat général" (6) is 0.092).

employer's belief reflect workers declarations at hiring, then the gain can be interpreted as a return to being effective at lying. Even if a causal mechanism explains the correlation, the magnitude of the estimated gain is probably influenced by selection biases: those who are effective at lying might have peculiar unobserved characteristics, and, with learning going on, some liars which are not grown to their lies get caught and leave their employer.

Endogeneity of beliefs In a conservative interpretation, employer's beliefs correlate with significant wage differentials because they correlate with some unobserved personal characteristics which are positively correlated with wages: smartness, assertiveness, self-esteem, are all plausible candidates.

Reverse causality In the extreme case, endogeneity of employers' beliefs arises not from their correlation with omitted unobserved variables, but directly from their correlation with productivity or wages. This would be the case if employers form a guess on the worker's diploma based on their wages or productivity.

I explore the issue of reverse causality in table 16 by distinguishing small and large establishments in running estimations. I expect that if there is a reverse causality problem, it is concentrated in small establishments where the information system on employees is more informal. Results indicate a small, but insignificant difference in the estimates of g by establishment size.

Table 16: Estimates of g by establishment size

A: No underreporting by workers

Establishment size $<$ 200

	x	est.	se
g	no	0.421	0.020
g	linear	0.434	0.018
g	nonparam.	0.423	0.020

Establishment size $\geqslant$ 200

	x	est.	se
g	no	0.397	0.021
g	linear	0.397	0.019
g	nonparam.	0.374	0.020

B: Independent underreporting by workers

Establishment size $<$ 200

sex	agecat	w^{00}	w^{01}	$s^{1\|1}$	N	$\hat{\phi}$	$\hat{g}$	se
men	30-34	2.422	2.553	0.944	1506	0.964	0.131	
	35-39	2.539	2.927	0.909	1687	0.972	0.388	
	40-44	2.634	3.115	0.886	1614	0.960	0.481	
	45-49	2.710	3.157	0.861	1502	0.978	0.447	
	50-54	2.791	3.189	0.835	1575	0.956	0.398	
	55-59	2.869	3.289	0.829	1038	0.918	0.420	
women	30-34	2.321	2.218	0.932	901	0.964	-0.103	
	35-39	2.416	2.672	0.913	929	0.972	0.256	
	40-44	2.506	2.569	0.900	808	0.960	0.064	
	45-49	2.567	2.807	0.822	805	0.978	0.240	
	50-54	2.629	2.877	0.761	778	0.956	0.247	
	55-59	2.732	2.806	0.765	398	0.918	0.074	
Total					13541	0.961	0.293	

Establishment size $\geqslant$ 200

sex	agecat	w^{00}	w^{01}	$s^{1\|1}$	N	$\hat{\phi}$	$\hat{g}$	se
men	30-34	2.507	2.597	0.962	1230	0.993	0.090	
	35-39	2.647	2.838	0.964	1470	0.973	0.191	
	40-44	2.783	3.049	0.946	1676	0.977	0.265	
	45-49	2.832	3.171	0.923	1564	0.971	0.339	
	50-54	2.916	3.243	0.898	1773	0.963	0.327	
	55-59	2.995	3.306	0.858	950	0.942	0.312	
women	30-34	2.430	2.506	0.958	542	0.993	0.076	
	35-39	2.524	2.594	0.924	543	0.973	0.070	
	40-44	2.583	2.764	0.923	596	0.977	0.181	
	45-49	2.626	2.878	0.917	566	0.971	0.252	
	50-54	2.723	3.051	0.886	536	0.963	0.328	
	55-59	2.799	3.167	0.776	274	0.942	0.368	
Total					11720	0.972	0.243	

Note: Sample size is 13541 for the left panel and 11720 for the right panel. Standard errors do not correct for survey design. In the linear specification, control variables include sex, age and age squared. The nonparametric estimator computes g separately for all sex/age combinations and averages over these estimates.

8 Conclusive remarks

- possible applications of the empirical strategy: measuring discrimination (nationality). Union status (Jakubson).

A Appendix

- Learning?
- False positives and false negatives.
- Impostors: Aleksander Kwasniewski[13], Rachida Dati's MBA[14], Alexis Debat[15]

[13] http://en.wikipedia.org/wiki/Aleksander_Kwa%C5%9Bniewski#Degree
[14] http://www.rue89.com/2007/10/25/rachida-dati-a-t-elle-menti-sur-ses-diplomes
[15] http://en.wikipedia.org/wiki/Alexis_Debat

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. 2006. *Misreported Schooling and Returns to Education: Evidence from the UK*. Institute for Fiscal Studies CEMMAP working paper CWP07/06.

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.

Flegal, Katherine M., Keyl, Penelope M., & Nieto, F. Javier. 1991. Differential Misclassification Arising from Nondifferential Errors in Exposure Measurement. *American Journal of Epidemiology*, **134**(10), 1233–1246.

Hu, Yingyao, & Lewbel, Arthur. 2008. *Identifying the returns to lying when the truth is unobserved*. Cemmap working paper CWP6/08.

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

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.