Misreported Schooling in Survey and Administrative Data and Returns to Educational Qualifications

Erich Battistin

University of Padova, IRVAPP and IZA

Michele De Nadai

University of Padova

Barbara Sianesi

Institute for Fiscal Studies

http://ftp.iza.org/dp6337.pdf

5th ESRC Research Methods Festival, July 2012

Motivation

Categorical vis-à-vis **continuous** measures of education:

more adequate when different educational routes may potentially have very different returns

Extent of misclassification in qualifications data (Kane *et al.*, 1999):

- **transcript errors**: transcript measures found to be subject to at least as much error as self-reported measures
- **misreporting**: respondents may lie, not know if the schooling they have had counts as a qualifications or not remember.
    - more likely for low levels of qualifications
    - over-reporting more likely than under-reporting

Estimates of returns can be heavily affected – Kane *et al.* (1999), Bound *et al.* (2000), Lewbel (2006), Hu (2012)

What we do

1. **Return from attaining any academic qualification** (compared to leaving without formal qualifications) allowing for misreported attainment – tricky because non-classical measurement error

2. **Extent of measurement error** in:
   - administrative information
   - self-reported information very close to completion
   - recall information 10 years later

   Temporal patterns of misreporting errors across survey instruments
   Decompose misreporting errors into systematic individual component and transitory random survey errors.

3. **How misclassification and omitted-variable biases interact** → calibration rules

4. **Semi-parametric estimation** approach based on balancing scores and mixture models (allows for arbitrarily heterogeneous individual returns)

What we find

1. Evidence on measurement error in 3 types of UK data on educational attainment

- Self-reports and transcript data: no source uniformly better
- For individuals, over-reporting is more of a problem; for transcripts, under-reporting
- Despite different underlying patterns of error, the two types of data are very similar in their overall reliability when information collected close to completion (80%), 3-4pp lower when based on recall
- Figures from just one wave not likely to reveal behaviour;
  Still bulk of correct classification can be attributed to persistency in individual reporting across waves (90% of measurement error in NCDS related to individual behaviour)

2. Evidence on true return and interplay between the two biases

- Return allowing for measurement error: 26.4% wage gain
  (statistically different from return ignoring misclassification)
- Ability *vs* measurement error biases
  - ⇒ Reports collected close to completion (school or individual): ignoring both leads to ↑ bias
    Calibration: 0.80·LFS-style estimate
  - ⇒ Reports rely on recall: the two biases cancel out

General Formulation of the Problem

$D^* \in \{0,1\}$ indicator of any academic qualifications

Y_0, Y_1 potential (log)wage if no quals and if any quals

$$Y = Y_0 + (\underbrace{Y_1 - Y_0}_{})D^*$$ observed individual (log)wage

causal effect of *D** on *Y* (here, individual return to qualifications)

$\Delta^*(x) \equiv E(Y_1 - Y_0 \mid D^* = 1, x)$ conditional treatment effect

$\Delta^* \equiv E(Y_1 - Y_0 \mid D^* = 1)$ **Average Treatment effect on the Treated** (ATT)

$$= E(Y_1 \mid D^* = 1) - \underbrace{E(Y_0 \mid D^* = 1)}_{\text{counterfactual}}$$

Assumptions – to identify counterfactual

Conditional Independence (CIA)

Conditional on X, D^* is independent of Y_0 and Y_1:

$$(Y_0, Y_1) \perp D^* \mid X$$

- abstract from omitted-variable bias to focus on impact of mismeasured quals
- rich data (NCDS) – building on Blundell, Dearden & Sianesi (2005)

Common Support (CS)

Individuals with and without the qualification can be found at all values of X:

$$0 < P(D^*=1 \mid X) < 1$$

(CIA)+(CS) → identification of Δ^* when observing (Y, D^*, X)

Misclassification and Multiple Measures

Observe $(Y, D_S^1, D_S^2, D_T, X)$; possibly $D_S^j \neq D^*$ (j=1,2) and $D_T \neq D^*$

Δ^* is not identified from raw data in general;
bias depends on the extent of misclassification:

Probabilities of exact classification for any measurement $W = \{D_S^1, D_S^2, D_T\}$
% of truth tellers or of individuals correctly classified in transcript files amongst those
with quals $f_{W|D^*X}[1|1,x]$
without quals $f_{W|D^*X}[0|0,x]$

Assumptions on Measurement Error

Non-Differential Misclassification given *X*

Any variables $\mathbf{D}_S$ and D_T which proxy D^* do not contain information to predict Y conditional on D^* and X:

$$f_{Y|D^*,\mathbf{D}_S,D_T,X}(y\,|\,d^*,\mathbf{d}_s,d_T,x) = f_{Y|D^*,X}(y\,|\,d^*,x)$$

Independent Sources of Error given *X*

$\mathbf{D}_S$ and D_T are independent given D^* and X:

$$f_{\mathbf{D}_S,D_T|D^*,X}(\mathbf{d}_s,d_T\,|\,d^*,x) = f_{\mathbf{D}_S|D^*,X}(\mathbf{d}_s\,|\,d^*,x)\,f_{D_T|D^*,X}(d_T\,|\,d^*,x)$$

Identification

Mixture representation

Under our assumptions, the distribution of observed wages conditional on *X* for 2x2x2 groups defined by D_S^1xD_S^2xD_T is a **mixture** of the *two* latent distributions

$$f_{Y_1|X}(y_1 \mid x)$$

$$f_{Y_0|X}(y_0 \mid x)$$

$$\boxed{f_{Y|\mathbf{D}_S, D_T, X}(y \mid \mathbf{d}_s, d_T, x) = [1 - p(\mathbf{d}_s, d_T, x)] f_{Y_0|X}(y \mid x) + p(\mathbf{d}_s, d_T, x) f_{Y_1|X}(y \mid x)}$$

with mixture weights $p(\mathbf{d}_s, d_T, x) \equiv f_{D^*|\mathbf{D}_S, D_T, X}(1 \mid \mathbf{d}_s, d_T, x)$

Mixture weights → get probabilities of exact classification relative to each measure (Bayes)

Mixture components → get $\Delta^*(x)$

Non-parametric identification of mixture weights and components

- Additional assumptions (conditional on X)
  - o Relevance of educational qualifications
  - o Informational content of D_T

- Intuition:
  Information on proportion of individuals classified differently by different (independent) measures can be combined with information on the difference in their earnings to estimate the distribution of reporting errors in *both* measures.

(a)	$E(Y \mid D_T=1, D_S=1)$	(c)	$E(Y \mid D_T=1, D_S=0)$
(b)	$E(Y \mid D_T=0, D_S=1)$	(d)	$E(Y \mid D_T=0, D_S=0)$

- Technically:
  Use D_T as a source of *instrumental variation* to define a large enough number of moment conditions given the unknowns in the mixture representation

- Multiple self-reported measurements introduce **over-identification** → additional moment restrictions that can be used to allow for correlation in self-reported measurements

Estimation of Returns

To date, estimation relied on fully parametric models (Kane *et al.*, 1999, Black *et al.*, 2000, Lewbel, 2005 and Hu, 2012)

Suggest **semi-parametric estimation** by restricting ourselves to a class of parametric mixtures:

- Assume log-normality of potential wages within cells (empirical evidence supports it)
- Exploit balancing scores to deal with curse of dimensionality (as in Battistin and Sianesi, 2010)
- Allow for arbitrarily heterogeneous returns
- Allow for correlated reports (e.g. reports from the same individuals over time)
- Variety of estimation procedures available (e.g. EM algorithm, Bayesian modelling)

Application to NCDS Data

- 1958 NCDS cohort: 2,716 working males, non-miss education
- ***Y*** = real gross hourly wage at 33 (in 1991)
- ***D**** = any academic qual (i.e. at least O-lev by age 20) vs leaving at 16 with none
  - independent measure for O- and A-levels only
  - academic quals are well defined and homogeneous
  - policy interest: main effect of ROSLA was to induce individuals to leave school with O-levels (Chevalier *et al*., 2003, Galindo-Rueda, 2004)
- ***X*** =
  - gender and age, ethnicity, region ("LFS-style controls")
  - math and reading ability at 7 and 11
  - family background (age and education, father's social class, mother's employment, number of siblings)
  - school type

NCDS: Measures of qualifications

- Obtained by age 23, ***self-reported at age 33*** (1991 sweep)
- Obtained by age 23, ***self-reported at age 23*** (1981 sweep)
- Obtained by age 20, ***admin*** (1978 School Files)

Wage returns to any academic qualifications by age 20

	(1)	(2)	(3)	Tests of equality		
	Transcript (schools)	**1981 wave** (at age 23)	**1991 wave** (at age 33)	(1)=(2)	(1)=(3)	(2)=(3)
Δ_{LFS}	0.332 (*0.015*)	0.333 (*0.016*)	0. 293 (*0.016*)		***	***
Δ_{FULL}	0.194 (*0.018*)	0.194 (*0.018*)	0.151 (*0.018*)		***	**

	Transcript files			
	Any		None	
1981 wave	**1991 wave** (at age 33)		**1991 wave** (at age 33)	
(at age 23)	Any	None	Any	None
Any	1445	103	148	70
None	24	25	120	781

Incidence of qualifications in the population

- transcript → 58.8%
- self-reports → 64% at age 23, 65% at age 33

Agreement rates

- overall → 82%
- transcript and self-reported 1981 → 90%
- transcript and self-reported 1991 → 85%
- self-reported 1981 and 1991 → 88%

Despite substantial formal agreement between measures, remaining divergences can lead to substantially and significantly different impact estimates.

If transcript were the "truth"

- more over- than under-reporting (at 23: 20% vs 3%; at 33: 25% vs 8%)
- errors get worse as individuals recall

Descriptive analysis of degree of concordance

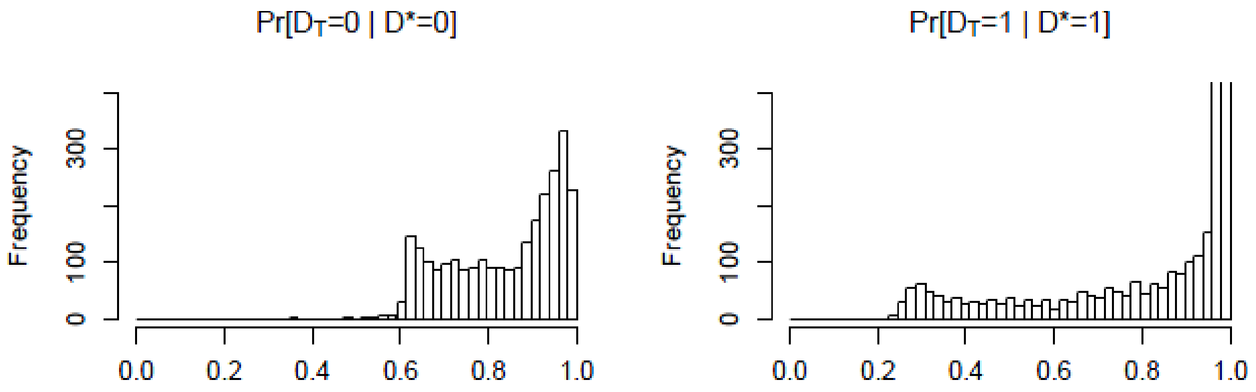
- very low predictive power of X
  - esp. for $P(D_{\mathrm{T}}=D_S^2)$ (3.9% of Var)
- professional father ↑ $P(D_S^1=D_S^2)$
- higher math ability ↑ $P(D_{\mathrm{S}}=D_{\mathrm{T}})$
- secondary modern and comprehensive types of school: < agreement rates

Results:
(1) Characterising Misclassification

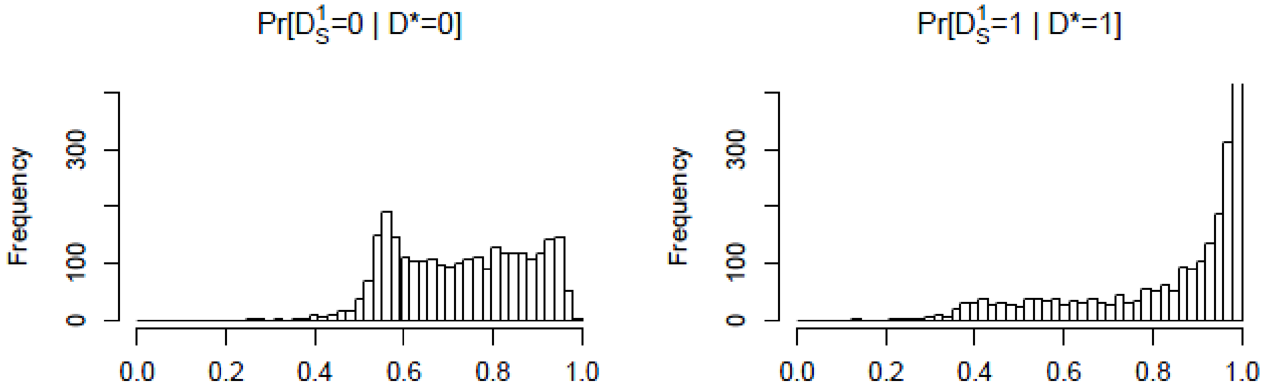
Probabilities of reporting correctly…

…not to have any quals … to have academic quals

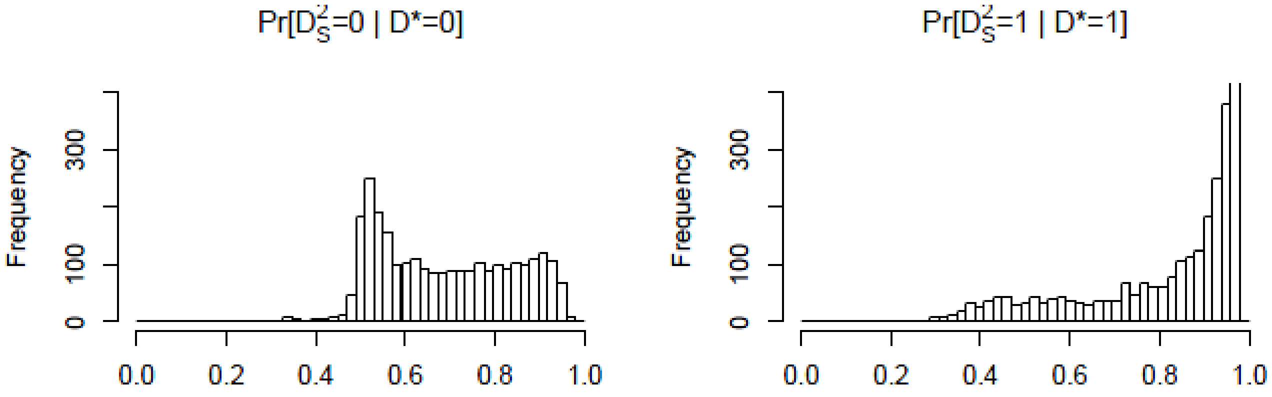
Transcript



Self-report, age 23



Self-report, age 33



	Transcript (schools)	*1981 Wave (age 23)*	*1991 Wave (age 33)*
Any qualification			
- prob exact classification	0.783	0.847	0.811
- prob under-reporting	0.217	0.153	0.189
No qualifications			
- prob exact classification	0.836	0.729	0.687
- prob over-reporting	0.164	0.271	0.313
Correct classification	0.800	0.803	0.765

Individuals

- more accurate than transcripts when they do have quals
- less accurate when they don't have any qualification
- both types of errors worsen over time
- though small effect of time (survey close to completion is only 3-4pp more accurate)

No source uniformly better (in line with the little US evidence)

- Individuals: over-reporting more important
- Transcripts: under-reporting more important, though more similar incidence of both types of error
- Despite different underlying patterns of measurement error, the two types of data are remarkably similar in their overall reliability, esp. when information collected close to completion

Incidence of qualifications in the population → $P(D^* = 1) = 64.1\%$

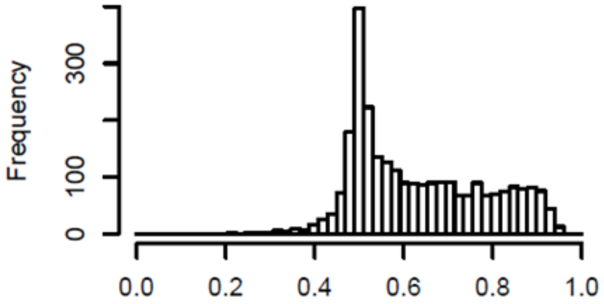
$P(D_T=1) = 58.8\%$
$P(D_S^1=1) = 64.0\%$
$P(D_S^2=1) = 65.0\%$

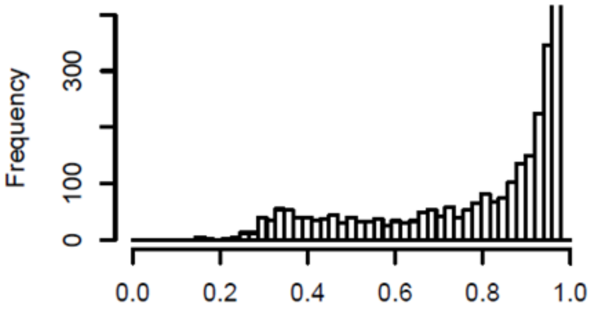
Temporal patterns and decomposition of misreporting errors

CONSISTENT TRUTH TELLERS

No qualifications **Any qualification**

$P(D_S^1{=}0, D_S^2{=}0 \mid D^*{=}0) = \mathbf{0.631}$ $P(D_S^1{=}1, D_S^2{=}1 \mid D^*{=}1) = \mathbf{0.769}$





Consistent truth tellers represent **72%** of the NCDS sample

Share of consistent truth tellers amongst those correctly reporting their attainment in a given survey wave:

	No quals	**Any quals**
% of individuals reporting correctly in wave 1 who will also report correctly in wave 2	86.6	90.8
% of individuals reporting correctly in wave 2 who had also reported correctly in wave 2	91.8	94.8

- Figures from just one wave may not reveal behaviour (those with or without the qual have different response patterns over time)
- However *bulk of correct classification can be attributed to some degree of persistency in the reporting of individuals* across waves; remaining error (5-13pp) is not systematic.

Formal test of independent measurements

Assumption that self-reported measurements in the two surveys waves are conditionally independent given D^* and X: clearly rejected (positive autocorrelation)

CONSISTENT OVER-REPORTERS (among those with no quals)

$P(D_S^1=1, D_S^2=1 \mid D^*=0)$ = **0.196**

- Sizeable but looking at only one wave (27.1% at w1, 31.3% at w2) would overstate it.
- 28 to 30% of over-reporting errors in a given wave result from non-systematic recording errors

CONSISTENT UNDER-REPORTERS (among those with quals)

$P(D_S^1=0, D_S^2=0 \mid D^*=1)$ = **0.112**

- Focusing on one wave only would overstate amount of over-reporting (15.3% at w1, 18.9% at w2)
- 27 to 40% of under-reporting errors in a given wave result from non-systematic recording errors – almost identical to share accounting for over-reporting

CONFUSED

No qualifications: **0.172** Any qualification: **0.118**

15% of the NCDS sample

Group affected by **RECALL BIAS**

$P(D_S^1=1, D_S^2=0 \mid D^*=1)$ = **0.077**

$P(D_S^1=1, D_S^2=0)$ = **0.050**

Summary

	No quals	**Any quals**	**NCDS**
Truth tellers	0.631	0.769	0.719
Over-reporters	0.196		0.070
Under-reporters		0.112	0.072
Confused	0.172	0.118	0.153
Recall bias		0.077	0.050

(2) Returns to Academic Quals

Δ^*	0.264
Δ^*_{LFS}	0.378

	Transcript	**1981 wave**	**1991 wave**
Δ_{LFS}	0.332	0.333	0.293
p-value: $\Delta_{\text{LFS}}=\Delta^*$	*0.000*	*0.000*	*0.070*
Δ_{FULL}	0.194	0.194	0.151
p-*value*: $\Delta_{\text{FULL}}=\Delta^*$	*0.000*	*0.000*	*0.000*

Δ_{LFS} (ignore ability and misclassification) *vs* Δ^*

A. *based on reports close to attainment*

- omitted ability (Δ^*_{LFS} *vs* Δ^*): 43% ↑ bias
  misclassification (Δ_{FULL} *vs* Δ^*): 27% ↓ bias
- no evidence of balancing bias: large ↑ bias (26%)
- calibration rule: 0.80 Δ_{LFS}

B. *based on reports relying on recall (>10 years)*

- omitted ability (Δ^*_{LFS} *vs* Δ^*): 43% ↑ bias
  misclassification (Δ_{FULL} *vs* Δ^*): 43% ↓ bias
- $\Delta_{\text{LFS}} \approx \Delta^*$
- measurement error in recall information seems strong enough to compensate for omitted ability bias
- no need to calibrate returns from the LFS