Misreported Schooling, Multiple Measures and Returns to Educational Qualifications

Erich Battistin

University of Padova and Institute for Fiscal Studies

Barbara Sianesi

Institute for Fiscal Studies

October 11, 2006

Introduction

'**Categorical**' vis-*á*-vis '**continuous**' measures of education: more adequate in countries where different educational routes may have potentially very different returns on earnings (e.g. UK).

Extent of misclassification in educational attainment measures (Kane *et al.*, 1999, using US data):

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

Estimates of returns can be heavily affected: Kane *et al.* (1999), Black *et al.* (2000) and Lewbel (2006).

What We (Will Eventually) Do

- Estimate returns to educational qualifications in the UK allowing for misreported attainment (tricky because of non-classical error - an IV strategy does not do the job).
- Estimate the extent of measurement error in:
  - administrative information.
  - self-reported information very close to completion.
  - recall information 10 years later.
- Explore how measurement error and ability biases interact: calibration rules.
- Propose a semi-parametric estimation approach based on balancing scores and mixture models. When multiple independent reports are available resembles a control function approach with exclusion restrictions.

General Applicability: Some Examples

- Policy schemes in which participation (or eligibility) has to be obtained from survey respondents, who have often been shown to have poor recall or awareness of the kind of schemes they are in (see e.g. Bound *et al.*, 2001).
- Government schemes where the researcher cannot directly observe actual take-up and has to 'impute' the treatment status (e.g. eligibility to some means-tested benefits).
- Situations in which the treatment status is derived by splitting the sample based on an underlying continuous variable which is itself potentially measured with error (e.g. income or consumption to define poverty status, or firm size to define some form of eligibility).

Plan of the Talk

- General formulation of the problem: inferring the causal effect of a certain treatment (e.g. having or not a certain educational qualification) on the outcome of interest (earnings) when the treatment status is potentially mismeasured.
- Non-parametric identification of the returns with/without multiple measures of educational attainment:
  - Partial identification (only one measure).
  - Point-identification (with multiple measures).
- Mixture representation of the problem and semi-parametric estimation of returns when point identification is achieved.
- Preliminary results using data from the British National Child Development Survey.

Potential Outcomes Notation

Let

- D^* be an indicator for having the **qualification of interest**.
- (Y_1, Y_0) be the two **potential wages** from having and not having the qualification of interest, respectively.
- $Y = Y_0 + D^*(Y_1 - Y_0)$ be the **observed individual wage**.
- $Y_1 - Y_0$ be the **individual return** to achieving the qualification of interest.
- the "average treatment effect on the treated"(**ATT**) be

$$E_{Y_1|D^*}[Y_1|1] - \underbrace{E_{Y_0|D^*}[Y_0|1]}_{counterfactual},$$

corresponding to the *average return for those who have chosen to undertake the qualification of interest.*

General formulation of the problem

Let the *average return for those who have chosen to undertake the qualification of interest* be defined as follows:

$$\boxed{\Delta^* \equiv E_{Y_1 - Y_0 | D^*}[Y_1 - Y_0 | 1].}$$

- **Unconfoundedness**: conditional on a set of observables X, the educational choice D^* is *independent* of the potential outcomes:

$$f_{Y_i | D^*, X}[y | d^*, x] = f_{Y_i | X}[y | x], \quad i = 0, 1.$$

- **Common Support**: individuals with and without the qualification of interest can be found at all values of X, that is:

$$0 < P_{D^* | X}[1 | x] < 1, \quad \forall x.$$

Point-identification of Δ^* follows from raw data $\{Y, D^*, X\}$.

Identification with mismeasured qualifications

Allow data on educational attainment to be *mismeasured*, so that $\{Y, D_A, X\}$ is available instead of $\{Y, D^*, X\}$, with $D_A \neq D^*$.

Δ^* is not identified from raw data in general, the sign of the bias depending on the extent of misclassification:

$$\lambda_0^A(x) \equiv P_{D^*|D_A,X}[0|0,x], \qquad \lambda_1^A(x) \equiv P_{D^*|D_A,X}[1|1,x].$$

- **Non-Differential Misclassification**: any variable D_A which proxies D^* does not contain information to predict Y conditional on the true measure D^* and X:

$$f_{Y|D^*,D_A,X}[y|d^*,a,x] = f_{Y|D^*,X}[y|d^*,x].$$

Δ^* is a known functional of $\{Y, D_A, X, \lambda_0^A(x), \lambda_1^A(x)\}$ (Battistin and Sianesi, 2006), so only partial identification is feasible.

Identification with two measurements

Assume to observe *two* measurements of attainment (D_A, D_B) for each individual, both potentially mismeasured:

$$\lambda_0^j(x) \equiv P_{D^*|D_j,X}[0|0,x], \quad j \in \{A,B\},$$

$$\lambda_1^j(x) \equiv P_{D^*|D_j,X}[1|1,x], \quad j \in \{A,B\}.$$

Typically obtained by combining complementary datasets (e.g. administrative records and self-reported data) or longitudinal information from different interview waves.

- **Non-Differential Misclassification (extended)**: any variables (D_A, D_B) which proxy D^* do not contain information to predict Y conditional on the true measure D^* and X:

$$f_{Y|D^*,D_A,D_B,X}[y|d^*,a,b,x] = f_{Y|D^*,X}[y|d^*,x].$$

Identification with two measurements (cont'd)

The observed wage distribution conditional of X within cells defined by (D_A, D_B) has the following mixture representation:

$$\begin{aligned} f_{Y|D_A,D_B,X}[y|a,b,x] &= [1-\phi_{a,b}(x)]f_{Y_0|X}[y|x] + \phi_{a,b}(x)f_{Y_1|X}[y|x], \\ \phi_{a,b}(x) &\equiv P_{D^*|D_A,D_B,X}[1|a,b,x]. \end{aligned}$$

Conditional on X:

- **mixture weights** denote the proportion of individuals with the qualification of interest amongst those in the 2×2 cells defined by $(D_A = a, D_B = b)$.
- **mixture components** denote the distribution of potential wages from having / not having the qualification of interest.

Identification with two measurements (cont'd)

- By the law of total probability, knowledge of the $\phi_{a,b}(x)$'s suffices to identify the extent of misreporting in either measure of educational attainment – i.e. the $\lambda_0^j(x)$'s and the $\lambda_1^j(x)$'s.
- For example, non-parametric identification of the misclassification probabilities is achieved when errors in D_A and D_B are *independent* of each other (Lewbel, 2006, and Mahajan, 2006).
- Knowledge of the misclassification probabilities in turn implies that Δ^* is over-identified, since it can be written as a functional of $\{Y, D_A, X, \lambda_0^A(x), \lambda_1^A(x)\}$ or $\{Y, D_B, X, \lambda_0^B(x), \lambda_1^B(x)\}$ (see Battistin and Sianesi, 2006).

Estimation: Main Idea

Rests upon the *two-component* mixture distribution (the conditioning on X is left implicit)

$$f_{Y|D_A,D_B}[y|a,b] = [1-\phi_{a,b}]f_{Y_0}[y] + \phi_{a,b}f_{Y_1}[y],$$

which is identified if its components are a *linearly independent set over the field of real numbers* (Yakowitz and Spragins, 1968).

- Assume that potential wages are log-normally distributed, which in turn implies (parametric) identification of mixture weights and components (e.g. Everitt and Hand, 1981).
- Economic relevance (Heckman and Honorè, 1980).
- Variety of estimation procedures available – we use the EM algorithm (Dempster, Laird and Rubin, 1977).

Estimation: Comments

- D_A and D_B can be *dependent* reports, including the case of being one report only, i.e. $D_A \equiv D_B$; with independent reports resembles a control function approach with an exclusion restriction.
- Does <u>not</u> work under *homogeneity*, i.e. if

$$f_{Y_1|X}[y|x] \equiv f_{Y_0|X}[y|x],$$

though it allows the case of $\Delta^*[x]$ being zero, where

$$\Delta^*[x] \equiv E_{Y_1|X}[Y_1|x] - E_{Y_0|X}[Y_0|x].$$

- Following Yakowitz and Spragins (1968), if the assumption of non-degenerate $\phi_{a,b}$'s is maintained (which is consistent with past evidence in the literature) testing for log-normality of the mixture distribution is equivalent to testing for homogeneity.

Curse of Dimensionality

Estimation hampered by the large number of X's required to assures validity of the unconfoundedness assumption.

- Use the *balancing property* of propensity scores to reduce the dimensionality problem (Imbens, 2000, and Lechner, 2001).
- Show that the mixture representation keeps holding if the $\phi_{a,b}$'s are left to vary with X only through an *index* $\mathcal{S}(X)$ suitably defined from (D_A, D_B, X). Amounts to assuming that individuals with the same value of (D_A, D_B, X) share the same probability of having the qualification of interest.

Estimation Steps

1. Estimate **mixture weights** $\phi_{a,b}(s)$ by assuming log-normality of potential wages conditional on strata defined from $\mathcal{S}(X) = s$.
2. Obtain *individual-specific* **misclassification probabilities**:

$$\begin{aligned} \lambda_1^A(x_i) &= \sum_b \phi_{1,b}(s_i) P_{D_B|D_A,X}[b|1,x_i], \\ \lambda_1^B(x_i) &= \sum_a \phi_{a,1}(s_i) P_{D_A|D_B,X}[a|1,x_i], \end{aligned}$$

   where $P_{D_B|D_A,X}[b|1,x]$ and $P_{D_A|D_B,X}[a|1,x]$ are estimated from binomial regressions (similar procedure for the $\lambda_0^j(x)$'s).
3. Use results in Battistin and Sianesi (2006) to back out the **parameter of interest** Δ^*.

Data

Use data for $2,716$ working males from the 1958 UK NCDS cohort to estimate the *average return to any academic qualification.*

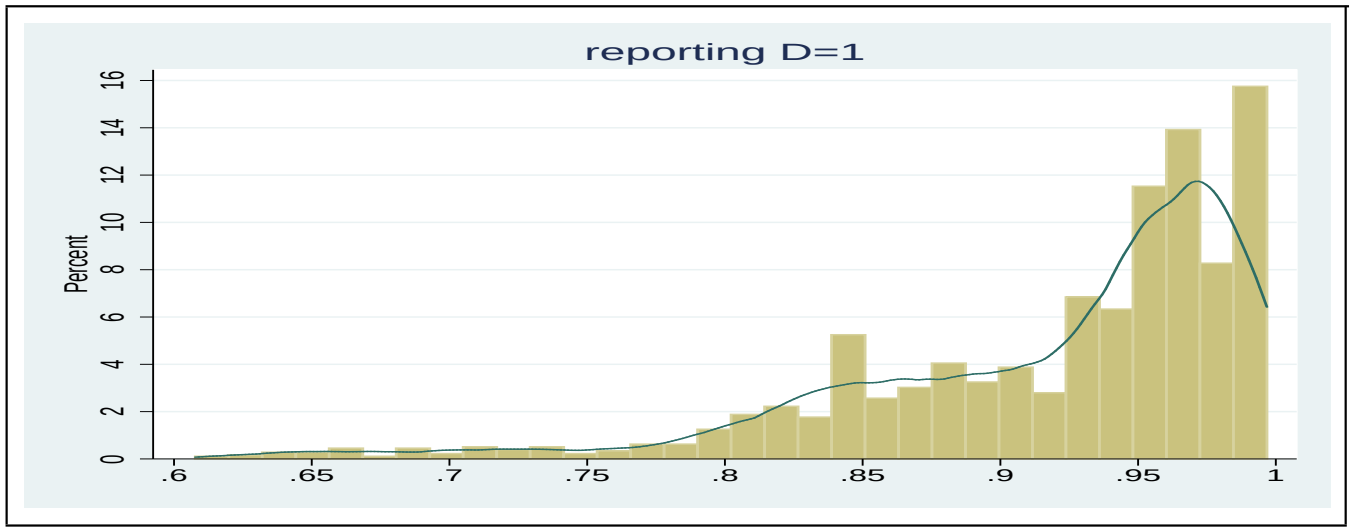
- Y: real gross hourly wage at age 33 (in 1991).
- D^*: dummy for *any academic qualification.*
- X: gender and age, ethnicity, region, math and reading ability tests at 7 and 11, mother's and father's education, mother's and father's age, father's social class when child was 16, mother's employment status when child was 16, number of siblings when child was 16, school type.
- Raw measures of educational attainment obtained from
  - self-reported data at age 23 (D_A).
  - exam transcripts (D_B).

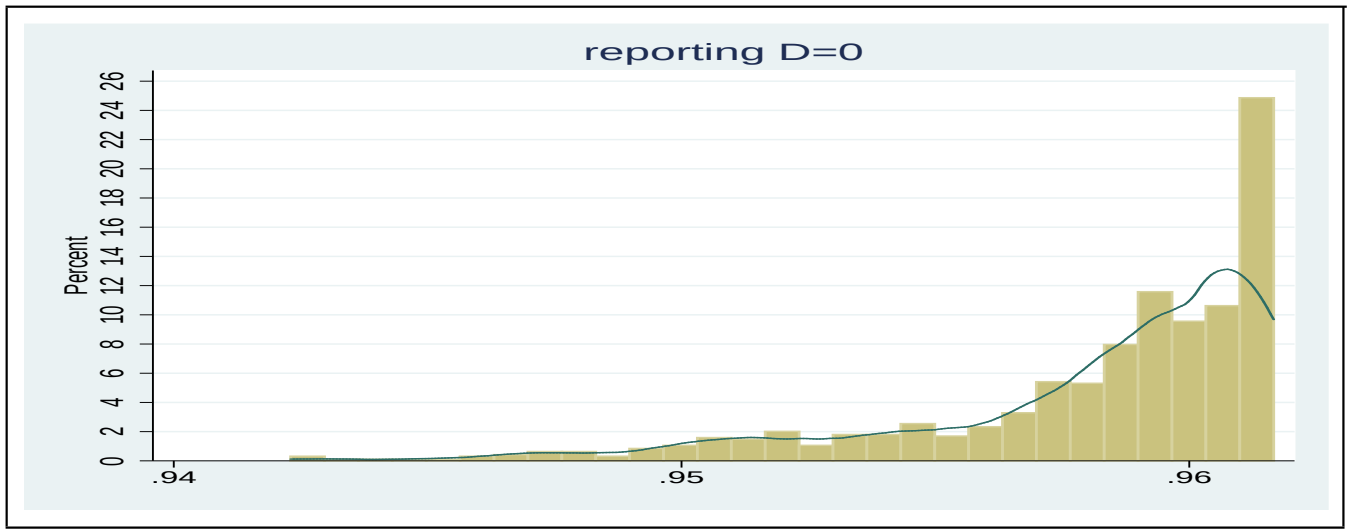
Data (cont'd)

Sample size: academic qualifications (cell frequencies in brackets)

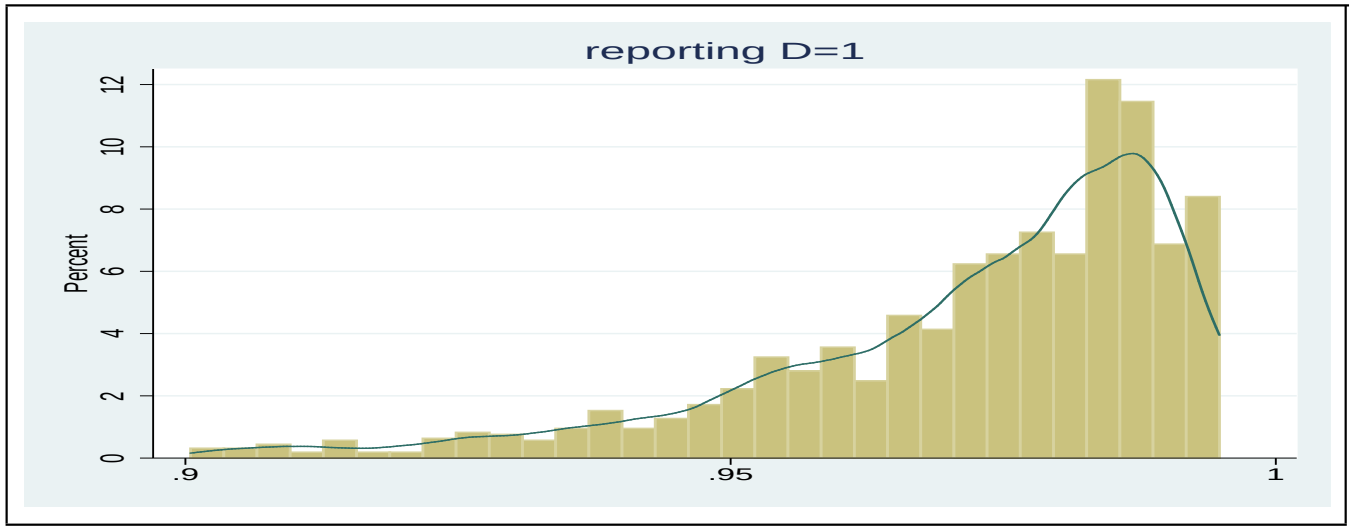
	Self-reported: D_A		
Admin: D_B	None	Any	
None	901 (33.2%)	218 (8.0%)	1,119 (42.2%)
Any	49 (1.8%)	1,548 (57.0%)	1,597 (58.8%)
	950 (35.0%)	1,766 (65.0%)	2,716 (100.0%)

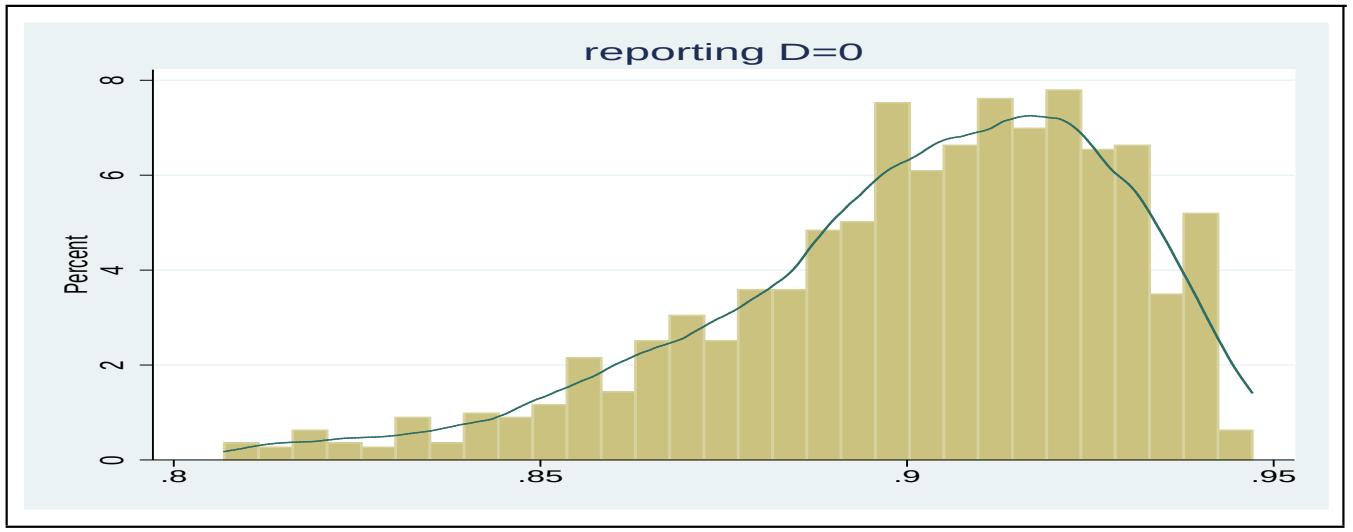
Results: Self-reported Data





Results: Administrative Data





Results: Extent of Misclassification

	Self-reported		Administrative	
	Mode	Std. Err.	Mode	Std. Err.
λ_1	0.970	0.119	0.988	0.120
λ_0	0.961	0.043	0.909	0.046

- No source appears to be uniformly better (in line with the little evidence available from the US).
- λ_0: individuals are 5.2 percentage points less likely than schools to under-report their qualifications.
- λ_1: individuals are almost 2 percentage points more likely than schools to over-report their attainment.
- Estimated $P[D^* = 1] \simeq 55\%$, while $P[D_A = 1] \simeq 65\%$ and $P[D_B = 1] \simeq 59\%$.

Results: Point Estimates of Returns

Parameter	Control for	Estimate	Std. Err.
Δ_{LFS}	nothing	0.329	0.015
Δ_{FULL}	ability bias	0.255	0.026
Δ^*_{LFS}	meas. err. bias	0.361	0.025
Δ^*_{FULL}	both	0.264	0.047

- Δ^*_{FULL} vs Δ_{LFS}: no evidence of balancing bias – ignoring both biases: large upward bias (6.5%).
- Calibration rule: 80% of Δ_{LFS}.
- Which bias matters most?
  - Δ^*_{FULL} vs Δ^*_{LFS} (ignore ability) $\Rightarrow$ 10% $\uparrow$ bias.
  - Δ^*_{FULL} vs Δ_{FULL} (ignore measurement error) $\Rightarrow$ 1% $\downarrow$ bias.