Journal of Econometrics 181 (2014) 136–150

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Misreported schooling, multiple measures and returns to educational qualifications☆

Erich Battistin [a,b,c], Michele De Nadai [d], Barbara Sianesi [e,*]

[a] Queen Mary University, UK
[b] IRVAPP, Italy
[c] IZA, Germany
[d] University of Padova, Italy
[e] Institute for Fiscal Studies, UK

ARTICLE INFO

Article history:
Received 17 September 2012
Received in revised form
3 March 2014
Accepted 8 March 2014
Available online 13 April 2014

JEL classification:
C10
I20
J31

Keywords:
Misclassification
Mixture models
Returns to educational qualifications
Treatment effects

ABSTRACT

We consider the identification and estimation of the average wage return to attaining educational qualifications when attainment is potentially measured with error. By exploiting two independent measures of qualifications, we identify the extent of misclassification in administrative and self-reported data on educational attainment. The availability of multiple self-reported educational measures additionally allows us to identify the temporal patterns of individual misreporting errors across survey waves. We provide the first reliable estimate of a highly policy relevant parameter for the UK, namely the return from attaining any academic qualification compared to leaving school at the minimum age without any formal qualification. We identify returns to qualifications under two alternative settings: a strong ignorability assumption and an exclusion restriction. All these results are obtained by casting the identification problem in terms of a mixture model, and using a semi-parametric estimation approach based on balancing scores, which allows for arbitrarily heterogeneous individual returns.

1. Introduction

The estimation of the return to education has probably become the most explored and prolific area in labour economics.[1] If a continuous years-of-schooling measure of education is affected by recording errors, standard results based on classical measurement error show that OLS estimates of the return to an additional year of education are downward biased, while appropriate IV methods applied to the linear regression model provide consistent estimates. If however a categorical qualification-based measure of education is affected by errors, any such error will necessarily vary with the true level of education, so that the assumption of classical measurement error cannot hold (see, for example Aigner, 1973). In this case, OLS estimates of the returns to qualifications are no longer necessarily downward biased, and the IV methodology cannot provide consistent estimates (see, for example Bound et al., 2001).

To date, empirical evidence on the importance of misreporting and returns to discrete educational levels is restricted to higher education in the US, where it was in fact shown that measurement error might play a non-negligible role (see Kane et al., 1999; Black

☆ This paper benefited from helpful discussions with Enrico Rettore and comments by audiences at Policy Studies Institute (London, September 2005), ADRES Conference on "Econometric Evaluation of Public Policies: Methods and Applications" (Paris, December 2005), Franco Modigliani Fellowship Workshop (Rome, February 2006), the 5th ESRC Research Methods Festival (Oxford, 2012) and European University Institute (Florence, 2013). Two editors and three referees further provided many insightful suggestions. Financial support from the ESRC under the research grant ES/I02574X/1 is gratefully acknowledged. Address for correspondence: Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK; School of Economics and Finance, Queen Mary, University of London (Mile End Campus), Mile End Road, London E1 4NS, UK and Department of Economics and Management, Via del Santo 33, 35123 Padova, Italy.

* Correspondence to: Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK. Tel.: +44 20 7291 48 00.

E-mail addresses: e.battistin@qmul.ac.uk (E. Battistin), denadai@stat.unipd.it (M. De Nadai), barbara_s@ifs.org.uk (B. Sianesi).

[1] Policymakers too have shown increasing interest, with estimated returns feeding into debates on national economic performance, educational policies, or the public funding of education. For an extensive discussion of the policy interest of the individual wage return from education, see Blundell et al. (2005a).

http://dx.doi.org/10.1016/j.jeconom.2014.03.002

et al., 2003; Lewbel, 2007). For the UK there are no estimates of the returns to educational qualifications that adequately correct for measurement error. This is of great concern, given the importance of focusing on discrete levels of educational qualifications[2] and given the widespread misconception amongst UK researchers and policymakers that the bias from measurement error (believed to be downward) and the so-called "ability bias" largely cancel each other out (Dearden, 1999; Dearden et al., 2002; McIntosh, 2006; Department for Business, Innovation and Skills, 2011).

This paper focuses on the returns to educational qualifications when attainment is potentially misrecorded[3] and offers a two-fold contribution. First, it provides the first reliable estimates of a highly policy relevant parameter for the UK, namely the return from attaining any academic qualification compared to leaving school at the minimum age without any formal qualification. Secondly, it estimates misclassification probabilities and patterns of misclassification, including the temporal correlations in misreporting by individuals across survey waves.

In order to overcome the bias introduced by misreported educational qualifications and to achieve point identification of the returns, (at least) two categorical reports of qualifications need to be available for the same individuals, both potentially affected by reporting error but coming from independent sources (for the proof of non-parametric identification, see Mahajan, 2006; Lewbel, 2007; Hu, 2008). Repeated measurements on educational qualifications are typically obtained by combining complementary datasets, for example exploiting administrative records and information self-reported by individuals. An additional appealing feature of this approach is that it provides estimates of the extent of misclassification in each educational measures, which is often of independent interest. As we will see, our case study employs additional measures on top of the minimum number required to achieve identification, thus allowing us to shed light on features of the measurement error process not unveiled in past studies.[4]

We focus on the return from attaining any academic qualification compared to leaving school at the minimum age of 16 without any formal qualification (the latter being akin to dropping out of high-school in the US). This return captures all the channels in which the initial decision to attain academic qualifications at the school-leaving age impacts on wages later on in life, in particular the contribution that attaining such qualifications gives to subsequent educational attainment. The policy relevance of this parameter for the UK is additionally highlighted by the finding that the main effect of changes in compulsory schooling was not to increase the length of schooling, but rather to induce individuals to leave school with an academic certification (Del Bono and Galindo-Rueda, 2004).

We rely on detailed longitudinal data for the male sample of the British National Child Development Survey (NCDS), which allows us to identify returns under two alternative settings. This data appears particularly suited to support the strong ignorability assumption that the observables are enough to control for the endogeneity of educational choices. This is because, in addition to detailed family background and school type variables, the NCDS contains extensive measures of both cognitive and non-cognitive traits at early ages. Under the strong ignorability assumption we also explore how the biases from measurement error and from omitted variables interact in the estimation of returns to educational qualifications, providing simple calibration rules that policy makers can apply to nationally representative datasets relying on self-reported qualifications and with no information on individual ability and family background (e.g. the Labour Force Survey). Alternatively, we identify returns for a specific group (the "compliers") exploiting an exclusion restriction.

Using the unique nature of our data we identify the extent of misclassification in *three* different data sources on educational qualifications: administrative school files, self-reported information very close to the dates of completion of the qualification, and self-reported recall information ten years later. To this end, we combine multiple measurements self-reported by individuals in the NCDS with administrative data on qualifications coming from school records. Compared to the existing articles in the literature, the availability of multiple self-reported measurements introduces a certain degree of over-identification, which allows us to isolate the extent of misreporting in school files from that of individuals, while allowing for persistence in the propensity to misreport across self-reported measurements. Thus, our setup gives us the unique chance of assessing the *temporal patterns* of misreporting errors across survey instruments and of decomposing misreporting errors into a systematic component linked to individuals' persistent behaviour and into a transitory part reflecting survey errors that occur independently of individuals in each cross-section survey wave.

On the methodological front we propose a semi-parametric estimation approach based on balancing scores and mixture models. We cast the estimation problem in terms of a mixture model, which combined with the propensity score defines a semi-parametric procedure that allows for arbitrarily heterogeneous individual returns. Given that the misclassification problem can be stated in terms of finite mixtures with a known number of components, we find this approach particularly suited for the case at hand (Hui and Walter, 1980, also propose an approach to misclassification of a dichotomous variable based on maximum likelihood). The general identification problem in the case of two reports has been considered, amongst others, by Kane et al. (1999), Black et al. (2000), Mahajan (2006), Lewbel (2007) and Hu (2008). We build upon these papers, and in particular upon Hu (2008), to show that the components of the mixture model are non-parametrically identified given the setup we consider. Specifically, we first show that all the quantities of interest are non-parametrically identified from the data through the availability of our repeated measurements on educational qualifications. The conditions required for this result are very general in nature, or at least are as restrictive as those commonly invoked in the relevant literature on misclassification. We then proceed with estimation, drawing from the statistical literature on finite mixtures to propose a flexible strategy based on Bayesian modelling.

We report a number of findings of substantive importance. All our results pertain to males only of the NCDS birth cohort. Individuals are found to be appreciably *less* accurate than transcript files when they do not have any academic qualification, but slightly

[2] In the UK educational system, individuals with the same number of years of schooling can have different educational outcomes; this not only obfuscates the interpretation of the return to one additional year, but imposing equality of yearly returns across educational stages was found to be overly restrictive (see Blundell et al., 2005b).

[3] Misrecorded education can arise from data transcript errors, as well as from misreporting, whereby survey respondents may either over-report their attainment, not know if the schooling they have had counts as a qualification or simply not remember.

[4] An alternative to dealing with misclassification bias which does not require repeated educational measures is to derive bounds on the causal effect of interest by making *a priori* assumptions on the misclassification probabilities (see, for example Kreider and Pepper, 2007; Molinari, 2008; Kreider et al., 2012). In previous work (Battistin and Sianesi, 2011) we based the analysis on one self-reported measure and could only provide partial identification of returns under strong ignorability. The bounds we suggested can be derived allowing for arbitrarily heterogeneous individual returns, while the corresponding sensitivity analysis is easy to implement and can provide an often quite informative robustness check. By contrast, in the current paper we achieve point identification not only of the returns, but also of the distribution of measurement error. Additionally, we discuss identification of policy effects under two scenarios that in our context seem appropriate (under strong ignorability and in the presence of a valid exclusion restriction).

more accurate than transcripts when they do in fact have academic qualifications. In line with the scant evidence available from the US, we thus find that no source is uniformly better. For individuals, over-reporting is by far the most important source of error. Under-reporting is more of a problem in transcript files. Notwithstanding their different underlying patterns of measurement error, transcript files and self-reported data appear to be remarkably similar in their overall reliability. This is especially so when information is collected close to the time of attainment of the educational qualification of interest. We estimate that the degree of accuracy in the reporting of educational qualifications in the NCDS is about 86% in both transcript files and self-reported data collected close to attainment of the qualification. This figure is 4 percentage points lower when educational attainment is recalled ten years later.

From estimating the share of individuals who consistently report correctly, over-report and under-report their educational qualification across survey waves of the NCDS, we find that figures from just one wave are not likely to reveal behaviour. Our results do however show that the bulk of correct classification can be attributed to some degree of persistency in the reporting of individuals across waves. We estimate that about 90% of measurement error in the NCDS is related to the behaviour of individuals; the remaining error is not systematic, and depends on random survey errors. We further provide strong evidence of positive autocorrelation in self-reported measurements conditional on true educational attainment. This finding in itself invalidates setups that base identification on repeated measurements by the same individuals. A piece of interesting evidence on survey errors is the incidence of recall errors among those with the qualification, which we estimate at 6.2%.

Our preferred estimate of the return from achieving any academic qualification for those who do so suggests a 26.7% wage gain. This figure is statistically different from that obtained from raw data without adjusting for measurement error. When educational records (from schools or individuals) are obtained relatively close to the completion of the qualification of interest, we find that ignoring both ability and misreporting biases would lead to strongly upward-biased estimates of returns. The resulting calibration rule to get an LFS-style estimate close to the true return suggests to multiply the "raw" estimate by 0.8. By contrast, when the educational information recorded in the data has been collected after over 10 years since completion, the two biases do seem to cancel each other out, with LFS-style estimates of the average return to academic qualifications being indeed very close to the true return.

The remainder of the paper is organized as follows. In Section 2 we allow for the possibility of misclassification in the treatment status in the general evaluation framework, and discuss identification of misclassification probabilities and of returns. Our estimation strategy is presented in Section 3. Section 4 discusses the informational content of the data, motivates the parameter of interest, and presents the evidence on raw returns and on the agreement rates between our multiple measurements. Section 5 presents our empirical results on the extent and features of misclassification, as well as on the true educational returns under the two alternative assumptions. We also explore how the biases from misclassification and from omitted variables interact in the estimation of such a return. Section 6 concludes. Proofs of identification that pertain to results presented in Section 3 are reported in Appendix A. A detailed description of the estimation method, as well as additional sensitivity checks are made available in the on-line appendix.

2. General formulation of the problem

In the potential-outcomes framework, interest lies in the causal impact of a given "treatment" on the Y. For reviews of the evaluation problem see, for example, Heckman et al. (1999), Imbens (2004) and Di Nardo and Lee (2011).[5] With our application in mind, in the following let the treatment be the qualification of interest and let the outcome be individual (log) wages. Let Y_1 and Y_0 denote the *potential* wages from having and not having the qualification of interest, respectively, and D^* the binary indicator for having such qualification.[6] The individual causal effect of (or return to) achieving the qualification is defined as $Y_1 - Y_0$. The *observed* individual wage can then be written as $Y = Y_0 + D^*(Y_1 - Y_0)$.

We will study settings in which average returns are identified from knowledge of the conditional expectations $E_{Y|D^*V}[Y|1, v]$ and $E_{Y|D^*V}[Y|0, v]$ for some observable variables V to be defined below.[7] We will consider two antipodes, one in which V is rich enough to control for the endogeneity of educational choices, and one in which V contains variables that affect the outcome only through their effect on educational choices. The former setting amounts to an *ignorability* assumption, while the latter sets out identification through *instrumental variables*. The two approaches are widely used in empirical work, and impose restrictions that define an identifying correspondence between data and average effects in the population. The analogue principle paves the way for estimation.

When realizations of misreported attainment are used in place of D^*, analogue estimators will typically deliver biased estimates of the average effect of interest (Bound et al., 2001). We address the identification problem in two different steps. First, using the availability of repeated measurements of educational attainment, we provide sufficient conditions to retrieve the quantities $E_{Y|D^*V}[Y|1, v]$ and $E_{Y|D^*V}[Y|0, v]$. We then use these quantities to estimate average returns under ignorability and with the availability of an exclusion restriction.

2.1. Misclassified educational qualifications

We assume the availability of *two* repeated measurements of educational qualifications self-reported by individuals at different points in time (D_S^1 and D_S^2), as well as of transcript records on the same individuals coming from the schools (D_T). The former two measurements need not be independent of each other, as they are most likely correlated through unobservables that affect the propensity of individuals to misreport.

For any measurement $W \in \{D_S^1, D_S^2, D_T\}$, define by $f_{W|D^*V}[1|1, v]$ the probability of correct self-reporting, or of correct classification in the transcript files, amongst those actually holding the qualification of interest. The corresponding probability amongst those without the qualification of interest is $f_{W|D^*V}[0|0, v]$. We refer to these terms as *probabilities of exact classification* for the measurement W. Similarly, letting $\mathbf{D}_S \equiv [D_S^1, D_S^2]$ denote the vector of self-reported measurements, define the probabilities $f_{\mathbf{D}_S|D^*V}[\mathbf{d}_S|1, v]$ and $f_{\mathbf{D}_S|D^*V}[\mathbf{d}_S|0, v]$ as the survey response patterns conditional on true educational attainment, separately for those having and not having the qualification of interest. The definitions employed accommodate for error heterogeneity through the observable characteristics V.

In the setting considered, data are informative on $(Y, \mathbf{D}_S, D_T, V)$. Throughout we maintain the assumption that the misclassification error in either measure is *non-differential*, that is conditional on a

[5] For the potential outcome framework, the main references are Fisher (1935), Neyman (1935), Roy (1951), Quandt (1972) and Rubin (1974).

[6] For this representation to be meaningful, the stable unit-treatment value assumption needs to be satisfied (Rubin, 1980), requiring that an individual's potential wages and the chosen qualification are independent of the qualification choices of other individuals in the population.

[7] A similar argument applies to identification of quantile treatment effects, by replacing conditional expectations with conditional distributions.

person's actual qualification and on other covariates, reporting errors are independent of wages (see Battistin and Sianesi, 2011, for a more detailed discussion of the implications of this assumption).

Assumption 1 (*Non-Differential Misclassification Given V*)**.** Any variables $\boldsymbol{D}_S$ and D_T which proxy D^* do not contain information to predict Y conditional on the true measure D^* and V:

$$f_{Y|D^*\boldsymbol{D}_S D_T V}[y|d^*, \boldsymbol{d}_S, d_T, v] = f_{Y|D^*V}[y|d^*, v].$$

It follows that the distribution of observed wages conditional on V for the $2 \times 2 \times 2$ groups defined by $D_S^1 \times D_S^2 \times D_T$ can be written as a *mixture* of *two* latent distributions:

$$f_{Y|\mathbf{D}_S D_T V}[y|\mathbf{d}_S, d_T, v] = [1 - p(\mathbf{d}_S, d_T, v)] f_{Y|D^*V}[y|0, v] + p(\mathbf{d}_S, d_T, v) f_{Y|D^*V}[y|1, v], \quad (1)$$

where the probability:

$$p(\mathbf{d}_S, d_T, v) \equiv f_{D^*|\mathbf{D}_S D_T V}[1|\mathbf{d}_S, d_T, v],$$

denotes the true proportion of individuals with the qualification of interest amongst those with $\mathbf{D}_S = \mathbf{d}_S$ and $D_T = d_T$ within cells defined by V. Knowledge of the mixture weights $p(\mathbf{d}_S, d_T, v)$'s suffices to identify the probabilities of *exact* classification relative to the self-reported measurements and transcript files.

2.2. Identification of misclassification probabilities

Key to our identification result, as to that in other papers, is the assumption that qualifications self-reported by individuals and transcript files are correlated only through the true measurement D^* (and the observables V).[8]

Assumption 2 (*Independent Sources of Error Given V*)**.** The measurements $\boldsymbol{D}_S$ and D_T are conditionally independent given D^* and V:

$$f_{\boldsymbol{D}_S D_T|D^*V}[\boldsymbol{d}_S, d_T|d^*, v] = f_{\boldsymbol{D}_S|D^*V}[\boldsymbol{d}_S|d^*, v] f_{D_T|D^*V}[d_T|d^*, v].$$

The general idea is to use D_T as a source of instrumental variation which, through Assumption 2, allows one to define a large enough number of moment conditions given the unknowns in the mixture representation (1). The availability of multiple self-reported measurements in our setting introduces a certain degree of over-identification, and allows us to isolate the extent of misreporting in school files from that of individuals while allowing for persistence in the propensity to misreport across self-reported measurements of educational qualifications. To the best of our knowledge, this is the first paper that looks into this problem.

The following additional assumptions are required for identification, and closely match those exploited in the relevant literature.

Assumption 3 (*Relevance of Educational Qualifications Given V*)**.** There is:

$$E_{Y|D^*V}[Y|1, v] \neq E_{Y|D^*V}[Y|0, v].$$

Intuitively this assumption is required to disentangle the mixture distributions in (1), and implies that the latent measurement D^* is relevant for the policy parameter under consideration (at all values V).[9] The next assumption requires that the measurement D_T contains enough information on the true educational qualification D^* given V or, more formally, that $f_{D^*|D_T V}[1|1, v] \neq f_{D^*|D_T V}[1|0, v]$ (see Chen et al., 2011). For the binary case considered in this paper, a sufficient condition for this to hold is the following.

Assumption 4 (*Informational Content of the Transcript Measurement Given V*)**.** The extent of misclassification in the measurement D_T is such that $f_{D^*|D_T V}[1|1, v] > 0.5$ and $f_{D^*|D_T V}[0|0, v] > 0.5$.

This assumption is indeed very reasonable, as it implies that information from the school files is more accurate than pure guessing once V is corrected for. Finally, a more technical assumption is needed to ensure identification, which is implied by a non-zero causal effect of the latent measurement D^* on the survey response patterns $\mathbf{D}_S$ given V.

Assumption 5 (*Relevance of Survey Instruments*)**.** For each value v on the support of V there is: $f_{\boldsymbol{D}_S D_T|V}[\boldsymbol{d}_S, d_T|v] \neq f_{\boldsymbol{D}_S|V}[\boldsymbol{d}_S|v] f_{D_T|V}[d_T|v]$.

The general identification result is summarized in the following theorem, the proof of which is given in Appendix A.

Theorem 1 (*Identification*)**.** *The mixture components* $f_{Y|D^*V}[y|0, v]$ *and* $f_{Y|D^*V}[y|1, v]$ *and the mixture weights* $p(\mathbf{d}_S, d_T, v)$ *are nonparametrically identified from the data* $(Y, \mathbf{D}_S, D_T, V)$ *under Assumptions* 1–5.

2.3. Identification of returns

We now address identification of returns from knowledge of the mixture components. We start by considering a setting in which the conditioning on a large set of observables is sufficient to retrieve the causal parameter of interest. As we shall discuss in the data section, the case study we consider makes this assumption rather plausible. We call these observables X, and we set $V \equiv X$ to state the following identifying restriction.[10]

Assumption 6 (*Strong Ignorability*)**.** Conditional on X, the educational choice D^* is independent of the two potential outcomes:

$$f_{Y_0, Y_1|D^*, X}[y_0, y_1|d^*, x] = f_{Y_0, Y_1|X}[y_0, y_1|x].$$

Moreover, individuals with and without the qualification of interest can be found at all values of X, that is:

$$0 < e^*(x) \equiv f_{D^*|X}[1|x] < 1, \quad \forall x$$

where $e^*(x)$ is the propensity score.

Under this assumption, the causal effect at $X = x$ can be retrieved by noting that $f_{Y|D^*X}[y|i, x] = f_{Y_i|X}[y|x]$ for $i \in \{0, 1\}$, so that

[8] As pointed out by Hu (2008) and Battistin and Sianesi (2011), the conditioning on a large set of V's makes both assumptions weaker than those reviewed in Bound et al. (2001).

[9] The requirement is stated in terms of conditional expectations. However, as we show in Appendix A, it could be formulated in more general terms by considering features of the conditional distribution $f_{Y|D^*V}[y|d^*, v]$. With our application in mind, sufficient conditions for Assumption 3 can be obtained from the idea of *Monotone Treatment Response* and *Monotone Treatment Selection* by Manski and Pepper (2000). The former assumption implies that each individual's wage is not decreasing in the educational level that may be potentially attained. The latter assumption states that, conditional on V, individuals who attain higher qualifications have wages that, on average, are not lower than those for individuals who attain lower qualifications. Taken together, the two assumptions imply $E_{Y|D^*V}[Y|0, v] \equiv E_{Y_0|D^*V}[Y_0|0, v] \leq E_{Y_1|D^*V}[Y_1|0, v] \leq E_{Y_1|D^*V}[Y_1|1, v] \equiv E_{Y|D^*V}[Y|1, v]$. Assumption 3 holds with at least one strict inequality sign in the last expression and conditional on a suitable set of observable characteristics V.

[10] The restriction is stronger than required, as restrictions on the conditional distribution of Y_0 would suffice to achieve identification of the policy parameter of interest.

average and quantile effects in the population can be obtained by integrating with respect to the distribution of X. Similarly, effects for individuals with the qualification of interest can be obtained by integrating with respect to the distribution of X in the $D^* = 1$ group. This distribution is identified from knowledge of the $p(\mathbf{d}_S, d_T, x)$'s. It follows that, under the conditions stated, one can retrieve causal parameters that — in the programme evaluation jargon — are referred to as average treatment effect (ATT) and the quantile treatment effect (QTE).

If Assumption 6 is not valid, identification of causal effects can be obtained under a valid exclusion restriction in the wage equation. We consider in what follows the case of a binary instrument Z for D^* and set $V \equiv [Z, X]$, where X denotes the same variables employed under Assumption 6. The formulation of the assumptions required builds upon the potential outcome framework specialized to instrumental variables by Imbens and Angrist (1994). This needs some refinements to the notation employed above. Accordingly, for the attainment dummy D_z^* and the potential outcomes $Y_{0,z}$ and $Y_{1,z}$ we make explicit the dependence on the realized value z of the instrument Z.

Assumption 7 (*Instrumental Variables*)**.** The following conditions hold conditional on X:

1. Z is as good as randomly assigned (independence condition):
$$f_{Y_{0,0}Y_{0,1}Y_{1,0}Y_{1,0}D_0^*D_1^*|Z}[\mathbf{y}_0, \mathbf{y}_1, \mathbf{d}^*|z] = f_{Y_{0,0}Y_{0,1}Y_{1,0}Y_{1,0}D_0^*D_1^*}[\mathbf{y}_0, \mathbf{y}_1, \mathbf{d}^*].$$
2. Potential outcomes do not depend on Z, that is $Y_{0,z} = Y_0$ and $Y_{1,z} = Y_1$ (exclusion restriction).
3. $f_{D^*|Z}[1|z]$ is a non-degenerate function of Z (rank condition).
4. by changing from $Z = 0$ to $Z = 1$, no individual would shift from having the qualification of interest to not having it (no defiers assumption).

Under Assumption 7 and conditional on X, an important result derived by Imbens and Rubin (1997, Eqs. 5 and 6) is that:

$$f_{Y_0^c}[y] = \frac{\phi_n + \phi_c}{\phi_c} f_{Y|D^*Z}[y|0, 0] - \frac{\phi_n}{\phi_c} f_{Y|D^*Z}[y|0, 1],$$
$$f_{Y_1^c}[y] = \frac{\phi_a + \phi_c}{\phi_c} f_{Y|D^*Z}[y|1, 1] - \frac{\phi_a}{\phi_c} f_{Y|D^*Z}[y|1, 0],$$

where (Y_0^c, Y_1^c) are potential outcomes for *compliers* (c), and ϕ_c, ϕ_a and ϕ_n are the shares of *compliers* (for whom $D_1^* > D_0^*$), *always takers* (a — those for whom $D_1^* = D_0^* = 1$) and *never takers* (n — those for whom $D_1^* = D_0^* = 0$) in the population. Note that such shares are identified from the conditional distribution of D^* given Z, since there is $\phi_n = f_{D^*|Z}[0|1]$, $\phi_a = f_{D^*|Z}[1|0]$ and $\phi_c = 1 - \phi_n - \phi_a$. These distributions are all identified from knowledge of the mixture weights. It therefore follows that average or quantile effects for compliers, namely those individuals responding in their choice of D^* to a change in Z, are easily obtained as functionals of identified quantities.

3. Estimation

Having derived the conditions for non-parametric identification of causal parameters and features of the error distribution across measurements, we now describe the strategy employed in the empirical section to estimate the quantities of interest. Two assumptions are maintained throughout. These impose restrictions on the heterogeneity of mixture weights and components on the one hand, and on the distribution of mixture components on the other.

As for restricting heterogeneity, we start by constructing functions of X using the concept of *balancing score*. Let $\mathcal{S}(X)$ be a balancing score such that the distribution of X within cells defined by $\mathcal{S}(x)$ is independent of $(\mathbf{D}_S, D_T)$:

$$f_{X|\mathbf{D}_S D_T \mathcal{S}(X)}[x|\mathbf{d}_S, d_T, s] = f_{X|\mathcal{S}(X)}[x|s].$$

To make the definition of $\mathcal{S}(X)$ operational, let G be a multinomial variable identifying the $2 \times 2 \times 2$ groups obtained from the cross tabulation of $(\mathbf{D}_S, D_T)$. Define the *propensity scores* obtained from the multinomial regression of G on the X's as $e_g(x) \equiv f_{G|X}[g|x]$ and $g = 1, \ldots, 8$. Results in Imbens (2000) and Lechner (2001) can be directly applied to conclude that the $e_g(x)$'s are balancing scores for $(\mathbf{D}_S, D_T)$. In words, this implies that individuals sharing the same vector of $e_g(x)$'s but characterized by different combinations of $(\mathbf{D}_S, D_T)$ are compositionally identical with respect to the vector of variables X. We impose in estimation that mixture components and mixture weights depend on X solely through the vector of the $e_g(x)$'s.[11]

Our second assumption is that the mixture components are normally distributed (see Heckman and Honore, 1990, for a similar approach). Given that the misclassification problem can be stated in terms of finite mixtures with a known number of components, we find this approach particularly suited for the case at hand as any finite mixture of univariate normal distributions is identifiable (see, for example Everitt and Hand, 1981).[12] Specifically, we proceed by assuming that, within cells defined by $\mathcal{S}(x)$, mixture components are normally distributed with cell-specific parameters. This amounts to assuming log-normality of wages conditional on D^*, the balancing score and – under Assumption 7 – the instrument Z: given the nature of the outcome variable, this appears to be a sound specification for the case at hand.

By denoting with $\mathbf{e}(x) = [1, e_2(x), \ldots, e_8(x)]'$ the 8×1 vector of balancing scores, under Assumption 7 we set:

$$f_{Y|D^*XZ}[y|i, x, z] \sim N(\boldsymbol{\theta}_i' \mathbf{e}(x) + \boldsymbol{\alpha}_i' \mathbf{e}(x) z, \sigma_i^2), \quad i = 0, 1$$
$$p(\mathbf{d}_S, d_T, x, z) = \Phi(\boldsymbol{\gamma}_g' \mathbf{e}(x) + \boldsymbol{\delta}_g' \mathbf{e}(x) z), \quad g = 1, \ldots, 8$$

where $\Phi(\cdot)$ is the standard normal distribution function. The former equation defines the 8×1 vectors of parameters $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$, and the scalars σ_0^2 and σ_1^2. The latter equation defines the 8×1 vector of parameters $\boldsymbol{\gamma}_g$ indexed to the group defined by $D_T \times D_S^1 \times D_S^2$. Under Assumption 6, changing the conditioning set is equivalent to imposing $\boldsymbol{\alpha}_i = 0$, $i = 0, 1$ and $\boldsymbol{\delta}_g = 0$, $g = 1, \ldots, 8$.

The mixture is estimated through a MCMC procedure, which is fully documented in the online appendix (see Appendix C). The mixture is estimated using two different conditioning sets: $V \equiv X$ under Assumption 6, which is our preferred option, and $V \equiv [Z, X]$ under Assumption 7. Our specification defines 82 unknowns under Assumption 6, and 162 unknowns under Assumption 7. We specify a joint prior distribution for these parameters, and we use a Gibbs sampling algorithm to obtain 2000 realizations from their joint posterior distribution. The posterior distributions for the unknown

[11] We show in the working paper version of this article (Battistin et al., 2011) that the latter restriction on weights together with Assumption 6 ensure that the mixture representation given X in (1) implies a mixture representation given $\mathcal{S}(X)$. Because of this, at least under Assumption 6 which is rather plausible for the case at hand, the restriction imposed on the mixture components is substantially weak.

[12] Perhaps the most natural and intuitive way of addressing the identification problem for mixtures of parametric distributions is found in Yakowitz and Spragins (1968), who show that a necessary and sufficient condition for the mixture to be identifiable is that the mixture components be a linearly independent set over the field of real numbers. This condition is met for the case of mixtures of normal distributions. Using the result by Yakowitz and Spragins (1968), it follows that our estimation procedure could be extended to more general families of parametric distributions. We further relax the normality assumption in the online appendix (see Appendix C), where we adopt an estimation method that does not rely on the normality of mixture components. Reassuringly, the results do not seem to be overly sensitive to the estimation method employed.

quantities of the mixture can easily be computed using these realizations. Knowledge of these quantities is in turn sufficient to obtain estimates of the misclassification probabilities and causal parameters discussed in Section 2.

4. Data and educational qualifications of interest

4.1. Data and sample

In our empirical analyses we use the National Child Development Survey (NCDS), a detailed longitudinal cohort study of all children born in Great Britain a week in March 1958. This dataset is particularly suited to implement the methods we propose, which rely either on strong ignorability (Assumption 6) or on an exclusion restriction (Assumption 7). For the former assumption we require very rich background information capturing all those factors that jointly determine the attainment of educational qualifications and wages. The NCDS data are quite unique in this respect, as in addition to detailed family background variables when the child was 16 (mother's and father's education and age, father's social class, mother's employment status and number of siblings) and school type variables (comprehensive, secondary modern, grammar, private and other school), they contain extensive cognition and personality tests at early ages. Specifically, at ages 7 and 11 respondents were administered a number of achievement tests (assessing math and verbal ability at 7 and math, reading, general verbal ability and general non-verbal ability at 11), as well as two widely used psychological scales measuring various aspects of child emotional and behavioural maladjustment (the Bristol Social Adjustment Guide, completed by the child's teacher, and the Rutter Behaviour Scale, completed by the child's parent). The unusual richness of the set of control variables can be used to motivate Assumption 6.

Alternatively, returns to education for specific groups can be identified based on an exclusion restriction. Again exploiting the informational richness of the NCDS, we use a combined measure of parental interest in the child's education. This dichotomous variable, used previously by Blundell et al. (2005b) to instrument participation in higher education, is equal to 1 if, according to the child's teacher, the mother and/or the father were very or overly interested in the education of their 7-year old child, and equal to 0 if both the father and the mother were judged by the teacher to have some or little interest in their child's education.

Our outcome is real gross hourly wages at age 33. To avoid the additional issue of selection into employment, we focus all our empirical analyses on males, further restricting attention to those in work (and with wage information) in 1991 and for whom neither of the three educational measure is ever missing.[13] These criteria leave us with a final sample of 2716 observations, which is the same sample used by Battistin and Sianesi (2011). Detailed summary statistics for the variables employed in our empirical exercise are reported in Appendix B.

4.2. Educational qualifications of interest

A highly policy relevant parameter, and the one we focus on in our application, is the return from attaining any academic qualification compared to leaving school at the minimum age of 16 without any formal qualification.[14] In the UK, at the time our sample members were making their education choices, the minimum level of academic qualification was the attainment of Ordinary levels (O levels). Special interest in estimating the returns to obtaining at least O levels arises from the finding that in the UK, reforms raising the minimum school leaving age have impacted on individuals achieving low academic qualifications, in particular O levels. Specifically, Chevalier et al. (2004) show that the main effect of the reform was to induce individuals to take O levels. Del Bono and Galindo-Rueda (2004) similarly show that changes in features of compulsory schooling have been biased towards the path of academic attainment; the main effect of the policy was not to increase the length of schooling, but rather to induce individuals to leave school with an academic certification. In such a context it is thus of great policy interest to estimate the returns to finishing school with O levels compared to leaving with no qualifications. Furthermore, looking as we do at the return to acquiring *at least* O levels compared to nothing captures all the channels in which the decision to attain O levels impacts on wages later on in life, in particular the potential contribution that attaining O levels may give to the subsequent attainment of Advanced levels (A levels) and then of higher education.[15]

As discussed below, the focus on academic qualifications such as O levels is driven by the availability of an independent school measure. They do however also offer clear advantages, in addition to their policy interest. First, O levels are well defined and homogeneous, with the central government traditionally determining their content and assessment. By contrast, the provision of vocational qualifications is much more varied and ill-defined, with a variety of private institutions shaping their content and assessment.[16] A second advantage of focusing on O levels is that they are almost universally taken through uninterrupted education, whereas vocational qualifications are often taken after having entered the labour market. It is thus more difficult to control for selection into post-school (vocational) qualifications, since one would ideally want to control also for the labour market history preceding the acquisition of the qualification.

4.3. Educational measurements in the NCDS

Non-parametric identification of misclassification probabilities requires access to at least two independent measurements of educational attainment (in the sense explained in Section 2.2). In the NCDS data, such measurements are offered by self-reported attainment and, for academic qualifications achieved by age 20, also by the School Files. Specifically, in 1978 the schools cohort members attended when aged 16 provided information on the results of public academic examinations entered up to 1978 (i.e. by age 20).[17]

Of particular interest to our purposes is that cohort members were asked twice to report the qualifications they had obtained as

[13] It is reassuring to note that the patterns that emerge from the following tables are the same irrespective of whether the sample is selected on the basis of non-missing educational information ever or non-missing wage information in 1991 (the latter obviously also restricting attention to those employed in 1991).

[14] Although the British system is quite distinct from the one in the US, one could regard the no-qualifications group as akin to the group of high-school drop-outs.

[15] In the British educational system, those students deciding to stay on past the minimum school leaving age of 16 can either continue along an academic route or else undertake a vocational qualification before entering the labour market. Until 1986, pupils choosing the former route could take O levels at 16 and then possibly move on to attain A levels at the end of secondary school at 18. A levels still represent the primary route into higher education.

[16] In fact, there is a wide assortment of options ranging from job-specific, competence-based qualifications to more generic work-related qualifications, providing a blend of capabilities and competencies in the most disparate fields.

[17] Similar details were collected from other institutions if pupils had taken such examinations elsewhere. Results were obtained for approximately 95% of those whose secondary school could be identified.

Table 1
Estimates of the returns to any academic qualification ignoring potential misclassification.

	(1)	(2)	(3)	Tests of equality		
	Transcript files from schools	1981 Wave (at age 23)	1991 Wave (at age 33)	(1) = (2)	(1) = (3)	(2) = (3)
LFS controls only:						
OLS	0.332	0.333	0.293		***	***
	(*0.016*)	(*0.016*)	(*0.015*)			
FILM	0.330	0.336	0.289		***	***
	(*0.016*)	(*0.016*)	(*0.014*)			
PSM	0.331	0.336	0.285		***	***
	(*0.015*)	(*0.017*)	(*0.015*)			
Full set of controls:						
OLS	0.174	0.174	0.131		***	***
	(*0.019*)	(*0.020*)	(*0.017*)			
FILM	0.199	0.220	0.123		**	***
	(*0.031*)	(*0.030*)	(*0.026*)			
PSM	0.204	0.245	0.120		***	***
	(*0.027*)	(*0.029*)	(*0.026*)			
IV	0.203	0.173	0.280			
	(*0.327*)	(*0.270*)	(*0.643*)			

Note. Reported are estimates of the average treatment effect on the treated (ATT) obtained by controlling only for the LFS set of variables (age, ethnicity and region) or the full set of variables including LFS-controls plus both cognitive and non-cognitive ability scores 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 and school type. Standard errors are in parentheses. Estimation methods considered are ordinary least squares (OLS), fully interacted linear matching (FILM), propensity score kernel matching (PSM). Also presented in the last row are 2SLS estimates of returns (IV), separately for the three measurements of educational attainment (see Section 4.4 for details). *Right-hand side panel*: the corresponding columns are significantly different at the 99% (***), 95% (**) and 90% (*) level, based on bootstrapped bias-corrected confidence intervals.

of March 1981 (i.e. by age 23): in the contemporaneous 1981 Survey (at age 23), as well as in the 1991 Survey (at age 33).[18] We can thus construct two separate self-reported measures of qualifications obtained up to March 1981, based either on responses in the 1981 or in the 1991 survey.

O level attainment is recorded by the schools by the time the individuals were aged 20, while it is self-reported by individuals by the time they were aged 23. In order to have repeated measurements of O level achievement by age 20 coming from both school records and NCDS survey reports, we thus need O level qualifications to be completed by age 20. The UK educational system is indeed such that O levels are obtained before age 20, with the official age being 16.[19]

For each individual we thus have three measurements, which can all be taken as proxies of having achieved any academic qualification by age 20. These are the measurements we use to implement the strategy described in Section 3.

4.4. Evidence from the raw data

Table 1 presents 'raw' estimates of the average wage return to any academic qualification, that is ignoring any potential misclassification of the educational indicator. Estimates are derived using:

- four different methods (simple dummy variable OLS, fully interacted regression model and propensity score matching relying on strong ignorability, as well as standard 2SLS relying on an exclusion restriction);
- two sets of control variables for the methods based on strong ignorability (the full set of observables including cognitive and non-cognitive ability and family background measures, and a subset mimicking what is available in Labour Force-style datasets); and
- most importantly for the aims of this paper, our three alternative education measures (school transcripts, self-reports at 23 and self-reports at 33).

We start by discussing the results of the models estimated under strong ignorability, noting that for those which allow for observed heterogeneity in returns (fully interacted regression and propensity score matching), the table reports estimates of the average return to academic qualifications for those who did acquire them (ATT). As in Blundell et al. (2005b), we find that while results change little in response to the method used to control for selection on observables, controlling for early cognitive and non-cognitive ability as well as family background measures is crucial, and significantly reduces the estimated return by between 27% and 58% to a 12% to 25% wage gain depending on the educational measure used. As to the latter, it is indeed striking that using an educational measure rather than another often gives rise to returns which exhibit the same magnitude of bias as from omitted controls. This is the case when comparing estimates using self-reported measures at different times (returns based on the concomitant measure being between 25% and 104% larger than those based on the recall measure) and when comparing estimates based on transcript vs recall information (returns based on recall being between 33% and 70% smaller than estimates based on school records). Estimates arising from measures obtained close to completion (i.e. transcript and self-reports at age 23) are by contrast not statistically different.

Our 2SLS models exploit the parental variable defined in Section 4.1.[20] Parental interest is found to be an important determinant of acquiring academic qualifications even conditional on our rich set of observables, particularly for the contemporaneous measures (the first-stage F-test statistic being 12 and 17 for the school measure and contemporaneous self-report, dropping to 6 for the later self-report). In the presence of return heterogeneity, IV models estimate a Local Average Treatment Effect (LATE), i.e. the average return for those who 'comply' with the instrument. For the two contemporaneous measures, the 'raw' IV point estimates for the compliers are in the range of the corresponding 'raw' return

[18] After having been asked about qualifications obtained since March 1981, cohort members were asked to "help us check our records are complete" in two steps. First, they had to identify on a card all the qualifications they had obtained in their lives (including any they had just told the interviewer about), and subsequently they had to identify any of these that had been obtained *before* March 1981.

[19] Indeed, in the NCDS only 5.7% of the O levels self-reported by the individuals at age 23 are reported to have been obtained after leaving school, and only a negligible share (1.3%) is self-reported to have been completed after age 20.

[20] This variable has been used as an instrument also by Blundell et al. (2005b) when looking at the returns to higher education; we discuss its validity in Section 5.3.2.

Table 2
Cross tabulation of the indicators of educational attainment (transcript files from schools and self-reported information from individuals at age 23 and at age 33; $N = 2716$).

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

Accordance between 1981 wave and 1991 wave ($\kappa_{D_S^1, D_S^2}$): 0.745
Accordance between 1981 wave and transcripts ($\kappa_{D_S^1, D_T}$): 0.792
Accordance between 1991 wave and transcripts ($\kappa_{D_S^2, D_T}$): 0.692
Accordance across all indicators ($\kappa_{D_T, D_S^1, D_S^2}$): 0.743

Note. Reported is the sample size of the $2 \times 2 \times 2$ cells defined from the cross tabulation $D_S^1 \times D_S^2 \times D_T$, where each indicator is a dummy variable for having *any* academic qualification vis-à-vis having *none*. For example, 781 is the number of individuals who, according to all measurements available, have no academic qualification at age 20. Also reported is the Fleiss's (1971) kappa coefficient of accordance (see Section 4.4 for further details) for the pairs (D_S^1, D_S^2), (D_S^1, D_T) and (D_S^2, D_T), and for the triple (D_S^1, D_S^2, D_T).

for the treated under strong ignorability: a 20% return based on the school measure and a 17% return based on the self-report close to completion. By contrast, the 'raw' IV estimate of 28% based on recall is much higher than the corresponding estimates of 12%–13% found under ignorability. Having moved to an IV framework has however imposed a very heavy price in terms of precision, as despite the good first stage none of the estimates is nowhere near statistical significance.

To investigate the substantial differences in estimated returns according to the educational report being used in the models based on selection on observables, Table 2 presents cross tabulations between the three underlying measurements. We find that the percentage of the sample where all three measures agree is 82%. Despite this reasonably high "agreement rate", there are still important differences between the information contained in the reports. Of particular interest for our results, the incidence of academic qualifications in the population is 58.8% according to transcript information, whilst according to self-reports it is considerably higher, around 65% in both interviews.

If we were to believe the school files, only 3.1% of those students who did achieve O-levels reported to have no academic qualifications at age 23. At age 33, when asked to recall the qualifications they had attained by age 23, individuals are observed to make more mistakes, with 8% of O-level achievers "forgetting" their attainment. Conversely, still taking the school files at face value, it appears that almost one fifth of those with no formal qualifications over-report their achievement when interviewed at age 23. As was the case with under-reporting, over-reporting behaviour seems to worsen when moving further away from the time the qualification was achieved. When relying on recall information, almost one fourth of individuals with no formal qualifications state to have some.

The highest agreement rates are observed between transcript files and self-reported information close to completion (an agreement rate of 90% and a kappa-statistic of 0.792[21]), while the lowest are found between transcript information and self-reported information based on recall (an agreement rate of 85% and a kappa-statistic of 0.692). The degree of congruence in information provided by the same individual 10 years apart falls in the middle (an agreement rate of 88% and a kappa-statistic of 0.745). The kappa statistics show a degree of agreement that Landis and Koch (1977) would view as substantial (kappa between 0.61–0.80).[22]

In conclusion, even though formal statistics like the kappa measure of inter-rater agreement may show that there is substantial agreement between educational measures, we have seen that remaining divergences in the resulting treatment indicators can lead to substantially and significantly different impact estimates — indeed of the same magnitude as not controlling for the rich set of variables available in the NCDS. Furthermore, taking the school files at face value, there appears to be much more over- than under-reporting, and reporting errors seem to get worse when individuals are asked to recall their qualifications. While it appears natural to take the school files as being closer to the "truth", this is however by no means an *a priori* correct assumption, and one which we assess empirically in the next section.

5. Results

5.1. Summary of the quantities retrieved

To ease readability in the discussion of the quantities that characterize misreporting, in the following the conditioning on observables V is left implicit. It is however important to highlight that the features of measurement error have been identified and estimated based on an unusually rich set of observed characteristics, including in particular an array of cognitive and non-cognitive ability measures (see Section 4.1 for a description of the data).

For each measurement $W \in \{D_S^1, D_S^2, D_T\}$, we consider the two probabilities of *exact classification* $f_{W|D^*}[0|0]$ and $f_{W|D^*}[1|1]$ (see Section 2.1 for their definition). Similarly, we define the percentage of *over-reporters* as $1 - f_{W|D^*}[0|0]$, and the percentage of *under-reporters* as $1 - f_{W|D^*}[1|1]$. For each measurement W, the probability of *correct classification* (equivalent to the event $W = D^*$) can be computed by averaging the two probabilities of exact classification:

$$f_{W|D^*}[0|0](1 - f_{D^*}[1]) + f_{W|D^*}[1|1]f_{D^*}[1].$$

The *extent of misclassification* in the measurement W is defined as one minus this quantity. Estimates of these quantities are presented in Table 3.[23]

The availability of repeated measurements coming from the same individuals allows us to define more structural parameters that reveal the individuals' propensity to misreport across waves. Errors in one survey wave are the result of purposive misreporting of individuals, or simply of survey errors that may occur independently of individual behaviour. These are substantially different sources of error, and so are their implications for the design of survey instruments aimed at recording educational attainment. We therefore focus on four different types of individuals. Consistent *truth tellers* are defined from the event $D_S^1 = D^*, D_S^2 = D^*$, namely as those individuals who self-report correctly their educational attainment across survey waves. They are made up of two groups: consistent truth tellers amongst those with the qualification (their share given by $f_{\mathbf{D}_S|D^*}[1, 1|1]$) and consistent truth tellers amongst those without the qualification (their share given by $f_{\mathbf{D}_S|D^*}[0, 0|0]$).

[21] The kappa-statistic measure of inter-rater agreement is scaled to be zero when the amount of agreement is what would be expected to be observed by chance and one when there is perfect agreement (see Fleiss, 1971).

[22] From a descriptive analysis of the determinants of concordance across indicators of educational attainment (see Mellow and Sider, 1983), our observed characteristics were found to have a very low power in predicting agreement rates.

[23] According to the procedure discussed in Section 3, we estimated the mixture model (1) under two alternative scenarios. Under Assumption 6, we set $V \equiv X$, while under Assumption 7 we impose $V \equiv [Z, X]$. We thus have two estimation sets for misclassification probabilities depending on the choice for V. The main text discusses results that were obtained imposing $V \equiv X$, which is our preferred choice. Results obtained under $V \equiv [Z, X]$ were qualitatively similar, and reported in the online appendix (see Appendix C).

Table 3
Probabilities of exact classification across survey instruments (transcript files from schools and self-reported information from individuals at age 23 and at age 33).

	Transcript files from schools	1981 Wave (at age 23)	1991 Wave (at age 33)
Probabilities of exact classification by recorded attainment:			
Any qualification	0.776	0.844	0.803
	(*0.042*)	(*0.045*)	(*0.039*)
No qualification	0.773	0.666	0.623
	(*0.075*)	(*0.064*)	(*0.059*)
Correct classification	0.868	0.863	0.82
	(*0.034*)	(*0.03*)	(*0.028*)

Note. The table presents estimates of the probabilities of exact classification for the three survey instruments. *Top Panel*: the row labelled *Any qualification* reports estimates for $f_{D_T|D^*}[1|1]$, $f_{D_S^1|D^*}[1|1]$ and $f_{D_S^2|D^*}[1|1]$, respectively; the row labelled *No qualification* reports estimates for $f_{D_T|D^*}[0|0]$, $f_{D_S^1|D^*}[0|0]$ and $f_{D_S^2|D^*}[0|0]$, respectively. *Bottom Panel*: estimates of the probabilities of correct classification obtained by averaging the two probabilities of exact classification (see Section 5.1 for definitions). Posterior standard deviations are reported in parentheses.

The percentage of these individuals can be computed as:

$$f_{\mathbf{D}_S|D^*}[0,0|0](1-f_{D^*}[1])+f_{\mathbf{D}_S|D^*}[1,1|1]f_{D^*}[1],$$

thus averaging probabilities that involve the survey response patterns. Similarly, one can define consistent *over-reporters* ($D_S^1 > D^*$, $D_S^2 > D^*$, their share being given by $f_{\mathbf{D}_S|D^*}[1,1|0]$), consistent *under-reporters* ($D_S^1 < D^*, D_S^2 < D^*$, their share being given by $f_{\mathbf{D}_S|D^*}[0,0|1]$) and the residual group of *confused* ($D_S^1 = 1-D^*$, $D_S^2 = D^*$ or $D_S^1 = D^*, D_S^2 = 1-D^*$), namely individuals with inconsistent response behaviour across survey waves. Estimates of these quantities are presented in Table 4. The comparison between the percentage of truth tellers, on the one hand, and the percentage of correct classification in each survey wave, on the other hand, reveals how much the latter results from behavioural attitudes of respondents or from survey errors.

Finally, we define the probability of *recall errors* from the event $D^* = 1, D_S^1 = D^*, D_S^2 = 1-D^*$, denoting individuals holding the qualification of interest who report so at age 23, but who do not recall having the qualification ten years later. The probability of this event can be computed as:

$$f_{\mathbf{D}_S|D^*}[1,0|1]f_{D^*}[1].$$

5.2. Characterizing the extent of misclassification

The first three panels of Fig. 1 present the distributions across individuals of the probabilities of exact classification, namely $f_{W|D^*X}[1|1,x]$ and $f_{W|D^*X}[0|0,x]$, for school files ($W = D_T$), for reports in 1981 ($W = D_S^1$) and for reports in 1991 ($W = D_S^2$). The probabilities of exact classification have been calculated for all individuals in our sample (i.e. males of the NCDS cohort) using the methodology described in Section 3. As for each individual our procedure yields 2000 realizations from the posterior distribution of the quantity of interest, all distributions in Fig. 1 are obtained by first taking the individual average of these realizations, and then plotting the distribution of such averages across individuals. The probabilities of exact classification by recorded attainment reported in Table 3 are simply the averages of these distributions.

Our results suggest that individuals are appreciably *less* accurate than transcripts when they do not have any academic qualification, and this is even more so when survey reports from the later 1991 wave are considered. Specifically, the bulk of the distributions on the left hand side column of Fig. 1 increasingly shifts towards lower values as one moves down the three indicators (D_T, D_S^1, D_S^2). The averages reported in the second row of Table 3 summarize the extent of misclassification/over-reporting for individuals without academic qualifications as being 23% in the school files, but as high as 33% and 38% in the 1981 and 1991 surveys. The degree of accuracy of self-reported measurements thus seems to be 10 percentage points lower when compared to concurrent transcript records in the case of no qualifications. Reporting a decade later is estimated as adding a further accuracy loss of 5 percentage point for individuals without academic qualifications.

Table 4
Extent of consistent misclassification across survey instruments (self-reported information from individuals at age 23 and at age 33).

	Academic qualification	
	Any	None
Probabilities of consistent misclassification:		
Truth tellers	0.761	0.573
	(*0.039*)	(*0.056*)
Over reporters		0.263
		(*0.071*)
Under reporters	0.113	
	(*0.045*)	
Confused	0.126	0.164
	(*0.02*)	(*0.03*)

Note. The table presents estimates of the percentage of individuals who consistently report correctly (*truth tellers*), over-report (*over-reporters*) and under-report (*under-reporters*) their educational qualification across survey waves. Presented also is the percentage of individuals with inconsistent response behaviour across survey waves (*confused*). Numbers in the first column refer to $f_{\mathbf{D}_S|D^*}[1,1|1]$, $f_{\mathbf{D}_S|D^*}[0,0|1]$ and the residual category, respectively. Numbers in the second column refer to $f_{\mathbf{D}_S|D^*}[0,0|0]$, $f_{\mathbf{D}_S|D^*}[1,1|0]$ and the residual category, respectively. See Section 5.1 for definitions. Posterior standard deviations are reported in parentheses.

On the other hand, it seems that individuals are slightly *more* accurate than transcripts when they do in fact have academic qualifications (see the right hand side column of Fig. 1, and the first row of Table 3). Misclassification/under-reporting for individuals with academic qualifications is 22% in the school files, 16% in the contemporaneous survey and 20% in the later survey. Individuals with qualifications are thus between 2% to 6% more likely than schools to report correctly their attainment, pointing to a survey wave effect of the same magnitude as the one uncovered for individuals without qualifications (the survey closer to completion being 4 percentage points more accurate than the later survey).

In line with the little evidence available from the US, no source thus appears to be uniformly better. For individuals, we find that over-reporting is by far the most important source of error and that both types of reporting error worsen over time. Under-reporting is more of a problem in transcript files, although the incidence of errors coming from under- and over-reporting is markedly more similar than when individuals are considered.

Notwithstanding their different underlying patterns of measurement error, the two types of data sources appear to be remarkably similar in their overall reliability, especially when the sources collect the information of interest close in time. Specifically, the extent of correct classification for school files is estimated at 86.8%, for the 1981 wave at 86.3% and for the 1991 wave at 82.0% (see the last row of Table 3). The numbers reported thus suggest that self-reported measurements close to completion are just as accurate as the administrative information coming from the schools. The degree of accuracy is however around 4 percentage points lower when the information is collected 10 years after the qualification was attained.

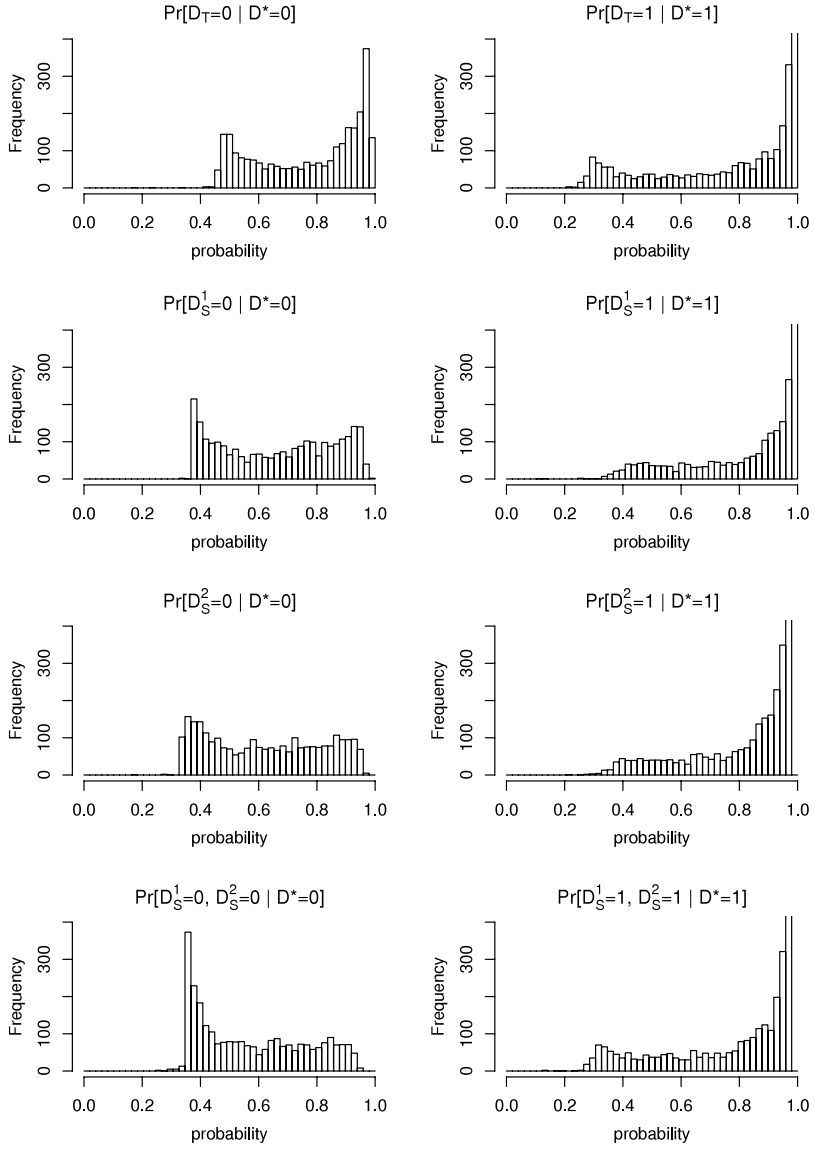

Fig. 1. Probabilities of exact classification in the indicators of educational attainment (transcript files from schools and self-reported information from individuals at age 23 and at age 33). **Notes**. *Top panel*: Probabilities of exact classification in *administrative* information, i.e. percentage of individuals having any academic qualification for whom schools report so (right hand side figure) and percentage of individuals without academic qualifications for whom schools report so (left hand side figure). *Central panels*: probabilities of exact classification in *self-reported* information at age 23 (1981 wave) and at age 33 (1991 wave), respectively, i.e. percentage of individuals reporting any academic qualification amongst those having so (right hand side figures) and percentage of individuals reporting *no* academic qualification amongst those without the qualification (left hand side figures). *Bottom panel*: probabilities of consistent exact classification in both *self-reported* information at age 23 (1981 wave) and at age 33 (1991 wave), i.e. percentage of individuals reporting any academic qualification in both waves amongst those having so (right hand side figure) and percentage of individuals reporting *no* academic qualification in both waves amongst those without the qualification (left hand side figure). Posterior distributions are presented throughout (see Section 3 for details).

Using the misclassification probabilities, we recovered an estimate of the true incidence of academic qualifications in the population, namely $f_{D^*}[1]$, of 62.9%, midway between the incidence according to the school files (58.8%) and either self-reported educational measure (64.0% in the 1981 wave and 65.0% in the 1991 wave).

The availability of two repeated measurements of qualifications which were self-reported by the same individuals at two points in time gives us the unique chance of assessing the temporal patterns of misreporting across survey instruments and of decomposing misreporting errors into a systematic component linked to individuals' persistent behaviour and into a transitory part reflecting survey errors that occur independently of individual behaviour in each cross section survey wave. Table 4 offers important insights on the nature of these errors.

First, the proportion of consistent truth-tellers, that is of those individuals who correctly self-report their educational attainment in both survey waves, is considerably higher amongst those who do have academic qualifications (76%) than amongst those who do not (57%). This is graphically corroborated by the corresponding distributions across individuals presented in the bottom panel of Fig. 1. Overall, we calculated that the percentage of truth tellers represents slightly more than two thirds (69%) of the NCDS sample.

Looking at the share of consistent truth tellers amongst those correctly reporting their attainment in a given survey wave, we find that among those who do have academic qualifications, 90% (=0.761/0.844) of individuals who report so correctly in wave 1 will also report correctly in wave 2 and 95% (=0.761/0.803) of individuals who reported correctly in wave 2 had also reported correctly in wave 1. Among those with no academic qualifications, the

corresponding ratios are around 3 percentage points lower (86% and 92%). Figures from just one survey round may thus not reveal behaviour, as we have shown that individuals with or without the qualification of interest have different survey response patterns over time. Our results do however show that the bulk of correct classification can be attributed to some degree of persistency in the reporting of individuals across waves, while the remaining error (about 5 to 14 percentage points depending on the measurement considered) is not systematic.

Our results further provide a formal test against the assumption that self-reported measurements in the 1981 and the 1991 surveys are conditionally independent given D^*. This would amount to assuming conditionally independent errors in the two survey measurements, thus ruling out possible correlation that may arise, for example, from unobserved individual propensity to misreport. Under the assumption stated, the covariance between D_S^1 and D_S^2, conditional on the true attainment D^*, would be zero, meaning that the probability of consistent classification in Table 4 should be equal to the product of the probabilities of exact classification in the two waves in Table 3. The evidence we find clearly points to a different pattern (for those with qualifications, $0.761 > 0.678 = 0.844 \times 0.803$; for those without qualifications, $0.573 > 0.415 = 0.666 \times 0.623$), highlighting the presence of positive autocorrelation in measurements after controlling for D^*.[24]

Consistent over-reporters appear to be an important fraction of the no-qualification sample: over one quarter (26.3%) of the NCDS members without any academic qualification over-report their attainment at both survey waves. The size of this group would be noticeably overstated if one were to consider only what happens in one survey wave (33.4% and 37.7% of the no-qualification samples in the 1981 and in the 1991 surveys, respectively). These two sets of results thus suggest that between 21% and 30% of over-reporting errors in a given wave are the results of non-systematic recording errors.

In survey data asking for a positive trait, one would expect the share of consistent under-reporters to be much lower than the one of over-reporters. Indeed, at 11.3%, it is more than half the size. As was the case for over-reporting, focusing on one survey wave alone would overstate the amount of under-reporting. Once we again combine the cross-sectional and panel results, we find that the share of under-reporting errors accounted for by non-systematic survey errors is quite similar to the one that accounted for over-reporting errors (28% and 43%), giving us confidence that we have indeed isolated the true random error component that occurs independently of individual behaviour.

The last group, the "confused", are those whose attainment is correctly recorded in one wave, but misrecorded in the other. This group makes up 14% of the NCDS sample, with slightly more "confused" among the no-qualification group (16.4%) than among the qualification group (12.6%). The most interesting subgroup amongst the "confused" is the group affected by recall bias, whose share is given by $f_{\mathbf{D}_S|D^*}[1, 0|1]$. We estimated the incidence of recall errors among those with the qualification at 6.2%, and in the NCDS sample at 1.1%.

5.3. Returns to any academic qualification

This section presents our estimates of the true return from achieving any academic qualification, focusing on the male sample of the NCDS. The bulk of the discussion of our empirical findings is devoted to our estimates of the average return to achieving any academic qualification for those individuals who have chosen to do so (ATT) under Assumption 6 of strong ignorability. We then briefly discuss our estimates of the average return for a specific group (LATE) under Assumption 7 that a valid instrumental variable is available.

5.3.1. Estimating returns under strong ignorability

As argued in Section 4, the NCDS contains very rich background information, including not only detailed family background variables, but also an array of cognitive and personality measures taken as early as age 7.[25] This makes this data particularly suited to exploring estimation of returns under Assumption 6, i.e. that controlling for the full set of observed variables is enough to directly correct for selection (or "ability") bias.

Our point of departure in this subsection are the 'raw' OLS estimates from Table 1 (discussed in Section 4.3) which do not account for misclassification in the educational measures. We focus on the OLS estimates as the fully interacted regression model (FILM) did not provide evidence of heterogeneous returns. OLS is also the standard methodological choice for the estimation of returns in the UK (see e.g. Dearden, 1999; Dearden et al., 2002; McIntosh, 2006; Jenkins et al., 2007; Department for Business, Innovation and Skills, 2011), making it the ideal benchmark for comparison with the method we propose.

Let Δ_{FULL} and Δ_{LFS} thus denote OLS estimates obtained from the raw data without controlling for misclassification and employing either the full set of controls available in the NCDS or the LFS-style variables. Estimates that are adjusted for misclassification following the procedure outlined in Section 3 are denoted by Δ^*_{FULL} and Δ^*_{LFS}. Under Assumption 6, Δ^*_{FULL} is the true ATT, i.e. $\Delta^*_{FULL} = E_{Y_1|D^*}[Y_1|1] - E_{Y_0|D^*}[Y_0|1]$.

Our estimate of Δ^*_{FULL} is a 26.7% wage gain from achieving at least O-levels, with a posterior standard deviation of 0.072 (see Table 5). When we correct for misrecording but only rely on the smaller set of controls (Δ^*_{LFS}), the estimated ATT is 37.8% with a posterior standard deviation of 0.042 (note that we use such limited set of variables both to estimate the misclassification probabilities and to then estimate the return). Taken together, these two results point to a 42% upward bias in estimated returns that do not fully control for selection into educational attainment.

Next, we compare our most reliable estimate, Δ^*_{FULL}, to OLS estimates based on either the full set or the LFS-set of controls. Interest in the latter comparison arises from the widespread practice in the UK of estimating returns using OLS regression based on LFS data and ignoring potential misclassification (see e.g. McIntosh, 2006; Jenkins et al., 2007; Department for Business, Innovation and Skills, 2011). In order to heuristically compare frequentist and Bayesian estimates, we construct p-values using the asymptotic distribution of the OLS estimator, calculating the probability of values larger, in absolute terms, than Δ^*_{FULL}. This amounts to assuming that the latter is the true value of the ATT. To ease readability, in the table we simply refer to these numbers as p-values for the statistical difference between Δ^*_{FULL} and Δ_{FULL}, or between Δ^*_{FULL} and Δ_{LFS}.

A first important result is that controlling for misclassification matters, and leads to an estimate of Δ^*_{FULL} which is statistically different from the ones of Δ_{FULL} based on either educational measure.

[24] Note also that this correlation cannot be explained by the observable characteristics X: the evidence discussed is against the assumption that D_S^1 and D_S^2 are conditionally independent given D^* and X, as there must be at least one value of X such that the latter assumption is violated. Fig. 6 in the online appendix (see Appendix C) presents the conditional distributions $f_{D_S^2|D^*D_S^1X}[a|1, b, x]$ and $f_{D_S^2|D^*D_S^1X}[a|0, b, x]$, visualizing the strong correlation across self-reports in the two survey waves.

[25] A large body of evidence has in fact underscored the importance of both cognitive and non-cognitive abilities in predicting many socio-economic outcomes (see e.g. Almlund et al., 2011 and Borghans et al., 2008 for extensive reviews on the powerful role of personality skills, and recent evidence by Heckman et al., 2013 on their importance in explaining the effects of an educational intervention).

Table 5
Comparison of estimates of returns to educational attainment.

Δ^*_{FULL}	0.267 (*0.072*)		
	Transcript files from schools	1981 Wave (at age 23)	1991 Wave (at age 33)
Δ_{LFS}	0.332 (*0.015*)	0.333 (*0.016*)	0.293 (*0.015*)
p-value: $\Delta_{LFS} = \Delta^*_{FULL}$	0.000	0.000	0.082
Δ_{FULL}	0.174 (*0.018*)	0.174 (*0.017*)	0.131 (*0.017*)
p-value: $\Delta_{FULL} = \Delta^*_{FULL}$	0.000	0.000	0.000

Note. The top panel of the table reports the ATT (Δ^*_{FULL}) computed as described in Section 3 under Assumption 6, which represents our most reliable estimate (posterior standard deviation is in parentheses). It is obtained using the full set of controls, and adjusting for misclassification. Also, reported is the OLS estimate of the same parameter from Table 1, using LFS controls (Δ_{LFS}) and the full set of controls available in the NCDS sample (Δ_{FULL}). *P*-values refer to the test of equality of the two estimates (see Section 5 for a description of how the test was implemented).

Specifically, controlling for ability bias but ignoring misclassification leads to a 35% downward bias in the case of concurrent reports and to a 51% downward bias when using reports based on recall. Ignoring misclassification would thus provide a highly misleading picture of how much people with academic qualifications have gained by investing in education.

Ignoring both omitted-ability bias and potential misclassification in recorded attainment close to completion (either in the school files or self-reported), we find a return to academic qualifications (Δ_{LFS}) of 33%, which is significantly different from Δ^*_{FULL}. We thus do not find any evidence of balancing biases; quite to the contrary, ignoring both biases leads to a sizeable upward bias in estimated returns of around one quarter (24%). This result is reassuringly consistent with the findings in Battistin and Sianesi (2011), who bound the ATT and find that ignoring both misreporting and omitted ability bias would generally lead to at times quite severely upward biased estimates of true returns. In a situation where educational records were obtained relatively close to the completion of the qualification of interest – irrespective of whether reported from school or individuals – we thus find that the policymaker or analyst cannot simply rely on measurement error to cancel out the ability bias. The resulting calibration rule to get the LFS-style estimate of the average return to academic qualifications for males close to the true return suggests to multiply the "raw" estimate by 0.8.

With educational information collected after over 10 years since completion we expect the relative importance of omitted variable bias and measurement error bias to shift, given that, in line with *a priori* expectations, we have found the recall measure to suffer from a larger extent of misclassification. Indeed, relying on the recall educational measure and controlling only for the LFS-style variables, the estimated raw return (Δ_{LFS}) is 29.3%, which is more than halved once we control for the full set of observables (Δ_{FULL} being equal to 13.1%). However, once we compare these estimates to the true return (Δ^*_{FULL}) of 26.7%, we find that the latter is very close and statistically indistinguishable (at the 90% level) from the *raw* estimate Δ_{LFS}. In this application, measurement error in recall information is thus strong enough to fully compensate for the upward bias induced by omitted ability controls. Specifically, while estimates that correct for misclassification but not for selection incur a 42% upward bias (compare Δ^*_{FULL} to Δ^*_{LFS}), controlling for selection but ignoring misclassification gives rise to a bias of roughly the same magnitude (51%) but of different sign. Hence in sharp contrast to a situation where information on education was obtained relatively close to attainment, when relying on recall information it seems indeed to be the case that the two biases cancel each other out. There thus seems to be no need for a calibration rule: LFS-style estimates of the average return to academic qualifications based on recall information on qualifications are indeed very close to the true return.

5.3.2. Estimating returns under an exclusion restriction

To instrument the acquisition of academic qualifications at the minimum school leaving age we use a combined measure of father's and mother's interest in the child's education as assessed by the child's teacher when the child was 7 years old (see Section 4). The identifying assumption of this model requires parental interest at age 7 to be legitimately excluded from potential wages at age 33 for given early cognitive and non-cognitive traits, early school performance (measured by test scores at age 11), family background and school type. One could in general argue that in addition to educational attainment, parental interest could affect other individual traits, such as motivation or self-esteem, that could in turn affect wages. In defending the exclusion restriction it is thus crucial to be able to condition on such types of socio-emotional characteristics, as well as on the parental choice of school and on the child's early school performance after parental interest was measured.

An important feature of an IV model is that it estimates a LATE, in our case the average return for those children who would acquire at least minimum academic qualifications only if their parents were very interested in their education, but who would otherwise stop at 16 without any qualifications. The estimated proportion of compliers is 2% in the population, our LATE thus pertains to a very small proportion of the population. As was the case for the 'raw' IV estimates in Section 4.4, our IV estimate that corrects for misclassification is very imprecisely estimated, showing a statistically insignificant 22.6% average wage gain for the compliers (with posterior standard deviation of 23.2). Following the same testing procedure of Section 5.3, a heuristic comparison of the raw and corrected IV estimates fails to uncover any significant differences. These results should only be taken as illustrative at best, as all estimates are highly imprecise and indeed none is individually significant.

6. Conclusions

In this paper we have provided reliable estimates of the returns to educational qualifications in the UK that allow for the possibility of misreported attainment under two alternative identifying assumptions: strong ignorability and an exclusion restriction. We have additionally identified the extent of misreporting in different types of commonly used data sources on educational qualifications: exam transcript files from schools and self-reported educational measures at different elapsed times after completion of the qualification of interest. We have thus provided estimates of the relative reliability of these different data sources, as well as of the temporal correlation in individual response patterns.

Under strong ignorability we have also produced some simple calibration rules as to how to correct returns estimated on data that rely on self-reported measures of qualifications and contain limited or no information on individual ability and family background characteristics (such as the Labour Force Survey).

Table 6
Summary statistics.

Variable	Mean	Std.dev.
Real log hourly wage	2.059	(0.426)
Any academic qualifications by age 23		
School report	0.588	(0.492)
Self report at age 23	0.650	(0.477)
Self report at age 33	0.640	(0.480)
White	0.972	(0.166)
Ability at 7		
Math test score	4.944	(2.840)
Verbal test score	21.41	(9.500)
BSAG total score all syndromes	7.797	(8.357)
First factor Rutter Behaviour Scale	0.061	(0.761)
Factor score cognitive and non-cognitive measures	−0.098	(0.969)
Any ability measure missing	0.144	(0.351)
Ability at 11		
Math test score	16.756	(11.55)
Verbal score on general ability test	20.276	(11.42)
Non verbal score on general ability test	19.491	(10.07)
Reading comprehension score	15.162	(8.153)
BSAG total score all syndromes	6.921	(8.147)
First factor Rutter Behaviour Scale	0.069	(0.708)
Factor score cognitive and non-cognitive measures	−0.030	(0.932)
Any ability measure missing	0.187	(0.390)
Parental background		
Father's years of education	7.634	(4.693)
Father's education missing	0.172	(0.377)
Mother's years of education	7.709	(4.435)
Mother's education missing	0.158	(0.365)
Father's age (child aged 16)	43.06	(13.72)
Father's age missing	0.076	(0.265)
Mother's age (child aged 16)	41.51	(10.76)
Mother's age missing	0.048	(0.214)
Father's social class (child aged 16)		
Professional	0.049	(0.217)
Intermediate	0.148	(0.355)
Skilled non-manual	0.083	(0.276)
Skilled manual	0.310	(0.463)
Semi-skilled non-man	0.011	(0.103)
Semi-skilled manual	0.098	(0.297)
Unskilled manual	0.029	(0.168)
Unknown/unempl/no father/missing	0.272	(0.445)
Mother employed (child aged 16)	0.540	(0.499)
Number of siblings	1.740	(1.772)
Number of siblings missing	0.072	(0.258)
School type at age 16		
Comprehensive	0.489	(0.500)
Secondary Modern	0.163	(0.370)
Grammar	0.108	(0.310)
Public	0.062	(0.241)
Other	0.018	(0.132)
Missing school information	0.160	(0.367)
Region at age 16		
North Western	0.104	(0.305)
North	0.070	(0.254)
East and West Riding	0.084	(0.277)
North Midlands	0.076	(0.265)
London and South East	0.144	(0.351)
Eastern	0.081	(0.272)
Southern	0.060	(0.238)
South Western	0.063	(0.243)
Midlands	0.085	(0.280)
Wales	0.059	(0.236)
Scotland	0.103	(0.304)
Other	0.072	(0.258)
Father and/or mother overly or very interested in child's education at age 7	0.436	(0.496)
Number of observations	2716	

Note. Reported are summary statistics from raw data from the British National Child Development Survey (see Section 4 for more details).

Results in this paper thus represent a new piece of evidence for appreciating the relative reliability of different sources of educational information, as well as for checking the robustness of current estimates of returns to the presence of misreported qualifications. Knowing the extent of misreporting also has obvious implications for the interpretation of other studies that use educational attainment as an outcome variable or for descriptive purposes.

Appendix A. Proof of non-parametric identification

The aim of this Appendix is to show that the setup considered in Section 2 is sufficient to non-parametrically identify the mixture components $f_{Y|D^*}[y|d^*]$ and the extent of misclassification in the data. We use the setting considered by Hu (2008) to allow for over-identification which, for the case at hand, arises because of

the availability of repeated measurements coming from the same individuals; for simplicity, the conditioning on $V = v$ will be left implicit throughout.

Let the following matrices constructed from *raw* data be defined:

$$\underset{2\times 4}{\mathcal{F}_{Y\mathbf{D}_S|D_T}} = \begin{bmatrix} f_{Y\mathbf{D}_S|D_T}[y,0,0|0] & f_{Y\mathbf{D}_S|D_T}[y,0,1|0] & f_{Y\mathbf{D}_S|D_T}[y,1,0|0] & f_{Y\mathbf{D}_S|D_T}[y,1,1|0] \\ f_{Y\mathbf{D}_S|D_T}[y,0,0|1] & f_{Y\mathbf{D}_S|D_T}[y,0,1|1] & f_{Y\mathbf{D}_S|D_T}[y,1,0|1] & f_{Y\mathbf{D}_S|D_T}[y,1,1|1] \end{bmatrix},$$

$$\underset{2\times 4}{\mathcal{F}_{\mathbf{D}_S|D_T}} = \begin{bmatrix} f_{\mathbf{D}_S|D_T}[0,0|0] & f_{\mathbf{D}_S|D_T}[0,1|0] & f_{\mathbf{D}_S|D_T}[1,0|0] & f_{\mathbf{D}_S|D_T}[1,1|0] \\ f_{\mathbf{D}_S|D_T}[0,0|1] & f_{\mathbf{D}_S|D_T}[0,1|1] & f_{\mathbf{D}_S|D_T}[1,0|1] & f_{\mathbf{D}_S|D_T}[1,1|1] \end{bmatrix}.$$

Define the following *latent* matrices:

$$\underset{2\times 4}{\mathcal{F}_{\mathbf{D}_S|D^*}} = \begin{bmatrix} f_{\mathbf{D}_S|D^*}[0,0|0] & f_{\mathbf{D}_S|D^*}[0,1|0] & f_{\mathbf{D}_S|D^*}[1,0|0] & f_{\mathbf{D}_S|D^*}[1,1|0] \\ f_{\mathbf{D}_S|D^*}[0,0|1] & f_{\mathbf{D}_S|D^*}[0,1|1] & f_{\mathbf{D}_S|D^*}[1,0|1] & f_{\mathbf{D}_S|D^*}[1,1|1] \end{bmatrix},$$

$$\underset{2\times 2}{\mathcal{F}_{D^*|D_T}} = \begin{bmatrix} f_{D^*|D_T}[0|0] & f_{D^*|D_T}[1|0] \\ f_{D^*|D_T}[0|1] & f_{D^*|D_T}[1|1] \end{bmatrix},$$

$$\underset{2\times 2}{\mathcal{F}_{Y|D^*}} = \begin{bmatrix} f_{Y|D^*}[y|0] & 0 \\ 0 & f_{Y|D^*}[y|1] \end{bmatrix},$$

which are characterized by 10 unknowns.

Using Assumptions 1 and 2 there is:

$$f_{Y\mathbf{D}_S|D_T}[y,\mathbf{d}_S|d_T] = \sum_{d^*=0}^{1} f_{Y|D^*}[y|d^*] f_{\mathbf{D}_S|D^*}[\mathbf{d}_S|d^*] f_{D^*|D_T}[d^*|d_T],$$

$$f_{\mathbf{D}_S|D_T}[\mathbf{d}_S|d_T] = \sum_{d^*=0}^{1} f_{\mathbf{D}_S|D^*}[\mathbf{d}_S|d^*] f_{D^*|D_T}[d^*|d_T],$$

or, in matrix notation:

$$\mathcal{F}_{Y\mathbf{D}_S|D_T} = \mathcal{F}_{D^*|D_T}\mathcal{F}_{Y|D^*}\mathcal{F}_{\mathbf{D}_S|D^*}, \tag{2}$$

$$\mathcal{F}_{\mathbf{D}_S|D_T} = \mathcal{F}_{D^*|D_T}\mathcal{F}_{\mathbf{D}_S|D^*}. \tag{3}$$

Now, under Assumption 4 the matrix $\mathcal{F}_{D^*|D_T}$ is nonsingular (i.e. full rank), so that from (3) there is:

$$\mathcal{F}_{\mathbf{D}_S|D^*} = \mathcal{F}^{-1}_{D^*|D_T}\mathcal{F}_{\mathbf{D}_S|D_T}, \tag{4}$$

which if substituted into (2) yields:

$$\mathcal{F}_{Y\mathbf{D}_S|D_T} = \mathcal{F}_{D^*|D_T}\mathcal{F}_{Y|D^*}\mathcal{F}^{-1}_{D^*|D_T}\mathcal{F}_{\mathbf{D}_S|D_T}.$$

Identification of $\mathcal{F}_{Y|D^*}$, $\mathcal{F}_{D^*|D_T}$ and $\mathcal{F}_{\mathbf{D}_S|D^*}$ is achieved by considering a particular type of generalized inverse, called the right Moore–Penrose inverse, which here always exists and is unique provided that the matrix to be inverted is of full rank (see, for example, Seber, 2008). Define:

$$\mathcal{A}^{+} \equiv \mathcal{A}'(\mathcal{A}\mathcal{A}')^{-1}.$$

The matrix $\mathcal{A}^{+}$ is known as the right Moore–Penrose inverse of the matrix $\mathcal{A}$ and has the property that $\mathcal{A}\mathcal{A}^{+}$ equals the identity matrix. It follows that:

$$\mathcal{F}_{Y\mathbf{D}_S|D_T}\mathcal{F}^{+}_{\mathbf{D}_S|D_T} = \mathcal{F}_{D^*|D_T}\mathcal{F}_{Y|D^*}\mathcal{F}^{-1}_{D^*|D_T},$$

where $\mathcal{F}_{\mathbf{D}_S|D_T}$ has full rank because of Assumption 5. By defining $\mathcal{M} \equiv \mathcal{F}_{Y\mathbf{D}_S|D_T}\mathcal{F}^{+}_{\mathbf{D}_S|D_T}$, mixture components $\mathcal{F}_{Y|D^*}$ and misclassification probabilities $\mathcal{F}_{D^*|D_T}$ can be obtained as eigenvalues and eigenvectors from the eigenvalue decomposition of $\mathcal{M}$.

Additional assumptions need to be imposed to establish a unique correspondence between the eigenvalues and the eigenvectors in this factorization. First, Assumption 3 guarantees that there exist two eigenvalues in the diagonal matrix $\mathcal{F}_{Y|D^*}$, and that these are distinguishable. Second, Assumption 4 ensures that $\mathcal{F}_{D^*|D_T}$ is a diagonally dominant matrix, hence determining the ordering of the eigenvalues and the eigenvectors. Third, in order to fully characterize the distribution $f_{D^*|D_T}[d^*|d_T]$, the eigenvectors need to be suitably normalized using the property that for any given d_T $Pr(D^* = 1|D_T = d_T) + Pr(D^* = 0|D_T = d_T) = 1$. Knowledge of the latter probabilities implies, via (4), identification of the $f_{\mathbf{D}_S|D^*}[\mathbf{d}_S|d^*]$'s and of the $f_{D^*}[d^*]$'s. This in turn implies identification of the mixture weights in (1).

The above argument may be generalized further to accommodate for D^*, D_S and D_T to be categorical random variables taking an arbitrary number of values as long as the conditional independence assumption between D_T and D_S is maintained. In this more general setting, the main complication lies in the fact that $\mathcal{F}_{D^*|D_T}$ is no longer a square matrix, and that the existence of its left generalized inverse, crucial to obtain Eq. (4) and defined by $\mathcal{A}^{-} = \mathcal{A}(\mathcal{A}'\mathcal{A})^{-1}$, is not guaranteed by the full rank condition stated above. It must also be the case that the number of columns of the matrix to be inverted is larger than the number of its corresponding rows. In our setup, this would amount to assuming that the support of the instrument D_T is larger than the support of the latent random variable D^*, an assumption which is standard in the literature on instrumental variables. Finally, a generalization of Assumptions 3 and 4 would be required to identify the ordering of eigenvalues and eigenvectors. This would introduce refinements to what discussed above that call for further research.

Appendix B

See Table 6.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2014.03.002.

References

Aigner, D., 1973. Regression with a binary independent variable subject to errors of observation. J. Econometrics 1, 49–60.

Almlund, M., Duckworth, A., Heckman, J., Kautz, T., 2011. Personality psychology and economics. In: Hanushek, S.M.E., Woessman, L. (Eds.), Handbook of the Economics of Education, vol. 4. Elsevier, Amsterdam, pp. 1–181.

Battistin, E., De Nadai, M., Sianesi, B., 2011. Multiple schooling, multiple measures and returns to educational qualification, IZA Discussion Paper, N. 6337.

Battistin, E., Sianesi, B., 2011. Misclassified treatment status and treatment effects: an application to returns to education in the UK. Rev. Econ. Stat. 93 (2), 495–509.

Black, D., Berger, M., Scott, F., 2000. Bounding parameter estimates with nonclassical measurement error. J. Amer. Statist. Assoc. 95 (451), 739–748.

Black, D., Sanders, S., Taylor, L., 2003. Measurement of higher education in the census and current population survey. J. Amer. Statist. Assoc. 98 (463), 545–554.

Blundell, R., Dearden, L., Sianesi, B., 2005a. Measuring the returns to education. In: What's the Good of Education? The Economics of Education in the UK. Princeton University Press, ISBN: 0691117349, pp. 117–145.

Blundell, R., Dearden, L., Sianesi, B., 2005b. Evaluating the effect of education on earnings: models, methods and results from the national child development survey. J. R. Stat. Soc. Ser. A 168 (3), 473–512.

Borghans, L., Duckworth, A., Heckman, J.J., ter Weel, B., 2008. The economics and psychology of personality traits. J. Hum. Resour. 43 (4), 972–1059.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics. Vol. 5. North-Holland, Amsterdam, pp. 3705–3843.

Chen, X., Hong, H., Nekipelov, D., 2011. Nonlinear models of measurement errors. J. Econ. Lit. 49 (4), 901–937.

Chevalier, A., Harmon, C., Walker, I., Zhu, Y., 2004. Does education raise productivity, or just reflect it? Econ. J. 114 (499), 499–517.

Dearden, L., 1999. Qualifications and earnings in Britain: how reliable are conventional OLS estimates of the returns to education? IFS Working Paper W99/7.

Dearden, L., McIntosh, S., Myck, M., Vignoles, A., 2002. The returns to academic and vocational qualifications in britain. Bull. Econ. Res. 54, 249–274.

Del Bono, E., Galindo-Rueda, F., 2004. Do a few months of compulsory schooling matter? The education and labour market impact of school leaving rules, IZA Discussion Paper No. 1233.

Department for Business, Innovation and Skills, 2011. Returns to Intermediate and Low Level Vocational Qualifications, BIS Research Paper Number 53, London, UK.

Di Nardo, J., Lee, D.S., 2011. Program evaluation and research designs. In: Handbook of Labor Economics, Vol. 4, Part A. pp. 463–536 (Chapter 5).

Everitt, B.S., Hand, D.J., 1981. Finite Mixture Distributions. Chapman and Hall, London.

Fisher, R.A., 1935. The Design of Experiments. Oliver & Boyd, Edinburgh.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378–382.

Heckman, J.J., Honore, B.E., 1990. The empirical content of the Roy model. Econometrica 58 (5), 1121–1149.
Heckman, J.J., Lalonde, R., Smith, J., 1999. The economics and econometrics of active labor market programs. In: Ashenfelter, A., Card, D. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier Science, Amsterdam.
Heckman, J.J., Pinto, R., Savelyev, P., 2013. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. Amer. Econ. Rev. 103 (6), 2052–2086.
Hu, Y., 2008. Identification and estimation of nonlinear models with misclassification error using instrumental variables: a general solution. J. Econometrics 144 (1), 27–61.
Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36 (1), 167–171.
Imbens, G.W., 2000. The role of the propensity score in estimating dose–response functions. Biometrika 87 (3), 706–710.
Imbens, G.W., 2004. Semiparametric estimation of average treatment effects under exogeneity: a review. Rev. Econ. Stat. 86, 4–29.
Imbens, G.W., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.
Imbens, G.W., Rubin, D.B., 1997. Estimating outcome distributions for compliers in instrumental variables models. Rev. Econom. Stud. 64, 555–574.
Jenkins, A., Greenwood, C., Vignoles, A., 2007. The returns to qualifications in England: updating the evidence base on level 2 and level 3 vocational qualifications, Centre for the Economics of Education Discussion Paper 89, London, UK.
Kane, T.J., Rouse, C., Staiger, D., 1999. Estimating returns to schooling when schooling is misreported, National Bureau of Economic Research Working Paper No. 7235.
Kreider, B., Pepper, J.V., 2007. Disability and employment: reevaluating the evidence in light of reporting errors. J. Amer. Statist. Assoc. 102 (478), 432–441.
Kreider, B., Pepper, J.V., Gundersen, C., Jolliffe, D., 2012. Identifying the effects of snap (food stamps) on child health outcomes when participation is endogenous and misreported. J. Amer. Statist. Assoc. 107 (499), 958–975.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.
Lechner, M., 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. Physica, Heidelberg, pp. 43–58.
Lewbel, A., 2007. Estimation of average treatment effects with misclassification. Econometrica 75 (2), 537–551.
Mahajan, A., 2006. Identification and estimation of regression models with misclassification. Econometrica 74 (3), 631–665.
Manski, C.F., Pepper, J.V., 2000. Monotone instrumental variables: with an application to the returns to schooling. Econometrica 68 (4), 997–1010.
McIntosh, S., 2006. Further analysis of the returns to academic and vocational qualifications. Oxf. Bull. Econ. Stat. 68 (2), 225–251.
Mellow, W., Sider, H., 1983. Accuracy of response in labor market surveys: evidence and implications. J. Labor Econom. 1 (4), 331–344.
Molinari, F., 2008. Partial identification of probability distributions with misclassified data. J. Econometrics 144 (1), 81–117.
Neyman, J., 1935. Statistical problems in agricultural experimentation. Suppl. J. R. Stat. Soc. 2, 107–180 (with co-operation by Iwaszkiewicz, K., and Kolodziejczyk, S.).
Quandt, R., 1972. Methods for estimating switching regressions. J. Amer. Statist. Assoc. 67, 306–310.
Roy, A., 1951. Some thoughts on the distribution of earnings. Oxf. Econ. Papers 3, 135–146.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomised and non-randomised studies. J. Educ. Psychol. 66, 688–701.
Rubin, D.B., 1980. Discussion of randomisation analysis of experimental data in the Fisher randomisation test' by Basu. J. Amer. Statist. Assoc. 75, 591–593.
Seber, G.A.F., 2008. A Matrix Handbook for Statisticians. Wiley, New Jersey.
Yakowitz, S.J., Spragins, J.D., 1968. On the identifiability of finite mixtures. Ann. Math. Stat. 39, 209–214.