Battistin, Erich; de Nadai, Michele; Sianesi, Barbara

Working Paper

Misreported schooling, multiple measures and returns to educational qualifications

IZA Discussion Papers, No. 6337

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Battistin, Erich; de Nadai, Michele; Sianesi, Barbara (2012) : Misreported schooling, multiple measures and returns to educational qualifications, IZA Discussion Papers, No. 6337, Institute for the Study of Labor (IZA), Bonn, https://nbn-resolving.de/urn:nbn:de:101:1-201205026030

This Version is available at:
https://hdl.handle.net/10419/58474

DISCUSSION PAPER SERIES

IZA DP No. 6337

Misreported Schooling, Multiple Measures and Returns to Educational Qualifications

Erich Battistin
Michele De Nadai
Barbara Sianesi

February 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Misreported Schooling, Multiple Measures and Returns to Educational Qualifications

Erich Battistin
University of Padova,
IRVAPP and IZA

Michele De Nadai
University of Padova

Barbara Sianesi
Institute for Fiscal Studies

Discussion Paper No. 6337
February 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 6337
February 2012

ABSTRACT

Misreported Schooling, Multiple Measures and Returns to Educational Qualifications*

We provide a number of contributions of policy, practical and methodological interest to the study of the returns to educational qualifications in the presence of misreporting. First, we provide 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. Second, we provide the academic and policy community with estimates of the accuracy and misclassification patterns of commonly used types of data on educational attainment: administrative files, self-reported information close to the date of completion of the qualification, and recall information ten years after completion. We are in the unique position to assess the temporal patterns of misreporting errors across survey waves, and to decompose misreporting errors into a systematic component linked to individuals' persistent behaviour and into a transitory part reflecting random survey errors. Third, by using the unique nature of our data, we assess how the biases from measurement error and from omitted ability and family background variables interact in the estimation of returns. On the methodological front, we propose a semi-parametric estimation approach based on balancing scores and mixture models, in particular allowing for arbitrarily heterogeneous individual returns.

JEL Classification: C10, I20, J31

Keywords: misclassification, mixture models, returns to educational qualifications, treatment effects

Corresponding author:

Erich Battistin
University of Padova
Department of Statistics
Via Cesare Battisti 243-5
35123 Padova
Italy
E-mail: erich.battistin@unipd.it

* Original draft February 2006. 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) and Franco Modigliani Fellowship Workshop (Rome, February 2006). Financial support from the ESRC under the research grant RES-000-22-1163 is gratefully acknowledged.

1 Introduction

Focusing on the returns to educational qualifications when attainment is potentially misreported, this paper offers a two-fold contribution. First, it provides 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. These results are obtained by casting the identification and estimation problem in terms of a mixture model, and using a semi-parametric estimation approach.

The measurement of the return to education, that is of the individual wage gain from investing in more education, has become probably the most explored and prolific area in labour economics.[1] Two important and interrelated issues arise as to the measurement of education. A first question is whether we can summarize it in the single, homogeneous measure of years of schooling. Although particularly convenient, this "one-factor" model assumes that the returns increase linearly with each additional year, irrespective of the level and type of educational qualifications the years refer to. In the US, grades generally follow years, and it has long been argued that the returns to an additional year are reasonably homogeneous (see for example Card, 1999). In the UK and other European countries, however, there are alternative nationally-based routes leading to quite different educational qualifications, and the importance of distinguishing between different types of qualifications is widely accepted. Blundell, Dearden and Sianesi (2005b) highlight the potential shortcoming of the "one-factor" model when applied to the UK's educational system, in which individuals with the same number of years of schooling have quite different educational outcomes. Not only would this obfuscate the interpretation of the return to one additional year, but imposing equality of yearly returns across educational stages was found to be overly restrictive.

A second important issue as to the measurement of education - and the one object of this paper - is the possibility of errors in recorded education and its consequences on the estimated returns. Misrecorded education could arise from data transcript errors, as well as from misreporting: survey respondents may either over-report their attainment, not know if the schooling they have had counts as a qualification or simply not remember. With the continuous years of schooling measure of education, standard results based on *classical* measurement error show that OLS estimates are downward biased,

[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. Reliable measures of returns to education are in fact needed to establish whether it is worthwhile for individuals to invest in more education (and in which type), to compare private and social returns to education, or to assess the relative value that different educational qualifications fetch on the labour market. For an extensive discussion of the policy interest of the individual wage return from education, see Blundell, Dearden and Sianesi (2005a).

and that appropriate IV methods applied to the linear regression model provide consistent estimates. Indeed, the trade-off between attenuation bias due to measurement error and upward bias due to omitted variables correlated with both schooling and wages (the so-called "ability bias") has been at the heart of the early studies on returns to years of schooling. The received wisdom has traditionally been that ability bias and measurement error bias largely cancel each other out (for a review see in particular Griliches, 1977, and Card, 1999; for a recent UK study see Bonjour *et al.*, 2003).

With the categorical qualification-based measure of education, however, any measurement error in educational qualifications will vary with the true level of education. Individuals in the lowest category can never under-report their education and individuals in the top category cannot over-report, so that the assumption of classical measurement error cannot hold (see, for example, Aigner, 1973). In the presence of misclassification, OLS estimates are not necessarily downward biased, so that the cancelling out of the ability and measurement error biases cannot be expected to hold in general. Moreover, it is now well understood that the IV methodology cannot provide consistent estimates of the returns to qualifications (see, for example, Bound, Brown and Mathiowetz, 2001).

To date, empirical evidence on the importance of these issues is restricted to the US, where it was in fact shown that measurement error might play a non-negligible role (see the results in Kane, Rouse and Staiger, 1999, Black, Sanders and Taylor, 2003, and Lewbel, 2007). For the UK there are no estimates of the returns to educational qualifications that adequately correct for measurement error.[2] This is of great concern, in view of the stronger emphasis on returns to discrete levels of educational qualifications in the UK and given the widespread belief amongst UK researchers and policymakers that ability and measurement error biases still cancel out (Dearden, 1999, Dearden *et al.*, 2002, and McIntosh, 2006).

A first possibility to overcome the bias induced by misreported educational qualifications is to derive bounds on the returns by making a priori assumptions on the misclassification probabilities (see, for example, Molinari, 2008). This approach only allows partial identification of returns. In previous work (Battistin and Sianesi, 2011) we suggest bounds that can be derived allowing for arbitrarily heterogeneous individual returns. The corresponding sensitivity analysis is easy to implement and can provide an often quite informative robustness check. The alternative approach is more demanding in terms of data requirements but, if feasible, allows point identification of the returns. An additional appealing feature is that it provides estimates of the extent of misclassification in the educational measures, which may often be of independent interest. What is needed is (at least) two categorical reports of educational qualifications for the same individuals, both potentially affected by reporting error but

[2]Ives (1984) only offers a descriptive study of the mismatch between self-reported and administrative information on qualifications in the NCDS, finding serious discrepancies particularly for the lower-level academic qualifications.

coming from independent sources (for the proof of non-parametric identification, see Mahajan, 2006, Lewbel, 2007, and 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.

In this paper, we build on the latter approach and provide a number of new contributions of considerable policy and practical relevance, as well as of methodological interest. First, we provide the first reliable estimates of the returns to educational qualifications in the UK that allow for the possibility of misreported attainment. We focus on the highly policy-relevant 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). The institutional details and the literature review relevant to motivate our interest for this parameter is discussed at length in Section 4. We rely on detailed longitudinal data from the British National Child Development Survey (NCDS), which allows us to control for a large set of family background and school type variables, as well as for ability tests taken by individuals at early ages.

Second, 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 papers 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.

Third, exploiting the information available in the NCDS data, we explore how the biases from measurement error and from omitted variables interact in the estimation of returns to educational qualifications. We produce a simple calibration rule to allow policy makers to use nationally representative data sets such as the Labour Force Survey to estimate returns to qualifications. These data totally rely on self-reported qualifications and do not contain any information on individual ability and family background.

Finally, on the methodological front we propose a semi-parametric estimation approach based

on balancing scores and mixture models. As far as we are aware we are the first ones to 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. 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 maintain throughout a unified general framework for the study of the impact of misreported treatment status on the estimation of causal treatment effects (Mahajan, 2006, Lewbel, 2007, and Molinari, 2008, and Battistin and Sianesi, 2011, are the only examples we are aware of). Our estimation method is thus of far wider interest, since the same issues arise in any application looking at the effect of a potentially misrecorded binary or categorical variable, such as eligibility to policy schemes, participation in (multiple) government programmes or work-related training.

We report a number of findings of substantive importance. Our results suggest that individuals are appreciably *less* accurate than transcript files when they don't have any academic qualification, but that they are slightly *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 80% in both transcript files and self-reported data collected close to attainment of the qualification. This figure is 3 to 4% 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 7.7%.

We estimate the true return from achieving any academic qualification for those who do so as a 26.4% 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 misreporing 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 the resulting identification problem. Our estimation strategy for the case at hand is presented in Section 3. Section 4 discusses how information in the data will allow us to implement this strategy under fairly weak assumptions. It then 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. We also explore how the biases from misclassification and from omitted variables interact in the estimation of such a return. Section 6 concludes.

2 General formulation of the problem

2.1 Identification when the educational qualification is observed

In the potential-outcomes framework, interest lies in the causal impact of a given "treatment" on an outcome of interest Y.[3] To fix ideas, and 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.[4] Let

[3]For reviews of the evaluation problem see Heckman, LaLonde and Smith (1999) and Imbens (2004). For the potential outcome framework, the main references are Fisher (1935), Neyman (1935), Roy (1951), Quandt (1972) and Rubin (1974).

[4]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.

D^* be a binary indicator for the qualification of interest, which we will later allow to be potentially observed with error amongst individuals. The individual causal effect of (or return to) achieving the qualification is defined as the difference between the two potential outcomes, $\beta \equiv Y_1 - Y_0$. The observed individual wage can then be written as $Y = Y_0 + D^*\beta$. We are interested in recovering the average return for those individuals who have chosen to undertake the qualification of interest, that is the *average effect of treatment on the treated* (ATT):[5]

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

In the absence of misreporting of D^*, identification of the counterfactual term $E_{Y_0|D^*}[Y_0|1]$ follows straightforwardly from the following two assumptions, which we will maintain throughout.

Assumption 1 *(**Unconfoundedness**) Conditional on a set of observable variables X, the educational choice D^* is independent of the two potential outcomes:*

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

For the plausibility of this assumption, which allows one to focus on the impact of measurement error in the reporting of D^*, we draw on Blundell, Dearden and Sianesi (2005b), who find the set of regressors X available in our NCDS data to be rich enough to control for the endogeneity of educational choices. To give empirical content to Assumption 1, we also require the following condition on the support of the X variables:

Assumption 2 *(**Common Support**) Individuals with and without the qualification of interest can be found at all values of X, that is:*

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

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

Under these two assumptions one can perform any type of non- or semi-parametric estimation of the conditional expectation function in the non-participation group, $E_{Y_0|D^*X}[Y_0|0, x]$, and then average it over the distribution of X in the participants' group (within the common support) to get the counterfactual term of interest. Conditions 1-2 together make the *strong ignorability* condition of Rosenbaum and Rubin (1983).

[5] In the remainder of this paper, $f_{Y|X}[y|x]$ and $E_{Y|X}[Y|x]$ will denote the conditional distribution and the conditional mean of Y given $X = x$, respectively. Also, we will use upper-case letters for random variables and lower-case letters for their realisations.

2.2 Misclassified educational qualification

When qualifications are misreported, either because individuals are left to self-report or because of transcript errors, the treatment information recorded in the data may differ from the actual status D^*. With our application in mind, we assume to have *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 transcript records on the same individuals coming from the schools (D_T). It is worth noting the the former two measurements need not be independent of each other, as most likely they may be correlated through unobservables that affect the propensity of individuals to misreport. More in general, neither of self-reported and transcript measurements needs to coincide with D^*.

For any measurement $W = \{D_S^1, D_S^2, D_T\}$, define by $f_{W|D^*X}[1|1,x]$ the percentage of truth tellers, or of individuals correctly classified in transcript files, amongst those actually holding the qualification of interest. The corresponding percentage amongst those *without* the qualification of interest is instead defined as $f_{W|D^*X}[0|0,x]$. In the remainder of this paper, we will 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 $f_{\mathbf{D}_S|D^*X}[\mathbf{d}_S|1,x]$ and $f_{\mathbf{D}_S|D^*X}[\mathbf{d}_S|0,x]$ as the survey response patterns conditional on educational attainment, separately for those having and *not* having the qualification of interest, respectively. The definitions employed accommodate for error heterogeneity through the observable characteristics X.[6]

Throughout our discussion we will assume that the misclassification error in either measure is *non-differential*, that is conditional on a 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). This assumption is stated more formally in what follows.

Assumption 3 *(**Non-Differential Misclassification Given** X) 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 X:*

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

As shown in Battistin and Sianesi (2011), even under Assumptions 1-3 causal inference drawn from any of the triples (Y, D_S^1, X), (Y, D_S^2, X) or (Y, D_T, X) will in general be invalid for the ATT, with the magnitude of the bias depending on the extent of misclassification in each measurement. In what follows, we will maintain the assumption of independent sources of error between self-reported measurements and transcript files, conditional on the observables X.

[6] Note that the probabilities of D^* conditional on D_T or $\mathbf{D}_S$ could also be employed, signaling the percentage of achievers conditional on the educational status as it is observed in raw data (see Battistin and Sianesi, 2011).

Assumption 4 *(**Independent Sources of Error Given** X) The measurements* $\boldsymbol{D}_S$ *and* D_T *are conditionally independent given* D^* *and* X*:*

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

The assumption implies that qualifications self-reported by individuals and transcript files from schools are correlated only through the true measurement D^* and the observables X. This, together with Assumption 3, are assumptions widely adopted in the relevant literature. However, as pointed out by Hu (2008) and Battistin and Sianesi (2011), the conditioning on a large set of X's makes them weaker than those reviewed in Bound, Brown and Mathiowetz (2001).

The general identification problem induced by misclassification can be formalised as follows. Under Assumption 3 and Assumption 4, the distribution of observed wages conditional on X 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: the distribution of wages in the presence of the qualification, i.e. Y_1, and the distribution of wages in the absence of the qualification, i.e. Y_0. The mixture is:

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

where the equality follows from Assumption 1 and the probability:

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

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 X.

The mixture representation implies two results worth mentioning. First, knowledge of the mixture probabilities $p(\mathbf{d}_S, d_T, x)$'s suffices to identify the probabilities of *exact* classification relative to the self-reported measurements and transcript files. The result trivially follows from the Bayes theorem, after noting that their computation involves distributions that are identified from the data. Second, knowledge of the mixture components allows identification of:

$$\Delta^*(x) \equiv E_{Y_1|D^*X}[Y_1|1, x] - E_{Y_0|D^*X}[Y_0|1, x],$$

which corresponds, under Assumption 1, to the causal effect of having the qualification of interest for individuals with $X = x$. As the ATT is obtained by integrating $\Delta^*(x)$ with respect to $f_{X|D^*}[x|1]$, and the latter term is identified from knowledge of the $p(\mathbf{d}_S, d_T, x)$'s,[7] it follows that that the ATT is identified if the mixture in (1) is identified (see Battistin and Sianesi, 2011, for the exact characterisation

[7] There is:

$$f_{X|D^*}[x|1] = \frac{f_{D^*|X}[1|x] f_X[x]}{\int f_{D^*|X}[1|x] f_X[x] dx},$$

of the relationship between the true ATT, the effect estimated using either misrecorded measure and the latter's misclassification probabilities).

In the next section we show that, for the case at hand, the information in the data is sufficient to retrieve non-parametrically both mixture weights and mixture components.

2.3 Identification in the presence of misclassification

With two reports, Kane, Rouse and Staiger (1999) and Black, Berger and Scott (2000) have developed a procedure to simultaneously estimate the returns to qualifications and the distribution of reporting error in each educational measure. Their approach moves from the specification of a parametric model. The general problem of non-parametric identification in the case of two reports has been investigated, amongst others, by Mahajan (2006), Lewbel (2007) and Hu (2008). The returns to qualifications and the extent of misclassification are point identified by assuming that the two available measurements come from *independent* sources of information. This implies that the extent of misclassification must be independent across measurements, and qualifies one of these - provided additional conditions hold - as an instrument-like variable for the problem.

We build upon this idea to show that the components of the mixture in (1) are non-parametrically identified given the setup that we consider. Key to our identification result is Assumption 4. Although *three* measurements of educational qualifications are available in our data, one can always reduce the dimensionality problem by generating a new variable $\tilde{D}$ which results from the combination of D_S^1 and D_S^2, for example by considering $\tilde{D} \equiv D_S^1 D_S^2$ or $\tilde{D} \equiv D_S^1(1 - D_S^2)$. In this case, the two new measurements $(\tilde{D}, D_T)$ are sufficient to retrieve the returns to qualifications non-parametrically as in Mahajan (2006), Lewbel (2007) and Hu (2008). The availability of multiple self-reported measurements introduces a certain degree of over-identification, and allows one 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 identification result builds upon the following additional assumptions, that closely match those exploited in the relevant literature (see, for example, Chen, Hong and Nekipelov 2011). The general idea behind identification is to use D_T as a source of instrumental variation which, through Assumption 4, allows one to define a large enough number of moment conditions given the unknowns in

---

which is identified using:

$$f_{D^*|X}[1|x] = \sum_{\mathbf{d}_S} \sum_{d_T} p(\mathbf{d}_S, d_T, x) f_{\mathbf{D}_S D_T|X}[\mathbf{d}_S, d_T|x],$$

if the $p(\mathbf{d}_S, d_T, x)$'s are known.

the mixture representation (1). The availability of multiple reports coming from the same individuals sets the stage for additional moment restrictions, that can be used to allow for correlation in self-reported measurements.

Assumption 5 *(**Relevance of Educational Qualifications Given** X) The causal effect of having the qualification of interest for individuals with $X = x$ is such that:*

$$\Delta^*(x) \neq 0.$$

This assumption implies that the latent measurement D^* is relevant for the policy parameter under consideration at all values X. Following the discussion in the previous section, 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 $f_{Y|D^*X}[y|d^*, x]$. Intuitively, this assumption is required to disentangle the mixture distributions in (1) when estimation is carried out from raw data.

The next assumption requires that the measurement D_T contains enough information on the true educational qualification D^* given X or, more formally, that $f_{D^*|D_TX}[1|1, x] \neq f_{D^*|D_TX}[1|0, x]$ (see Chen, Hong and Nekipelov 2011). For the binary case considered in this paper, a sufficient condition for this to hold is the following.

Assumption 6 *(**Informational Content of the Transcript Measurement Given** X) The extent of misclassification in the measurement D_T is such that $f_{D^*|D_TX}[1|1, x] > 0.5$ and $f_{D^*|D_TX}[0|0, x] > 0.5$.*

This assumption is typically invoked in the literature and is indeed very reasonable, as it implies that information from the school files is more accurate than pure guesses once X 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 X (see Appendix A).

Assumption 7 *(**Relevance of Survey Instruments**) For each value x on the support of X there is: $f_{\boldsymbol{D}_S D_T|X}[\boldsymbol{d}_S d_T|x] \neq f_{\boldsymbol{D}_S|X}[\boldsymbol{d}_S|x] f_{D_T|X}[d_T|x]$.*

The general identification result can be summarized in the following theorem, for which the proof is given in Appendix A, and particularizes to the setup considered in our application previous results by Hu (2008).

Theorem 1 ***(Identification)*** *The mixture components* $f_{Y_0|X}[y|x]$ *and* $f_{Y_1|X}[y|x]$ *and the mixture weights* $p(\boldsymbol{d}_S, d_T, x)$ *are non-parametrically identified from the data* $(Y, \boldsymbol{D}_S, D_T, X)$ *under Assumptions 1 - 7.*

3 Estimation

Having proved that information on $(Y, \mathbf{D}_S, D_T, X)$ ensures non-parametric identification of the ATT 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 key assumptions will be maintained throughout the estimation process. First, we will assume that the mixture components are normally distributed, and propose a method that estimates (1) directly via MCMC. 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. Note also that this is in the spirit of the work by Heckman and Honore (1990), where it is shown that under normality it is possible to estimate the distribution of potential wages in the Roy model from a single cross-section of data (see also the discussion by Heckman, 2001). To reduce the dimensionality problem that results from having a large number of X's, we implement a semi-parametric estimator that makes use of the concept of balancing scores taken from the programme evaluation literature (see Battistin and Sianesi, 2011, for an application of the same idea). The second assumption we make is that the mixture weights are heterogeneous across individuals only through functions of the X's that can be estimated from raw data. The estimation procedure employed will be discussed in the remainder of this section.

3.1 The curse of dimensionality

The main problem that hampers estimation of the ATT is the curse of dimensionality arising from a large number of regressors in X. In this section we propose a method to reduce the dimensionality of the problem based on the properties of *balancing scores* (see Theorem 1 by Rosembaum and Rubin, 1983, and Imbens, 2000). Let $\mathcal{S}(X)$ be a balancing score such that the distribution of X within cells defined by $\mathcal{S}(x)$ is independent of $(\mathbf{D}_S, D_T)$:

$$f_{X|\mathbf{D}_S D_T \mathcal{S}(X)}[x|\mathbf{d}_S, d_T, s] = f_{X|\mathcal{S}(X)}[x|s]. \tag{2}$$

In what follows, we discuss under which conditions the mixture representation given X in (1) implies a mixture representation given $\mathcal{S}(X)$. The idea is most simply put across by assuming that the $p(\mathbf{d}_S, d_T, x)$'s do not vary with X, that is by assuming $p(\mathbf{d}_S, d_T, x) = p(\mathbf{d}_S, d_T)$.

By using (2) and from the fact that X is finer than $\mathcal{S}(X)$ we can write:

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

Using (1) and the fact that the $p(\mathbf{d}_S, d_T, x)$'s do not vary with X, the term on the right-hand-side of the last expression can be written as:

$$[1 - p(\mathbf{d}_S, d_T)] \int f_{Y_0|X}[y|x] f_{X|\mathcal{S}(X)}[x|s] dx + p(\mathbf{d}_S, d_T) \int f_{Y_1|X}[y|x] f_{X|\mathcal{S}(X)}[x|s] dx,$$

so that there is:

$$f_{Y|\mathbf{D}_S D_T \mathcal{S}(X)}[y|\mathbf{d}_S, d_T, s] = [1 - p(\mathbf{d}_S, d_T)] f_{Y_0|\mathcal{S}(X)}[y|s] + p(\mathbf{d}_S, d_T) f_{Y_1|\mathcal{S}(X)}[y|s],$$

where the last relationship again follows from X being finer than $\mathcal{S}(x)$. Accordingly, the distribution of wages conditional on $\mathcal{S}(X) = s$ for the group defined by all combinations of $(\mathbf{D}_S, D_T)$ is again a mixture of two latent distributions. The components of this mixture are weighted means of the components in (1) taken over individuals with $\mathcal{S}(X) = s$, with mixture weights given by $p(\mathbf{d}_S, d_T)$. Note that the same representation would hold if the $p(\mathbf{d}_S, d_T, x)$'s were left to vary with X only through the index $\mathcal{S}(x)$, i.e. by assuming $p(\mathbf{d}_S, d_T, x) = p(\mathbf{d}_S, d_T, s)$:

$$f_{Y|\mathbf{D}_S D_T \mathcal{S}(X)}[y|\mathbf{d}_S, d_T, s] = [1 - p(\mathbf{d}_S, d_T, s)] f_{Y_0|\mathcal{S}(X)}[y|s] + p(\mathbf{d}_S, d_T, s) f_{Y_1|\mathcal{S}(X)}[y|s]. \tag{3}$$

The identification problem is similar to the one described in the previous section: if (3) can be recovered from raw data, then one could identify the extent of misreporting and, therefore, the ATT.

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 $e_g(x) \equiv f_{G|X}[g|x]$. 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. This extends to the multinomial case the original idea introduced by Rosenbaum and Rubin (1983) for the binary case.

3.2 Bayesian modelling and inference

In the previous section we have shown that, for the case at hand, the mixture representation holds conditionally on the $e_g(x)$'s if these are the only factors driving heterogeneity of the $p(\mathbf{d}_S, d_T, x)$'s. We now build on this assumption to estimate the mixture (3).

We will assume throughout 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 the balancing score: given the nature of the outcome variable, this appears to be a sound specification for the case at hand. Importantly, it can be shown that any finite mixture of univariate normal distributions is identifiable (see, for example, Everitt and Hand, 1981) and this has some implications that are discussed in what follows.[8]

Under the hypothesis of no returns conditional on $\mathcal{S}(x)$ the two mixture components coincide, and thus the mixture representation is invalid. This is known as the problem of *homogeneity*, and is ruled out by Assumption 5. Note that testing homogeneity, that is testing no mixture against a mixture of two distributions, is a non-regular problem, in that the null hypothesis belongs to the boundary of the parameter space. However, using the results in Yakowitz and Spragins (1968), it follows that any non-degenerate finite mixture of normal distributions cannot itself be normal. It follows that, in our application, testing Assumption 5 under the maintained assumption of normal components and non-degenerate $p(\mathbf{d}_S, d_T, x)$'s amounts to testing normality of the observed wage distributions.[9]

The mixture in (3) is estimated through a MCMC procedure, which is fully documented in Appendix B and whose main features can be described as follows.[10] Let $\mathbf{e}(x) = [1, e_2(x), \ldots, e_8(x)]'$ be the 8×1 vector containing the balancing scores. We set:

$$\begin{aligned} Y_i|\mathbf{e}(x) &\sim N(\boldsymbol{\theta}_{\boldsymbol{i}}'\boldsymbol{e}(x), \sigma_i^2), \qquad i = 0, 1 \\ p(\boldsymbol{d}_S, d_T, \boldsymbol{e}(x)) &= \Phi(\boldsymbol{\gamma}_{\boldsymbol{g}}'\boldsymbol{e}(x)), \qquad g = 1, \ldots, 8 \end{aligned}$$

for mixture components and mixture weights, respectively, where $\Phi(\cdot)$ is the standard normal distribution function. The former equation defines the 8×1 vectors of parameters $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$, and the scalars σ_0^2 and σ_1^2. The latter equation defines the 8×1 vector $\boldsymbol{\gamma_g}$ for any combination $D_T \times D_S^1 \times D_S^2$. Overall, this specification defines 82 unknowns that fully characterise the mixture (3).

We specify a joint prior distribution for these parameters, and we use a Gibbs sampling algorithm to obtain $2,000$ realizations from their joint posterior distribution. The posterior distributions for the unknown quantities of the mixture representation (3) can easily be computed using these realizations.

[8] 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.

[9] We implemented a simple test for this hypothesis through a suitably defined set of regressions. Within cells defined by the cross tabulation of the three measurements of educational attainment, we regressed logged wages on the balancing scores, and tested for the normality of residuals. The results of this procedure, which are available upon request, are overall against the normality of logged wages.

[10] It is worth noting that the estimation results proved informationally similar to those obtained in a previous version of this paper, where maximum likelihood estimation via the EM algorithm was employed.

Knowledge of these quantities is in turn sufficient to obtain estimates of the probabilities of exact classification and of the ATT.

4 Data and educational qualifications of interest

4.1 Data

In this paper we only consider methods relying on Assumption 1, and we thus require very rich background information capturing all those factors that jointly determine the attainment of educational qualifications and wages. We use the uniquely rich data from the British National Child Development Survey (NCDS), a detailed longitudinal cohort study of all children born in a week in March 1958 which contains extensive and commonly administered ability tests at early ages (mathematics and reading ability at ages 7 and 11), accurately measured family background (parental education and social class) and school type variables. In fact, Blundell, Dearden and Sianesi (2005b) could not find evidence of remaining selection bias for the higher education *versus* anything less decision once controlling for the same variables we use in this paper. We thus invoke their conclusion in assuming that there are enough variables to be able to control directly for selection.

Our outcome is real gross hourly wages at age 33. As to educational attainment, of particular interest to our purposes is that cohort members were asked to report the qualifications they had obtained as of March 1981 not only in the 1981 Survey (at age 23), but also in the 1991 Survey (at age 33).[11] We can thus construct two separates measures of qualifications obtained up to March 1981, based either on responses in the 1981 or in the 1991 survey. Furthermore, 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).[12] For each individual we thus have *three* measurements, which - as we argue in the next section - can all be taken as proxies of educational qualifications acquired by age 20. These are the measurements that we will consider to implement the strategy that was described in Section 3.

We focus 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 $2,716$ observations, which is the same sample used by Battistin and Sianesi

[11] 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.

[12] 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.

[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).

(2011).

4.2 Educational qualifications of interest

Non-parametric identification of the 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 by the School Files, the latter however only recording academic qualifications and only those achieved by age 20 - that is Ordinary Levels (O levels) and Advanced Levels (A levels).[14]

Although driven by the availability of an independent school measure for O and A levels only, focusing on academic qualifications offers clear advantages, and allows one to estimate highly policy relevant parameters. First, academic qualifications are well defined and homogenous, 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.[15] A second advantage of focusing on O and A 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.

A highly policy relevant parameter, and the one we focus on in our application, is the return from attaining any academic qualification (that is, from acquiring at least O levels) compared to leaving school at the minimum age of 16 without any formal qualification.[16] Special interest in 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. In particular, Chevalier *et al.* (2003) 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.

[14] 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.

[15] 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.

[16] 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. The "None" category also includes very low-level qualifications at NVQ level 1 or less, in particular the academic CSE grade 2 to 5 qualifications. Students at 16 could take the lower-level Certificates of Secondary Education (CSE) or the more academically demanding O levels. The top grade (grade 1) achieved on a CSE was considered equivalent to an O level grade C. Most CSE students tended to leave school at 16.

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. Indeed, Blundell, Dearden and Sianesi (2005b) found a non-negligible return of 18% for those who did leave with O levels and of 13% for those who dropped out at 16 without any qualifications. Furthermore, the return to acquiring at least O levels compared to nothing captures all the channels in which the attainment of O levels at 16 can impact on wages later on in life, in particular the potential contribution that attaining O levels may give to the attainment of A levels and then of higher education.

Having defined the parameter of interest, it is important to highlight the condition that allows us to have repeated measurements of achievement at age 20 coming from both school records and NCDS survey reports. As 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, we 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.[17]

4.3 Evidence from the raw data

Table 1 presents estimates of the individual wage returns to any academic qualification using three different methods (simple dummy variable OLS, fully interacted regression model and propensity score matching), two sets of control variables (the full set of observables including 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 measures of the treatment of interest, i.e. of having obtained any academic qualification by age 20.

As in Blundell, Dearden and Sianesi (2005b), we find that while results change little in response to the method used to control for selection on observables, controlling for ability test scores at an early age and detailed family background measures is crucial, and significantly reduces the returns to a 15 to 28% wage gain depending on the educational measure used. As to the latter, it is indeed striking that in the more flexible models (fully interacted linear model and matching), using an educational measure rather than another gives rise to returns which exhibit the same magnitude of bias as from omitted controls. In particular, matching yields estimates of returns which range from as low as 14.2% (self-reported 13 years after attainment) to as high as 27.8% (self-reported 3 years after attainment), with returns estimated from school files falling in between (23.9%). For all three estimation methods - and irrespective of the set of control variables being used - the estimates using self-reported measures at different times, as well as those using transcript vs recall information are significantly different (99%

[17]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.

level - see the right hand side panel of Table 1). Estimates arising from measures obtained close to completion (i.e. transcript and self-reports at age 23) are by contrast not statistically different.

To investigate such substantial differences in estimated returns according to the educational report being used, Table 2 presents cross tabulations between the three underlying measurements. We find that the percentage of the sample where the three measures all agree is 82%. In what follows, we will refer to this statistic as the "agreement rate". Despite this being quite high, 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[18]), 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).

One can follow Mellow and Sider (1983) and perform a descriptive analysis of the determinants of concordance across indicators of educational attainment. In results not shown, we find that only a couple of measured characteristics seem to matter in predicting agreement rates. In particular, having a father whose social class is professional is associated with a higher probability of agreement between the two individual's self-reports, and consequently among all three reports. Higher ability as measured by mathematical test scores at 11 is associated with a higher probability of agreement between self-reported and school information, the link being particularly strong close to completion, but remaining significant 10 years on. This association also means that high ability individuals have

[18]The kappa-statistic measure of interrater agreement is scaled to be 0 when the amount of agreement is what would be expected to be observed by chance and 1 when there is perfect agreement (see Fleiss, 1971).

a higher likelihood of the three measures being congruent. Finally, school-type variables were found to be associated with the degree of concordance, with some types of schools (secondary modern and comprehensive) being associated with lower overall agreement rates. Overall, observed characteristics are thus found to have a very low predictive power of the degree of concordance, this being particularly the case when trying to infer the likelihood that information from the school files close to completion agrees with self-reported information 10 years later (all the control variables jointly explain 3.9% of the variance). By contrast, observables matter more in modelling the probabillity that individuals and schools agree close to the attainment of the qualification of interest.

In conclusion, even though formal statistics like the kappa measure of interrater 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 will be assessed empirically in the next section.

5 Results

This Section presents our empirical results on the extent and features of misclassification, as well as on the true return to academic qualifications, that is one which takes into account the misreporting uncovered in the data. We also explore how the biases from misclassification and from omitted variables interact in the estimation of such a return. We first define the quantities needed to characterize misreporting across survey and transcript measurements. To ease readability, the conditioning on observables X will be left implicit throughout.[19]

5.1 Summary of the quantities retrieved

For each measurement $W = \{D_S^1, D_S^2, D_T\}$, we start by considering the two probabilities of *exact classification* $f_{W|D^*}[0|0]$ and $f_{W|D^*}[1|1]$ (see Section 2.2 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 $W = D^*$)

[19] As discussed in Section 3, heterogeneity along observable dimensions is modeled by conditioning on the propensity scores $e_g(x)$. Thus, one can always view the quantities that will follow as the result of averaging out individual heterogeneity using the distribution of the $e_g(x)$'s in the population.

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 will be presented in Table 3.

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 among those with the qualification (their share given by $f_{\mathbf{D}_S|D^*}[1,1|1]$) and consistent truth tellers among those without the qualification (their share given by $f_{\mathbf{D}_S|D^*}[0,0|0]$). 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 will be 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, should reveal 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 don't 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 Characterising the extent of misclassification

The first three panels of Figure 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 the sample using the methodology described in Section 3.2. As for each individual our procedure yields $2,000$ realizations from the posterior distribution of the quantity of interest, all distributions in Figure 1 are obtained by first taking the individual average of these realizations, and then plotting the distribution of such averages across individuals. The distributions on the left hand side are in general more disperse than the corresponding distributions on the right hand side. 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 don't 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 Figure 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 summarise the extent of misclassification/over-reporting for individuals without academic qualifications as being 16% in the school files, but as high as 27% and 31% in the 1981 and 1991 surveys. Thus while the degree of accuracy of self-reported measurements seems to be between 11% to 15% lower when compared to transcript records, we find only a small effect of the time of reporting for individuals without the qualification of interest (i.e. the survey closer to completion is only 4% percentage points more accurate than the survey 10 years later).

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 Figure 1, and the first row of Table 3). Individuals with qualifications are between 3% to 7% more likely than schools to report correctly their attainment, again pointing to little, or no, survey wave effect.

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 80%, for the 1981 wave at 80.3% and for the 1991 wave at 76.5% (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% lower when the information is collected up to 10 years after the qualification was attained.

Using the misclassification probabilities, we recovered an estimate of the true incidence of academic qualifications in the population, namely $f_{D^*}[1]$, of 64.1%. Interestingly, while being substantially higher than the incidence according to school files (58.8%), this estimate coincides with the incidence according to 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 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 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 (76.9%) than amongst those who do not (63.1%). This is graphically corroborated by the corresponding distributions across individuals presented in the bottom panel of Figure 1. Overall, we calculated that the percentage of truth tellers represents almost three quarters (71.9%) 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, 90.8% (= 0.769/0.847) of individuals who report so correctly in wave 1 will also report correctly in wave 2 and 94.8% (= 0.769/0.811) 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 lower (86.6% and 91.8%). 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 show however 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 13 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 that we find clearly points to a different pattern (for those with the qualification, $0.769 > 0.687 = 0.847 \times 0.811$; for those without the qualification, $0.631 > 0.501 = 0.729 \times 0.687$), highlighting the presence of positive autocorrelation in measurements after controlling for D^*.[20]

Consistent over-reporters appear to be an important fraction of the no-qualification sample: one in five (19.6%) 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 (27.1% and 31.3% of the no-qualification samples in the 1981 and in the 1991 surveys, respectively). These two sets of results thus suggest that around one third (28 to 37%) 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 11.2%, it is almost half the size. As was the case for over-reporting, focusing on one survey wave alone would overstate the amount of under-reporting. Interestingly, 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 almost identical to the one that accounted for over-reporting errors (27 to 40%), 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 (17%) than among the qualification group (12%). 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 7.7%, and in the NCDS sample at 5%.

5.3 Returns to any academic qualification

With the misclassification probabilities in hand, we can then proceed to estimate the true ATT from achieving any academic qualification as outlined in Section 3. Throughout this section, the following notation will be employed. Δ^*_{FULL} and Δ^*_{LFS} denote estimates that are adjusted for misclassification

[20] 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 conditional independent given D^* and X, as there must be at least one value of X such that the latter assumption is violated. Figure 6 in 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.

and employ either full set of controls available in the NCDS or the LFS-style variables. Similarly, estimates obtained from raw data without controlling for misclassification will be denoted by Δ_{FULL} and Δ_{LFS}.

The most reliable estimate for the ATT (Δ^*_{FULL}) is a 26.4% wage gain from achieving at least O-levels, with a posterior standard deviation of 0.065. 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 0.043 (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 43% upward bias in estimated returns that do not fully control for selection into educational attainment.

To put these estimates in context, Table 5 displays the new results together with our OLS estimates from Table 1. In the following, we focus on the OLS estimates as the fully interacted regression model (FILM) did not provide evidence for heterogeneous returns. It follows that in the remainder of this section Δ_{FULL} or Δ_{LFS} will refer to point estimates obtained through OLS regressions. In order to heuristically compare frequentist and Bayesian estimates, we constructed 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}.

5.3.1 Estimating returns based on educational reports that were obtained relatively close to the attainment of the qualification of interest

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%. Correcting for selection bias using our rich set of observed background characteristics reduces the estimated ATT (Δ_{FULL}) to 19%. The value Δ^*_{FULL} thus appears to be bound below from the estimate that controls only for selection bias and above from the LFS-style estimate that controls for neither source of bias. Both these estimates are significantly different from the true return and would provide a misleading picture of how much people with academic qualifications have gained by investing in education.

What can we say about the relative importance of omitted ability and measurement error biases, and about the possibilities that the two cancel out when the qualification is recorded close to its attainment? By comparing the true return (Δ^*_{FULL}) to the one ignoring both types of potential biases (Δ_{LFS}), we 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 over one quarter (26%). This result

is reassuringly consistent with the findings in Battistin and Sianesi (2011), who bound the ATT of interest semi-parametrically and find that ignoring both misreporting and omitted ability bias would generally lead to at times quite severely upward biased estimates of true returns.

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. It has to be noted that these conclusions apply equally to education measured by the school as well as self-reported by the individuals themselves.

As to the relative importance of ability and measurement error biases, we find that while both sources of bias give rise to estimates that are significantly different from the true return, the bias arising from omitted ability controls is larger. In particular, we have shown above how estimates that correct for measurement error but not for omitted ability incur a 43% upward bias, whilst controlling for ability but ignoring misclassification error in concurrent reports leads to a 27% downward bias both in the case of self-reported measure and of school transcripts.

To conclude, in a situation where educational records were obtained relatively close to the completion of the qualification of interest, we find that the policymaker or analyst cannot simply rely on measurement error to cancel out the ability bias.

5.3.2 Estimating returns based on educational reports that rely on recalling the attainment of the qualification of interest over more than 10 years

We now turn to consider a situation in which the educational information recorded in the data has been collected after over 10 years since completion. Since in line with *a priori* expectations we have found the recall measure to suffer from a larger extent of measurement error, we now expect the relative importance of omitted variable bias and measurement error bias to shift.

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 almost halved once we control for the full set of observables (Δ_{FULL} being equal to 15.1%). However, once we compare these estimates to the true return (Δ^*_{FULL}) of 26.4%, 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 43% upward bias (compare Δ^*_{FULL} to Δ^*_{LFS}), controlling for selection but ignoring misclassification gives rise to a bias of exactly the same size (43%) 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.

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. We have additionally identified the extent of misreporting in different types of commonly used data sources on educational qualifications: exam transcript files 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.

We have also provided evidence on the relative importance of ability and measurement error biases, and 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).

Results in this paper thus represent a new piece of evidence for the UK policy community, which will allow one to appreciate the relative reliability of different sources of educational information as well as check 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.

TABLE 1. Estimates of the returns to any academic qualification.

	(1) Transcript files from schools	(2) 1981 Wave (at age 23)	(3) 1991 Wave (at age 33)	tests of equality (1)=(2)	(1)=(3)	(2)=(3)
LFS controls only:						
OLS	0.332	0.333	0.293		***	***
	(0.015)	*(0.016)*	*(0.016)*			
FILM	0.330	0.336	0.289		***	***
	(0.015)	*(0.016)*	*(0.016)*			
PSM	0.331	0.336	0.285		***	***
	(0.015)	*(0.015)*	*(0.016)*			
Full set of controls:						
OLS	0.194	0.194	0.151		***	**
	(0.018)	*(0.018)*	*(0.018)*			
FILM	0.216	0.241	0.137		**	***
	(0.030)	*(0.033)*	*(0.026)*			
PSM	0.239	0.278	0.142	*	***	***
	(0.033)	*(0.032)*	*(0.025)*			

Note. Reported are estimates of the average treatment effect on the treated (ATT) obtained by controlling only for the LFS set of variables (gender and age, ethnicity and region) or the full set of variables used by Blundell, Dearden and Sianesi (2005) (LFS-controls plus math and reading ability test 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) and propensity score kernel matching (PSM). *Right-hand side panel*: the corresponding columns are significantly different at the 99% (***), 95% (**) and 90% (*) level, based on bootstrapped bias-corrected confidence intervals.

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	**1991 Wave (at age 33)**		**1991 Wave (at age 33)**	
(at age 23)	**Any**	**None**	**Any**	**None**
Any	1445	103	148	70
None	24	25	120	781
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 (κ_{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 4.3 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).

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	0.783	0.847	0.811
	(0.029)	*(0.03)*	*(0.028)*
No qualification	0.836	0.729	0.687
	(0.067)	*(0.061)*	*(0.057)*
Correct classification	0.800	0.803	0.765
	(0.033)	*(0.033)*	*(0.031)*

Note. The table presents estimates of the probabilities of exact classification for the three survey instruments. *Top Panel*: the row labeled *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 labeled *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.

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	0.769	0.631
	(0.028)	*(0.053)*
Over reporters		0.196
		(0.065)
Under reporters	0.112	
	(0.03)	
Confused	0.118	0.172
	(0.017)	*(0.028)*

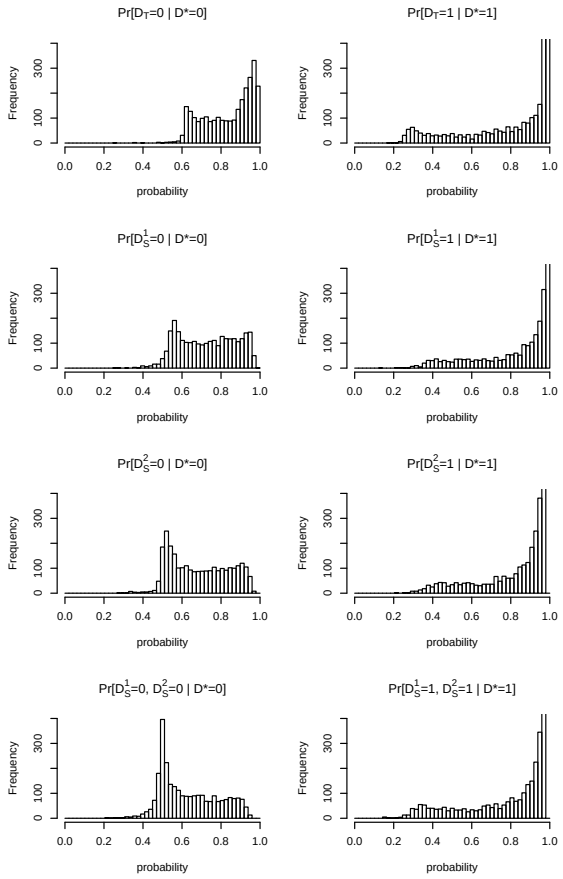
Note. The table presents estimates of the percentage of individuals who consistently report correctly (*truth tellers*), over-report (*over-reporters*) and under-report (*under-reporters*) their educational qualification across survey waves. Presented also is the percentage of individuals with inconsistent 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.

TABLE 5. Comparison of estimates of returns to educational attainment.

$\boldsymbol{\Delta^*_{FULL}}$		0.264 *(0.065)*	
	Transcript files from schools	**1981 Wave (at age 23)**	**1991 Wave (at age 33)**
$\boldsymbol{\Delta_{LFS}}$	0.332 *(0.015)*	0.333 *(0.016)*	0.293 *(0.016)*
p-value: $\Delta_{LFS} = \Delta^*_{FULL}$	0.000	0.000	0.070
$\boldsymbol{\Delta_{FULL}}$	0.194 *(0.018)*	0.194 *(0.018)*	0.151 *(0.018)*
p-value: $\Delta_{FULL} = \Delta^*_{FULL}$	0.000	0.000	0.000

Note. The top panel of the table reports the ATT computed as described in Section 3 (Δ^*_{FULL}), 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 are to test the equality of the two estimates (see Section 5 for a description of how the test was implemented).

FIGURE 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).

References

[1] Aigner, D. (1973), *Regression with a Binary Independent Variable Subject to Errors of Observation*, Journal of Econometrics, 1, 49-60.

[2] Battistin, E. and Sianesi, B. (2011), *Misclassified Treatment Status and Treatment Effects: An Application to Returns to Education in the UK*, Review of Economics and Statistics, 93, 2, 495-509.

[3] Black, D., Berger, M. , and Scott, F. (2000), *Bounding Parameter Estimates with Non-Classical Measurement Error*, Journal of the American Statistical Association, 95, 451, 739-48.

[4] Black, D., Sanders, S., and Taylor, L. (2003), *Measurement of Higher Education in the Census and Current Population Survey*, Journal of the American Statistical Association, 98, 463, 545-554.

[5] Blundell, R., Dearden, L., and 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, pp. 117-145. ISBN 0691117349.

[6] Blundell, R., Dearden, L., and Sianesi, B. (2005b), *Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey*, Journal of the Royal Statistical Society A, 168, 3, 473-512.

[7] Bonjour, D., Cherkas, L., Haskel, J., Hawkes, D., and Spector, T. (2003), *Returns to Education: Evidence from UK Twins*, American Economic Review, 93, 5, 1799-1812.

[8] Bound, J., Brown, C., and Mathiowetz, N. (2001), *Measurement error in survey data*, in J.J. Heckman and E. Leamer (eds.), *Handbook of Econometrics. Vol. 5*, Amsterdam: North-Holland, 3705-3843.

[9] Card, D. (1999), *The Causal Effect of Education on Earnings*, Handbook of Labor Economics, Volume 3, Ashenfelter, A. and Card, D. (eds.), Amsterdam: Elsevier Science.

[10] Chen, X., Hong, H., and Nekipelov, D. (2011), *Nonlinear Models of Measurement Errors*, forthcoming in the Journal of Economic Literature.

[11] Chevalier, A., Harmon, C., Walker, I., and Zhu, Y. (2004), *Does education raise productivity, or Just Reflect It?*, The Economic Journal, 114, 499, 499-517.

[12] Dearden, L. (1999), *Qualifications and earnings in Britain: how reliable are conventional OLS estimates of the returns to education?*, IFS working paper W99/7.

[13] Dearden, L., McIntosh, S., Myck, M., and Vignoles, A. (2002), *The Returns to Academic and Vocational Qualifications in Britain*, Bulletin of Economic Research, 54, 249-274.

[14] Del Bono, E., and 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.

[15] Everitt, B.S., and Hand, D.J. (1981), Finite Mixture Distributions, Chapman and Hall: London

[16] Fisher, R.A. (1935), *The Design of Experiments*, Edinburgh: Oliver&Boyd.

[17] Fleiss, J.L. (1971) *Measuring nominal scale agreement among many raters*, Psychological Bulletin, Vol. 76, No. 5 pp. 378Ű-382.

[18] Griliches, Z. (1977), *Estimating the returns to schooling: some econometric problems*, Econometrica, 45, 1-22.

[19] Heckman, J.J. (2001), *Micro data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture*, Journal of Political Economy, Vol. 109, No. 4, 673-748.

[20] Heckman, J.J., and Honore, B.E. (1990), *The Empirical Content of the Roy Model*, Econometrica, Vol. 58, No. 5, 1121-1149.

[21] Heckman, J.J., Lalonde, R., and Smith, J. (1999), *The Economics and Econometrics of Active Labor Market Programs*, Handbook of Labor Economics, Volume 3, Ashenfelter, A. and Card, D. (eds.), Amsterdam: Elsevier Science.

[22] Hu, Y. (2008),*Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution*, Journal of Econometrics, Vol. 144, No. 1, 27-61.

[23] Imbens, G.W. (2000), *The Role of the Propensity Score in Estimating Dose-Response Functions*, Biometrika, 87, 3, 706-710.

[24] Imbens, G.W. (2004), *Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review*, Review of Economics and Statistics, 86, 4-29.

[25] Ives, R. (1984), *School reports and self-reports of examination results*, Survey Methods Newsletter, Winter 1984/85, 4-5.

[26] Kane, T.J., Rouse, C., and Staiger, D. (1999), *Estimating Returns to Schooling when Schooling is Misreported*, National Bureau of Economic Research Working Paper No. 7235.

[27] Landis, J.R., and Koch, G.G. (1977), *The Measurement of Observer Agreement for Categorical Data.*, Biometrics, 33, 159-174.

[28] Lechner, M. (2001), *Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption*, in M. Lechner and F. Pfeiffer (Eds.), Econometric Evaluation of Labour Market Policies, Heidelberg: Physica, 43-58.

[29] Lewbel, A. (2007), *Estimation of Average Treatment Effects With Misclassification*, Econometrica, 75, 2, 537-551.

[30] Mahajan, A. (2006), *Identification and Estimation of Regression Models with Misclassification*, Econometrica, 74, 3, 631-665.

[31] McIntosh, S. (2006), *Further Analysis of the Returns to Academic and Vocational Qualifications*, Oxford Bulletin of Economics and Statistics, 68, 2, 225-251.

[32] Mellow, W., and Sider, H. (1983). *Accuracy of Response in Labor Market Surveys: Evidence and Implications*, Journal of Labor Economics, 1(4), 331-344.

[33] Molinari, F. (2008), *Partial Identification of Probability Distributions with Misclassified Data*, Journal of Econometrics, 144, 1, 81-117.

[34] Neyman, J. (with co-operation by Iwaszkiewicz, K., and Kolodziejczyk, S.) (1935), *Statistical Problems in Agricultural Experimentation*, Supplement of the Journal of the Royal Statistical Society, 2, 107-180.

[35] Quandt, R. (1972), *Methods for Estimating Switching Regressions*, Journal of the American Statistical Association, 67, 306-310.

[36] Roy, A. (1951), *Some Thoughts on the Distribution of Earnings*, Oxford Economic Papers, 3, 135-146.

[37] Rosenbaum, P.R., and Rubin, D.B. (1983), *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika, Vol. 70, No. 1, 41-55.

[38] Rubin, D.B. (1974), *Estimating Causal Effects of Treatments in Randomised and Non-randomised Studies*, Journal of Educational Psychology, 66, 688-701.

[39] Rubin, D.B. (1980), *Discussion of 'Randomisation analysis of experimental data in the Fisher randomisation test' by Basu*, Journal of the American Statistical Association, 75, 591–3.

[40] Yakowitz, S.J., and Spragins, J.D. (1968), On the Identifiability of Finite Mixtures, Annals of Mathematical Statistics, 39, 209-214.

ADDENDA
(not for publication)

A 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. The result in what follows generalizes 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 $X = x$ will be left implicit throughout.

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

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

Define the following *latent* matrices:

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

which are characterized by 10 unknowns.

Using Assumption 3 and assumption 4 there is:

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

or, in matrix notation:

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

$$\mathcal{F}_{\mathbf{D}_S|D_T} = \mathcal{F}_{D^*|D_T}\mathcal{F}_{\mathbf{D}_S|D^*}. \tag{5}$$

Now, under Assumption 6 the matrix $\mathcal{F}_{D^*|D_T}$ is nonsingular (i.e. full rank), so that from (5) there is:

$$\mathcal{F}_{\mathbf{D}_S|D^*} = \mathcal{F}^{-1}_{D^*|D_T}\mathcal{F}_{\mathbf{D}_S|D_T}, \tag{6}$$

which if substituted into (4) 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 $\mathcal{F}_{\boldsymbol{D}_S|D^*}$ 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 7. The above expression defines a *singular value decomposition* (see Seber, 2008, page 334), implying that the mixture components on the main diagonal of the *latent* matrix $\mathcal{F}_{Y|D^*}$ can be obtained as the singular values of the *known* matrix $\mathcal{M} \equiv \mathcal{F}_{Y\mathbf{D}_S|D_T}\mathcal{F}^+_{\mathbf{D}_S|D_T}$. Additional assumptions need to be imposed to establish a correspondence between the eigenvalues and the eigenvectors of $\mathcal{M}$, the latter here been represented by the the matrix of misclassification probabilities $\mathcal{F}_{D^*|D_T}$.

Assumption 5 ensures that there exist an ordering of the eigenvalues in the diagonal matrix $\mathcal{F}_{Y|D^*}$, while Assumption 6 guarantees that $\mathcal{F}_{D^*|D_T}$ is a diagonally dominant matrix, hence characterizing the order of the eigenvectors. The above argument proves identification of the $f_{Y|D^*}[y|d^*]$'s, and of the $f_{D^*|D_T}[d^*|d_T]$'s. Knowledge of the latter probabilities implies, via (6), 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 independence assumption between D_T and D_S is maintained. The proof would proceed along the same lines. 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 equation (6) 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.

Additional References

[1] Seber, G.A.F (2008). A Matrix Handbook for Statisticians. Wiley: New Jersey.

B Appendix B - Description of the MCMC algorithm

B.1 Model setup

In what follows we describe the MCMC procedure used to estimate weights and components of the mixture model. Let $\boldsymbol{e}(x) = (1, e_1(x), \ldots, e_{K-1}(x))'$ be the vector of propensity scores obtained from a multinomial regression of G, defined as in Section 3.2, on a set of conditioning variables X. In our empirical application there is $K = 8$.

Assume that the indicator function:

$$D^*|\boldsymbol{d}_S, d_T, \boldsymbol{e}(x) \sim Be\left(p(\boldsymbol{d}_S, d_T, \boldsymbol{e}(x))\right),$$

is distributed as a Bernoulli random variable. Note that, in the setup considered, D^* is a *latent* quantity. We will assume throughout that the mixture components are normally distributed as explained in Section 3, namely:

$$Y_i|\boldsymbol{e}(x) \sim \mathcal{N}\left(\mu_i(\boldsymbol{e}(x)), \sigma_i^2\right), \quad \text{for } i = 0, 1.$$

Note that the propensity score is, by construction, only affecting the mean of the potential outcome distribution, while retaining the assumption of homoscedasticity across individuals. The functions $p(\boldsymbol{d}_S, d_T, \boldsymbol{e}(x))$ and $\mu_j(\boldsymbol{e}(x))$ we select are as follows:

$$\begin{aligned} p(\boldsymbol{d}_S, d_T, \boldsymbol{e}(x)) &= p_g(\boldsymbol{e}(x)) = \Phi(\boldsymbol{\gamma}_{\boldsymbol{g}}'\boldsymbol{e}(x)), \\ \mu_i(\boldsymbol{e}(x)) &= \boldsymbol{\theta}_{\boldsymbol{i}}'\boldsymbol{e}(x), \qquad i = 0, 1, \end{aligned}$$

where $\boldsymbol{\gamma_g}$ and $\boldsymbol{\theta_j}$ are K-dimensional parameter vectors and $\Phi(\cdot)$ is the standard normal cumulative density function. Note that the subscript g, defining the combination of $\boldsymbol{d}_S$ and d_T considered, is introduced in the definition of $p_g(\boldsymbol{e}(x))$ to simplify notation. This setup defines the following vector of parameters: $\boldsymbol{\xi} = \left\{\boldsymbol{\theta_0}, \sigma_0^2, \boldsymbol{\theta_1}, \sigma_1^2, \boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_K}\right\}$.

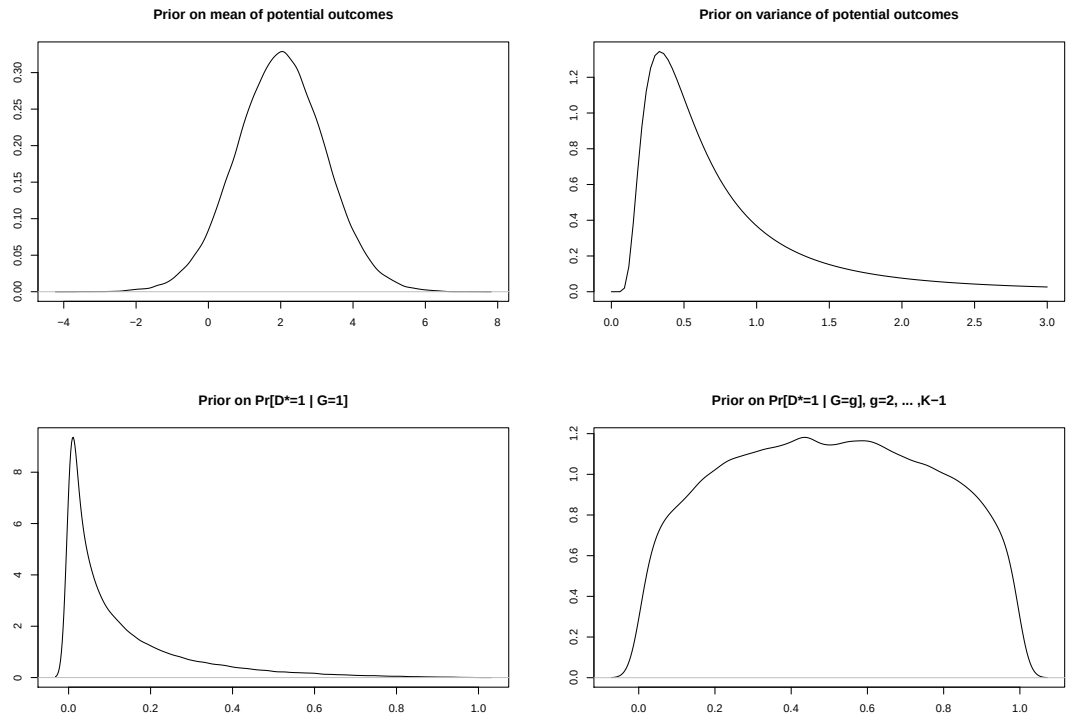
B.2 MCMC algorithm

The goal of the MCMC algorithm is to approximate the posterior distribution of $\boldsymbol{\xi}$ given the data. For the case at hand, given a starting point $\boldsymbol{\xi}^{(0)}$ we will generate a Markov chain whose invariant distribution is the posterior. The large number of parameters makes it convenient to update the Markov chain one component at a time. The Gibbs sampler implements this idea by drawing each component (conditional on the others) from the corresponding full conditional distribution.

B.3 Prior distributions

To ease computation and to obtain closed form solutions for the *full conditional distributions*, we considered the following priors for the parameters in $\boldsymbol{\xi}$.

FIGURE 2. Marginal prior distributions



- **Means of mixture components**. For $i = 0, 1$ we set $\boldsymbol{\theta_i} \sim N_K(\boldsymbol{\psi}, V)$, with $\boldsymbol{\psi} = (2, 0, 0, 0, 0, 0, 0, 0)$ and $V = I$, with I being the identity matrix. Such choice is made so that the resulting marginal prior distribution for the mean of the potential outcomes[21], is centered around the mean of the observed outcome Y. The variance of such prior distribution is also chosen to be sufficiently large (basically spanning the observed range of Y) so that we are not imposing any strong prior knowledge on the value of $\mu_i(\boldsymbol{e}(x))$, $i = 0, 1$. The top left panel of Figure 2 reports the shape of such prior density.

- **Variances of mixture components**. For $i = 0, 1$ we set $\sigma_i^2 \sim IG(\alpha, \beta)$, that is an inverse gamma distribution with density:[22]

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{\beta}{x}}.$$

[21] This is defined as:

$$\int \mu_j(\boldsymbol{e}(x)) dx.$$

[22] If the random variable Z has a gamma distribution with parameters α and β:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-\beta x},$$

then Z^{-1} has the inverse gamma distribution with parameters α and $1/\beta$. The density is always finite, its integral is finite if $\alpha > 0$, and is the conjugate prior distribution for the variance parameter of a normal distribution. To simulate from an inverse gamma, one has to draw samples from a gamma variate, namely X, and then compute $1/X$.

The values of the shape α and scale β parameters were chosen to 2 and 1, respectively. The corresponding density function is reported in the top right panel of Figure 2.

- **Index probability**. We set $\boldsymbol{\gamma_g} \sim N_K(\boldsymbol{\zeta_g}, W)$. We select multivariate normal priors following Albert and Chib (1993), so to ease sampling from the full conditional distributions (see below). In the application we set $W = 0.5I$, where I is defined as above. Note that we adopt different priors distribution for each group defined by the combination of $\boldsymbol{d}_S$ and d_T, summarized by $g = 1, \ldots, K$, so to include prior knowledge on the corresponding probabilities $p_g(\boldsymbol{e}(x))$. In particular $\boldsymbol{\zeta_1} = (-1.5, 0, 0, 0, 0, 0, 0, 0, 0)$, so that the prior density on the marginal probability of observing D^* equal to one given that all reported measures are zero, i.e. $f_{D^*|\boldsymbol{D}_S D_T}[1|0, 0, 0]$, is the one plotted in the bottom left panel of Figure 2, which give most of its weight to values around zero. Similarly $\boldsymbol{\zeta_K} = (1.5, 0, 0, 0, 0, 0, 0, 0, 0)$, so that the prior density for the probability of observing D^* equal to one given that all reported measures are one, i.e. $f_{D^*|\boldsymbol{D}_S D_T}[1|1, 1, 1]$, is symmetric around 0.5 with respect to the former, hence giving most of its weights around 1. Finally $\boldsymbol{\zeta_g} = (0, 0, 0, 0, 0, 0, 0, 0)$ for $g = 2, \ldots, K-1$, resulting in a prior on the marginal probability density given by the bottom right panel of Figure 2.

B.4 Full conditional distributions

The choice made on the prior distributions for the parameters involved implies that the full *conditional* distributions can be derived as follows.

B.4.1 Latent state D_i^*

Given the prior distributions outlined above, then for all i:

$$Pr[D^* = 1|\boldsymbol{d}_S, d_T, \boldsymbol{e}(x), \boldsymbol{\xi}] \propto p(\boldsymbol{d}_S, d_T, \boldsymbol{e}(x)) \exp\left\{-\frac{(y - \mu_1(\boldsymbol{e}(x)))^2}{2\sigma_1^2}\right\}, \tag{7}$$

$$Pr[D^* = 0|\boldsymbol{d}_S, d_T, \boldsymbol{e}(x), \boldsymbol{\xi}] \propto (1 - p(\boldsymbol{d}_S, d_T, \boldsymbol{e}(x))) \exp\left\{-\frac{(y - \mu_0(\boldsymbol{e}(x)))^2}{2\sigma_0^2}\right\}, \tag{8}$$

so that one could easily draw values from such conditional distributions.

B.4.2 Index probability $\boldsymbol{\gamma}_g$

For any $g = 1, \ldots, K$ define the latent random variable T as:

$$T|\boldsymbol{\xi}, \boldsymbol{e}(x), g \sim \mathcal{N}(\boldsymbol{\gamma_g'} \boldsymbol{e}(x)) \text{ truncated at left (right) by 0 if } D^* = 1 \ (D^* = 0). \tag{9}$$

The conditional posterior distribution of $\boldsymbol{\gamma_g}$ is then a multivariate normal:

$$\boldsymbol{\gamma_g}|\boldsymbol{\xi}, \boldsymbol{e}(x) \sim \mathcal{N}_K(\tilde{\boldsymbol{\zeta}}_{\boldsymbol{g}}, \tilde{W}),$$

where $\tilde{\boldsymbol{\zeta_g}} = (W^{-1} + E_g' E_g)^{-1}(W^{-1}\boldsymbol{\zeta_g} + E_g T_g)$ and $\tilde{W} = (W^{-1} + E_g' E_g)^{-1}$, with E_g and T_g being the matrices corresponding to $\boldsymbol{e}(x)$ and T, respectively, including only rows for which there is $G = g$.

B.4.3 Conditional means of mixture components $\boldsymbol{\theta}_i$

The conditional posterior for $\boldsymbol{\theta_i}$ is multivariate normal with mean vector:

$$\tilde{\boldsymbol{\psi}}_{\boldsymbol{j}} = (V^{-1} + S_j' S_j)^{-1}(V^{-1}\boldsymbol{\psi} + S_j' y_j),$$

and variance:

$$\tilde{V}_j = (V^{-1} + S_j' S_j)^{-1},$$

where S_j and y_j are the matrix obtained from $\boldsymbol{e}(x)$ and y only including rows for which there is $D^* = j$.

B.4.4 Conditional variances of mixture components σ_j^2

The conditional full posterior distribution for σ_i^2 is inverse gamma with parameters:

$$\begin{aligned} \tilde{\alpha} &= \alpha + n_j/2, \\ \tilde{\beta} &= \beta + 0.5\left(y_j' y_j + \boldsymbol{\psi}' V \boldsymbol{\psi} - \tilde{\boldsymbol{\psi}}_{\boldsymbol{j}}{}' \tilde{V}_j^{-1} \tilde{\boldsymbol{\psi}}_{\boldsymbol{j}}\right), \end{aligned}$$

where n_j is the number of observations for which there is $D^* = j$.

B.4.5 Algorithm

The sampler alternates two main steps. First, it draws from the distribution of the latent indicators D^* given the model parameters $\boldsymbol{\xi}$; then, it draws from the model parameters $\boldsymbol{\xi}$ given the indicators D^*. Convergence to the posterior distribution is obtained after a burn-in period set by a certain number of iterations ($10,000$ in our application). All draws after convergence refer to the posterior distribution and are, by construction, autocorrelated. In our application, the number of random draws was set to $2,000$.

The algorithm consists of the following steps (with t denoting the generic iteration):

- Initialize the chain at $\left(\boldsymbol{\theta}_{\boldsymbol{j}}^{(0)}, \boldsymbol{\sigma}_{\boldsymbol{j}}^{(0)} \boldsymbol{\gamma}_{\boldsymbol{g}}^{(0)}\right)$, for $j = 0, 1$ and $g = 1, \ldots, K$.
- `for t= 1, ..., iLoop`
  - Simulate $D^{*(t)} | \boldsymbol{\theta}_{\boldsymbol{i}}^{(t-1)}, \boldsymbol{\sigma}_{\boldsymbol{i}}^{(t-1} \boldsymbol{\gamma}_g^{(t-1)}$ from a Bernoulli random variable as in equations (7) and (8).
  - Simulate $T_i^{(t)} | \boldsymbol{D^*}^{(t)}, \boldsymbol{\gamma}_g^{(t-1)}$, $i = 1, \ldots, N$, from a truncated Normal distribution according to (9).

– Simulate $\boldsymbol{\gamma}_g^{(t)}|\boldsymbol{T}^{(t)}, \boldsymbol{D}^{*(t)}$ from a multivariate Normal distribution with mean $\tilde{\boldsymbol{\zeta}}_g$ and variance-covariance matrix $\tilde{D}$, for $g = 1, \ldots, K$.

– Simulate $\boldsymbol{\theta}_j^{(t)}|\boldsymbol{D}^{*(t)}, \boldsymbol{\theta}_j^{(t-1)}, \boldsymbol{\sigma}_j^{(t-1)}$ from a multivariate Normal distribution with mean $\tilde{\boldsymbol{\psi}}_j$ and variance-covariance matrix $\tilde{V}_j$, $j = 0, 1$;

– Permutation sampling step:[23]

∗ if $\left(\sum_{i=1}^{N} \boldsymbol{\theta}_0' \boldsymbol{e}(x) < \sum_{i=1}^{N} \boldsymbol{\theta}_1' \boldsymbol{e}(x)\right)$ then $(\boldsymbol{\theta}_0^{(t)}, \sigma_0^{2\ (t)})$ and $(\boldsymbol{\theta}_1^{(t)}, \sigma_1^{2\ (t)})$ are interchanged;

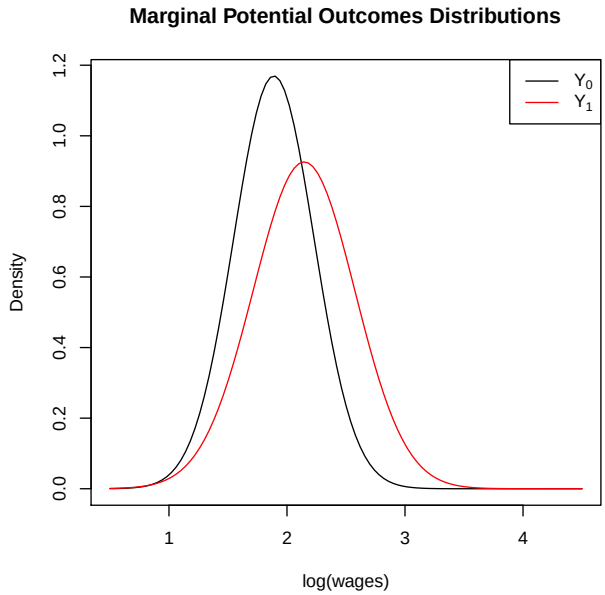
– Simulate $\sigma_j^{2\ (t)}|\boldsymbol{D}^{*(t)}, \boldsymbol{\theta}_j^{(t)}, \boldsymbol{\sigma}_j^{(t-1)}$ from an inverse gamma random variable with parameters $\tilde{\alpha}_j$ and $\tilde{\beta}_j$, $j = 0, 1$;

– t=t+1.

- `End for`

In Figure 3 we report the mixture components as they result from the algorithm. Reported in Figure 4 are the posterior distributions of the variance parameter of potential outcomes. The evidence in the figures is suggestive of heterogenous returns across individuals. Finally in Figure 5 we graph the marginal posterior distribution of the ATT as obtained by the algorithm using the full set of controls.

FIGURE 3. Estimated distributions of potential outcomes



[23]The mixture distribution is unchanged if the group labels are permuted, and thus the parameter space should be defined to clear up any ambiguity. We did so by imposing that the marginal means of the mixture components are in non-decreasing order (see Frühwirth-Schnatter, 2001).

FIGURE 4. Posterior distributions of the variance parameter of potential outcomes

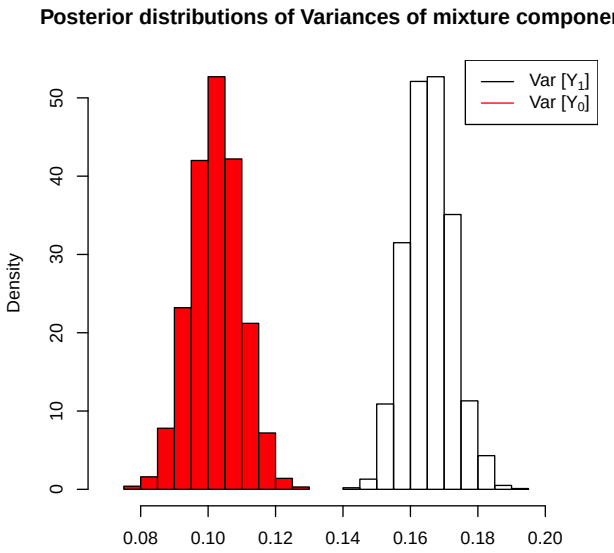
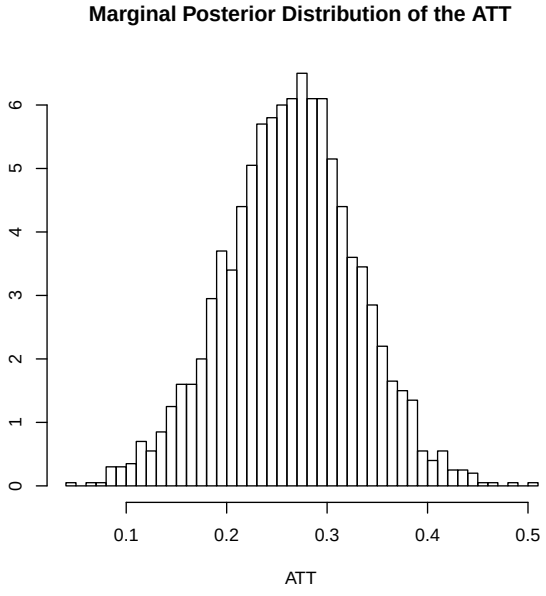


FIGURE 5. Marginal posterior distribution of the ATT



Additional References

[1] Albert, J.H., Chib, S. (1993), *Bayesian Analysis of Binary and Polychotomous Response Data*, Journal of the American Statistical Association, 88, 422, 669-679.

[2] Frühwirth-Schnatter, S. (2001) *Markov chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models*, Journal of the American Statistical Association. 96, 453, 194-209.

C Appendix C - Miscellanea

TABLE 6. Sample selection

	sample size
NCDS birth cohort	17,000
Non-missing education	
1978 Exam Files	14,331
1981 Survey	12,537
1991 Survey	11,407
None missing	8,504
Males with non-missing wage in 1991	3,639
Non-missing wage in 1991 and education ever	2,716

Note. Reported is the sample size after each selection criterium is applied starting from raw data from the British National Child Development Survey (see Section 4 for more details). The last row reports the sample size of the working dataset. The selection criteria adopted are those in Blundell, Dearden and Sianesi (2005) and Battistin and Sianesi (2011).

FIGURE 6. Exact classification probabilities in the 1991 Survey given for different values reported in the 1981 Survey