

Misclassified Treatment Status and Treatment Effects: An Application to Returns to Education in the UK*

Erich Battistin

University of Padova and Institute for Fiscal Studies

Barbara Sianesi

Institute for Fiscal Studies

27th September 2007

*First draft September 2004. This paper benefited from helpful discussions with Richard Blundell, Andrew Chesher, Andrea Ichino, Guido Imbens, Costas Meghir, Francesca Molinari, Peter Mueser, Enrico Rettore, Richard Spady and Yu Xie and comments by audiences at Padova (July 2004), “XIX Italian Conference of Labour Economics” (September 2004), “Cemmap Metrics Lunch Seminar” (November 2004), Bank of Italy (April 2005), Second World Conference SOLE/EALE (June 2005), “The Empirical Evaluation of Labour Market Programmes” (Nuremberg, June 2005), ESWC (August 2005), Policy Studies Institute (September 2005), CEMAPRE (Lisbon, November 2005), ADRES Conference on “Econometric Evaluation of Public Policies: Methods and Applications” (December 2005), Franco Modigliani Fellowship Workshop (February 2006), CEE (February 2006), Michigan (March 2006), Kentucky (March 2006), Maryland (March 2006), European University Institute (April 2006) and NBER Education Meeting (April 2006). Financial support from the ESRC under the research grant RES-000-22-1163 and through the ESRC Centre for Microeconomic Analysis of Public Policy at the IFS is gratefully acknowledged. Address for correspondence: Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE - UK and Department of Statistics, Via Cesare Battisti 243-5, 35123 Padova - Italy. E-mail: erich.battistin@unipd.it and barbara.s@ifs.org.uk.

Abstract

In this paper we study the impact of misreported treatment status on the estimation of causal treatment effects. We characterise the bias introduced by misclassification on the average treatment effect on the treated under the conditional independence assumption. Although the bias of matching-type estimators computed from misclassified data cannot in general be signed, we show that for the case of binary treatments the bias is most likely to be downward if misclassification does not depend on variables entering the selection-on-observables assumption, or only depends on such variables via the propensity score index. We provide results for the case of multiple treatments, showing that the bias arising from misclassification is still relatively easy to characterise but yet far more difficult to be signed even for relatively simple error processes. The empirical problem that motivates our paper is the estimation of the wage returns to a number of educational qualifications for the UK allowing for misreporting in the level of schooling. We provide results to bound the returns to broad educational qualifications semi-parametrically, and by using the unique nature of our data we assess the plausibility for the two biases from measurement error and from omitted variables to cancel out.

JEL Codes: C10, I20, J31.

Keywords: Bounds, Measurement Error, Misclassification, Programme Evaluation, Returns to Educational Qualifications, Treatment Effect.

1 Introduction

Countless theoretical and applied work has addressed itself to the evaluation problem, that is to the measurement of the causal impact of a generic ‘treatment’ on one or more outcomes of interest (for a review see Heckman and Robb, 1985, and Heckman, LaLonde and Smith, 1999).

While endogeneity of treatment status has been the main preoccupation of both theoretical and empirical research, measurement error in recorded treatment status and its consequences for the estimation of causal effects has received far less attention. An exception has been the literature on unions, which has traditionally been sensitive to the issue of misreporting of union status (see, for example, Card, 1996). Recently, the identification problem introduced by measurement error within a causal framework has been reconsidered in a number of studies, including Molinari (2007), Lewbel (2007), Mahajan (2006) and Battistin and Chesher (2007).

In empirical applications the possibility of misrecorded treatment status is indeed far from negligible. Examples include: the returns to work-related training, where the occurrence of training is typically self-reported by individuals who are asked to recall whether they have undertaken any course for work purposes; the effects of programmes (or policy schemes) in which participation (or eligibility) has to be obtained from survey respondents, who have often been shown to have rather poor recall or awareness of the kind of schemes they are in (see e.g. Bound, Brown and Mathiowetz, 2001); the effects of government schemes where the researcher cannot directly observe or measure actual take-up and has to ‘impute’ the treatment status (e.g. eligibility to some means-tested benefits); or a randomized study where the extent of actual compliance (in terms of participants failing to take the treatment and/or controls taking an alternative one) is not recorded in the data. A more general class of example applications comprises those situations in which the treatment status is derived by splitting the sample based on an underlying continuous variable which is itself potentially measured with error (e.g. income or consumption to define poverty status, or firm size to define some form of eligibility).

In this paper we study the impact of misreported treatment status on the estimation of causal treatment effects under the conditional independence assumption.¹ In particular, we characterize the bias introduced by misclassification on the average treatment effect on the treated (ATT). To the best of our knowledge, all papers dealing with this issue only consider

¹This assumption is often also referred to as selection on observables, or unconfoundedness. See Imbens (2004) for a review of semiparametric estimation of treatment effects under this assumption.

the conditional treatment effect (i.e. for a given value of the observable characteristics – see Lewbel, 2007). Especially in applied work, however, interest mostly lies in recovering the average effect of the treatment on the sub-population of participants - ATT in what follows. To retrieve this parameter, the conditional effect needs to be integrated over the (unobserved) distribution of characteristics in the truly treated group. We show that knowledge of the misclassification probabilities is enough to back out the true ATT from the raw data. Furthermore, although the resulting bias of matching-type estimators for the ATT computed from misclassified data cannot in general be signed, we show that the bias is most likely to be downward if misclassification does not depend on the variables entering the conditional independence assumption, or if misclassification only depends on such variables via the propensity score index.

We further extend our approach to a multiple-treatment framework, which becomes necessary if interest lies in estimating the incremental impacts of multiple treatments and in fact the impacts of binary but more narrowly-defined treatments that do not split up the entire population. In either of such cases, account needs to be taken of the potential misclassification in the reporting of all treatment levels, not just in the two ones being considered. Furthermore, as we argue in more detail in Section 6, the move to a multiple treatment framework is often necessary just to be able to justify the non-differential misclassification assumption widely invoked in the literature (see Bound, Brown and Mathiowetz, 2001).

When an instrument-like variable, or an additional, independent measure of the treatment becomes available, point identification of the misclassification probabilities can be achieved (see Mahajan, 2006, Lewbel, 2007, and Hu, 2007). In our companion paper (Battistin and Sianesi, 2006b) we exploit the characterization of the bias developed in this paper together with repeated measurements of individual qualifications to arrive at point estimates of the ATT that allow for measurement error.

This paper focuses on a bounding approach and corresponding sensitivity analysis that can provide an often quite informative robustness check when neither multiple measures nor instruments for the treatment status are available. Specifically, the characterization of the bias we provide straightforwardly leads to partial identification results that do not rely on additional information. The resulting bounds for the ATT(s) can be tightened by making a priori assumptions on the extent of misclassification. Such restrictions on the nature of reporting errors can be obtained by running small-scale validation studies, or by looking at results from previous

research or behavioural theories that seem reasonable for the phenomenon under investigation; alternatively, bounds can be calculated for a range of plausible values of reporting errors to assess the robustness of the evaluation inference to the presence of misclassification.

Furthermore, we suggest to exploit the propensity score calculated from raw data (that is, the probability of receiving treatment conditional on individual characteristics – see Rosenbaum and Rubin, 1983) to derive bounds for the ATT in a non/semi-parametric way. As far as we are aware, we are the first ones to use the propensity score to relax several parametric assumptions typically made in this literature. First, we allow for arbitrarily heterogeneous individual treatment effects. Second, we allow for arbitrary non-linearities in the no-treatment outcome equation. Finally, we allow misreporting probabilities to depend on individual characteristics, although we restrict such variation to be such that individuals with the same probability of reporting receipt of treatment have the same probabilities of misreporting it. By contrast, Black, Berger and Scott (2000) ignore the presence of individual characteristics; Kane, Rouse and Staiger (1999) assume linearity, impact homogeneity and constant misclassification; while Lewbel (2006) imposes some parametric structure to ease estimation.

The causal effects that motivated this paper are the wage returns to educational qualifications in the UK. While the estimation of the return to education is amongst the most explored and prolific areas in labour economics and attracts constant policy interest (for a discussion, see Blundell, Dearden and Sianesi, 2004), there is a real possibility of errors in education data: in addition to data transcript errors, survey respondents may over-report their attainment, not remember, or just not know if the schooling they have had counts as a qualification. As to the latter, the British education system is remarkably complex, with a plethora of different, often changing subqualifications classified in broader levels, often based on obtained grades.

The received wisdom from the studies on the returns to years of education has traditionally been that the upward bias from omitted ‘ability’ variables and the downward bias from (classical) measurement error largely cancel each other out (for a review see Griliches, 1977, and Card, 1999; for a recent UK study see Bonjour *et al.*, 2003). Such a continuous years-of-schooling measure, although particularly convenient, imposes the restriction that the returns increase linearly with each additional year, irrespective of the level and type of educational qualifications the years refer to. In the UK and other European countries, however, there are alternative nationally-based routes leading to quite different educational outcomes, and the

importance of distinguishing between different types of qualifications and allowing each to have a separate effect on earnings is widely accepted (see Blundell, Dearden and Sianesi, 2005). With a categorical qualification-based measure of education, however, as mentioned above the assumption of classical measurement error cannot hold, as individuals in the lowest category can never under-report their education level and individuals in the top category cannot over-report. Given that OLS estimates are not necessarily downward biased, the cancelling out of the ability and measurement error biases cannot be expected to hold in general. Moreover, the IV methodology cannot provide consistent estimates of the returns to qualifications.

To date, empirical evidence on the importance of these issues for the estimation of returns to education is restricted to the US, where it was in fact shown that measurement error might play a non-negligible role, as we briefly review in the next section. For the UK there are no estimates of the returns to educational qualifications that adequately correct for measurement error. 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, 1999b, Dearden et al., 2002, and McIntosh, 2004).

In our empirical application we demonstrate the usefulness of our proposed sensitivity analysis by applying it to bound the returns to a number of educational qualifications in the UK semi-parametrically, both in a binary and multiple-treatment setting. To motivate the conditional independence assumption we rely on Blundell, Dearden and Sianesi (2005) who could not find any strong evidence of remaining selection bias given the information available in the National Child Development Survey data. Further, by exploiting this uniquely rich data we assess the plausibility that the biases from measurement error and from omitted variables cancel out in the estimation of returns. If this were the case, to estimate up-to-date returns to qualifications policy-makers could simply rely on Labour Force Survey-type datasets, which totally rely on recall about individuals' and do not contain any information on individual ability and family background.

To preview our results, we show that the resulting bounds are sometimes wide but generally point to reasonable ranges of positive values for average returns to schooling among the schooled. For the range of educational qualifications considered, we show that the claim sometimes made that measurement error roughly cancels out selection bias is not supported. As long as the

extent of misreporting is not too severe, this result appears to be rather robust across the educational categories considered and if we allow the propensity to misreport to depend on individual characteristics. We find that the sign of the bias introduced by misreporting depends on the educational qualification considered and is more difficult to pin down in a multiple-treatment setting, thus preventing us from making a general statement about the relationship between the returns estimated from true and error ridden data. However, in the majority of cases that we deal with in this paper we observe that returns estimated from raw data are lower than the range of values for the causal effect of interest implied by our bounding approach.

The remainder of the paper is organized as follows. In Section 2 we start by reviewing the evidence on measurement error and returns to educational qualifications. Section 3 sets out the general evaluation framework, while Section 4 introduces the possibility of misclassification in the treatment status. In Section 5 we show the consequences that such reporting errors might have for the estimation of causal treatment effects. In Section 6 we discuss how to extend our identification strategy to deal with misreporting of categorical treatments. Section 7 briefly describes the data we use and defines our parameters of interest. In Section 8 we first sketch our strategy for partial identification of causal effects in the presence of misclassification, before presenting and discussing our results. Section 9 concludes. For all proofs except the one of Proposition 2, we refer to Battistin and Sianesi (2006a).

2 Returns and misreporting: the evidence so far

Whilst use of years of completed education has a long history in the US, for the UK most authors prefer qualification-based measures of educational attainment. Recent examples include Dearden (1999a), Blundell *et al.* (2000), Gosling, Machin and Meghir (2000) and Blundell, Dearden and Sianesi (2005).²

However despite the importance of schooling both as an outcome and as an explanatory variable, hardly any effort has been devoted to assessing either the accuracy of widely used survey reports of educational attainment in the UK, or the impact that misreporting might have on estimated returns to education.³ To date, the only work in the latter direction is

²For a review and summary of some recent work on returns to qualifications, see Sianesi (2003).

³Ives (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.

Dearden (1999b) and Dearden *et al.* (2000 and 2002), who however ignore the non-classical nature of measurement error caused by misreporting of discrete qualifications and conclude that measurement error bias and omitted ability bias largely cancel out in the estimation of returns. Indeed, some recent work based on the UK Labour Force Survey (e.g. McIntosh, 2004) at times appeals to this result.

As a benchmark it is thus worth considering the evidence on categorical education measures available for the US, most of which being provided by the study by Kane, Rouse and Staiger (1999) (see also the work referred to by Card, 1999). Overall, misreporting was found to be more likely to happen for low levels of qualification, with over-reporting being more likely than under-reporting (see also Black, Sanders and Taylor, 2003) and events such as degree completion being more accurately reported than completed years of college. Interestingly, transcript measures were often found to be subject to at least as much – and at times even more! – measurement error as self-reported survey measures. Estimates of returns that ignore such misclassification were found to be severely biased, either upwards or downwards depending on the educational level of interest. Similarly, the application in Lewbel (2006) points to seriously inaccurate transcript information as to degree attainment and finds that allowing for misclassification leads to a 5-fold increase in the estimated return to college.

3 Identification in the absence of misclassification

The specific evaluation problem we have in mind and to which we shall refer throughout is the identification and estimation of the causal effects of educational qualifications on individual (log) wages.

Consider the binary treatment case within the potential outcome framework⁴, where treatment status is defined by either having achieved the educational level of interest ($D^* = 1$) or not ($D^* = 0$). Examples include completing college compared to not doing so, or attaining any qualification compared to dropping out of high-school with none. The generalization to the multiple-treatment case, though notationally more demanding, proceeds along the same lines and will be considered in Section 6.

The individual causal effect (or return) of achieving the qualification is defined as the dif-

⁴For reviews of the evaluation problem see Heckman, LaLonde and Smith (1999) and Imbens (2004). For the potential outcome framework, see in particular Roy (1951), Quandt (1972) and Rubin (1974).

ference between two potential outcomes: the wage if the individual were to achieve the qualification, Y_1 , and the wage if the individual were not to achieve it, Y_0 . This set-up is extremely general, in particular it does not assume that the returns are homogeneous across individuals.⁵ Note that the observed individual wage can be written as $Y = Y_0 + D^*(Y_1 - Y_0)$.

In this paper we focus on the average return to a qualification for those who have obtained it, i.e. the average effect of treatment on the treated (ATT in the following):

$$\Delta^* \equiv E(Y_1 - Y_0 | D^* = 1) = E(Y_1 | D^* = 1) - E(Y_0 | D^* = 1). \quad (1)$$

This is the relevant parameter when the treatment is voluntary, which is the case for the achievement of (post-compulsory) educational qualifications, and is also the one needed for a cost-benefit analysis. In fact, it is the one which has traditionally received most attention in the evaluation literature.

To identify the ATT, the average unobserved counterfactual $E(Y_0 | D^* = 1)$ needs to be somehow constructed on the basis of some untestable identifying assumptions. As we aim to characterize the impact of misreporting of D^* , in what follows we assume that the educational choice D^* is otherwise exogenous conditional on a set of observable variables X :

Assumption 1 (*Mean Independence Assumption*): $E(Y_0 | D^*, X) = E(Y_0 | X)$.

This assumption requires the evaluator to observe all those characteristics that jointly affect the decision to acquire the qualification of interest and potential wages in the absence of that educational investment. Its plausibility for our empirical application, and in particular the issue of ‘ability bias’, will be addressed in the data section.

To give empirical content to Assumption 1, we also require that for all values X in the population of those with $D^* = 1$ there are also non-participants in the qualification of interest:

Assumption 2 (*Common Support*): $e^*(x) \equiv Pr(D^* = 1 | X = x) < 1 \quad \forall x$, where $e^*(x)$ is the propensity score.

Under Assumptions 1 and 2, the ATT is identified as:

$$\Delta^* = \int \Delta^*(x) f(x | D^* = 1) dx, \quad \text{where} \quad (2)$$

⁵For this representation to be meaningful, the stable unit-treatment value assumption needs however to be satisfied (Rubin, 1980), requiring that an individual’s potential wages as well as the chosen education level are independent from the schooling choices of other individuals in the population.

$$\Delta^*(x) \equiv E(Y_1 - Y_0|x) = E(Y|D^* = 1, x) - E(Y|D^* = 0, x)$$

is the conditional treatment effect, that is the average treatment effect (or average return) for individuals with characteristics $X = x$.

4 Misclassified treatment status

4.1 General formulation of the problem

This section extends the potential outcome framework to study the consequences for the identification of causal effects of allowing for misclassified treatment status. Specifically, either because individuals are left to self-report their qualifications or because of transcript errors, the treatment status $D \in \{0, 1\}$ recorded in the data may differ from the actual status D^* .

In the absence of measurement error, data are informative about (Y, D^*, X) ; as seen above, estimators based on Assumptions 1 and 2 establish a correspondence between this triple and the parameter of interest in (1). By contrast when qualifications are misreported, data are informative about the distribution of measurement-error contaminated variables. If measurement error is ignored, or not perceived, causal effects will thus be inferred using realizations of (Y, D, X) as if they were realizations of (Y, D^*, X) , and estimators of (2) will therefore be constructed as:

$$\int_{\mathcal{S}} \Delta(x) f(x|D = 1) dx \equiv \Delta, \quad \text{where} \quad (3)$$

$$\Delta(x) \equiv E(Y|D = 1, x) - E(Y|D = 0, x),$$

$$\mathcal{S} \equiv \{x : e(x) \equiv Pr(D = 1|X = x) < 1\}.$$

Hence \mathcal{S} is the observed common support for the self-reported participants in education and $e(x)$ is the propensity score calculated from the mismeasured qualification D . It is worth noting that estimators of the ATT based on $e(x)$ (e.g. estimators based on propensity score matching or re-weighting) are equivalent to the estimator defined by the empirical analogue of (3), as we have that, with obvious notation:

$$\Delta = \int_{\mathcal{S}} \Delta[e(x)] f[e(x)|D = 1] de.$$

The result straightforwardly follows from x being finer than $e(x)$ and by the balancing property of the propensity score (see Rosenbaum and Rubin, 1983). Showing that the effect estimated

from raw data (3) differs from the true effect is thus sufficient to conclude that any estimation procedure exploiting the propensity score calculated from raw data would in general lead to incorrect inference.

Since some individuals with $D^* = 0$ will erroneously be misclassified as participants on the basis of the error-affected indicator D and only part of those individuals reporting $D = 1$ have actually got the qualification of interest, the estimation of causal effects based on (Y, D, X) will in general be biased for treatment effects, with the magnitude of this bias depending on the extent of misclassification. This is shown in Section 5.3, where we derive the difference between Δ^* , the true causal parameter of interest, and Δ , the parameter that would instead be identified from the observable triple (Y, D, X) .

4.2 The misclassification probabilities

In what follows we build on Molinari (2007) to introduce the required notation and an assumption on the classification errors that will be maintained throughout. Define the (mis)classification probabilities as

$$\lambda_{ij}(x) \equiv Pr(D^* = i | D = j, x), \quad i, j \in \{0, 1\},$$

which may in general depend on (part of) X .⁶

In the binary case, there are two types of misclassification: $\lambda_{10}(x)$, the proportion of true participants amongst those reporting $D = 0$; and $\lambda_{01}(x)$, the proportion of true non-participants amongst those with $D = 1$. Of recurrent use will be the probabilities of exact classification:

$$\lambda_{00}(x) \equiv \lambda_0(x) = Pr(D^* = 0 | D = 0, x),$$

$$\lambda_{11}(x) \equiv \lambda_1(x) = Pr(D^* = 1 | D = 1, x),$$

where for ease of notation only one subscript is retained.⁷ It is convenient to collect the (mis)classification probabilities into the matrix of (mis)classification probabilities:

$$\Pi(x) = \begin{bmatrix} \lambda_0(x) & 1 - \lambda_0(x) \\ 1 - \lambda_1(x) & \lambda_1(x) \end{bmatrix}.$$

⁶For example, in the general setting treatment probabilities may depend in large part upon the implementation details of a particular social program while misreporting patterns could be directly related to individual characteristics, that may not necessarily play a role in the decision or in the eligibility to participate in that program.

⁷Note that the (mis)classification probabilities have thus been defined conditional on what individuals report, which we find more easily interpretable. These probabilities can also be defined conditional on the true (and unobserved) treatment status: $\gamma_1 = Pr(D = 1 | D^* = 1)$ and $\gamma_0 = Pr(D = 0 | D^* = 0)$. These γ 's are linked to our λ 's via Bayes' Theorem.

Throughout our discussion, we will assume that the classification error is non-differential, as this can help us write down relatively detailed but still manageable models (see Bound, Brown and Mathiowetz, 2001). Accordingly, we will maintain the assumption that conditional on a person's actual qualification and on the covariates entering the conditional independence assumption (1), any variable D which proxies D^* does not contain information to predict the outcome of interest:

Assumption 3 (*Non-Differential Misclassification*): $E(Y|D^*, D, X) = E(Y|D^*, X)$.

For the binary treatment case, the assumption that reporting errors are independent of earnings conditional on D^* and X amounts to requiring that:

$$\begin{aligned} E(Y_0|D^* = 0, D = 1, X) &= E(Y_0|D^* = 0, D = 0, X) \quad \text{and} \\ E(Y_1|D^* = 1, D = 1, X) &= E(Y_1|D^* = 1, D = 0, X). \end{aligned}$$

These two conditions highlight how Assumption 3 would not hold if an individual's propensity to misreport treatment status is related to outcomes. For example, it would be violated if those graduates ($D^* = 1$) who experience a very low Y_1 - either because they have received a negative productivity shock to their no-education earnings and/or because they have reaped a very low if not negative return from their degree - are more inclined to deny possessing the qualification. In addition to such type of behaviour by respondents, there is a more technical consideration that would in fact guarantee a violation of this assumption. If in defining the treatment indicator one were to ignore a feature of the treatment that affects both its effect and recall precision, Assumption 3 would by construction break down. The obvious solution to this is to refine the treatment to fully reflect the feature causing the violation, thus extending the framework to look at the treatment components separately. We further elaborate on this issue in Section 6, where we turn to multiple treatments.

Under Assumption 3, individuals for whom we observe $D = d$ are a mixture of participants ($D^* = 1$) and non participants ($D^* = 0$), with mixing weights given by the (mis)classification probabilities. This result can be written compactly in matrix algebra notation as

$$\begin{bmatrix} E(Y|D = 0, x) \\ E(Y|D = 1, x) \end{bmatrix} = \Pi(x) \begin{bmatrix} E(Y|D^* = 0, x) \\ E(Y|D^* = 1, x) \end{bmatrix},$$

from which we have that

$$\Pi^{-1}(x) \begin{bmatrix} E(Y|D = 0, x) \\ E(Y|D = 1, x) \end{bmatrix} = \begin{bmatrix} E(Y|D^* = 0, x) \\ E(Y|D^* = 1, x) \end{bmatrix}, \quad (4)$$

provided that $\det[\Pi(x)] = \lambda_0(x) + \lambda_1(x) - 1 \neq 0$. We formalize this condition requiring misclassification to be such that:

Assumption 4 (*Informative Recorded Treatment Status*): $\lambda_1(x) + \lambda_0(x) \neq 1 \forall x$.

Assumption 4 appears reasonable. It requires that conditional on X , the proportion of true graduates among those who self-report having a degree be different from the proportion of true graduates among those who self-report not having a degree; or in other words, that the marginal effect of recorded status D on true status D^* conditional on X is non-zero: $Pr(D^* = 1|D = 1, X) \neq Pr(D^* = 1|D = 0, X)$, the latter expression implying some dependence between the true latent variable D^* and its surrogate D .⁸ Note that $\lambda_1(x) + \lambda_0(x) > 1$ represents the most likely case, as it is implied by the assumption that observations on D are more accurate than pure guesses once X is corrected for, i.e. $\lambda_1(x) > 0.5$ and $\lambda_0(x) > 0.5$.

5 The bias introduced by misclassification

5.1 Bias on the conditional treatment effect

In deriving how the parameter that can be recovered from the observed data (3) compares to the causal parameter of interest (2), we start by considering the nature of the inconsistency for estimating the causal treatment effect conditional on X . This bias can be straightforwardly characterized using (4). The result in (5) coincides with the result in Lewbel (2007), and more in general follows from Aigner (1973).

Proposition 1 (*Bias on the Conditional Treatment Effect*) *If Assumptions 1 to 4 are satisfied, it follows that*

$$\Delta^*(x) = \frac{\Delta(x)}{\lambda_0(x) + \lambda_1(x) - 1}. \quad (5)$$

⁸Assumption 4 only requires inequality; it is however convenient to spell out here the two possible cases:

$$\begin{aligned} 4\text{-}(a) \quad \lambda_1(x) + \lambda_0(x) > 1 &\Leftrightarrow Pr(D^* = 1|D = 1, X) > Pr(D^* = 1|D = 0, X), \\ 4\text{-}(b) \quad \lambda_1(x) + \lambda_0(x) < 1 &\Leftrightarrow Pr(D^* = 1|D = 1, X) < Pr(D^* = 1|D = 0, X). \end{aligned}$$

Case 4-(b) represents a case of such extensive misclassification for it to be more likely to randomly draw a true graduate from the group reporting no degree than from the group reporting a degree. By contrast, case 4-(a) is a situation of limited misclassification in the sense that, given X , the proportion of true graduates among those reporting to have a degree is higher than the proportion of true graduates among those reporting not to have a degree.

Accordingly, the estimates of $\Delta^*(x)$ based on the triple (Y, D, X) are always biased towards zero, but possibly with the opposite sign if the measurement error is very strong (the denominator being negative in the case of extreme misclassification). In terms of the conditional treatment effect, therefore, an attenuation bias result still holds. An interesting implication of (5) is that $\Delta(x) = 0 \Leftrightarrow \Delta^*(x) = 0$, so that the raw difference in observed outcomes given X being zero actually implies that the true conditional treatment effect is zero. Finally, if there is no misclassification (that is, $\lambda_0(x) = \lambda_1(x) = 1$), then of course $\Delta(x) = \Delta^*(x)$; and if there is complete reversal in the classification (that is, $\lambda_0(x) = \lambda_1(x) = 0$), then $\Delta(x) = -\Delta^*(x)$.

5.2 Support condition

We have thus far defined the bias for the treatment effect conditional on a given value of the vector X . In order to characterize the bias for the ATT, we have to integrate over the distribution of X in the treated group, which brings us to discuss support issues. Although Assumption 2 implies that the true score $e^*(x)$ is strictly below one, misclassification can cause the observed score $e(x)$ to take on values at the boundary. If this were the case, true participants not belonging to the observed \mathcal{S} would be discarded so that the ATT estimated from (Y, D, X) would refer to a different population of participants than the population of participants the true ATT refers to.

To see this, by using the law of iterated expectations to write $e^*(x)$ in terms of $e(x)$ and solving for $e(x)$, there is:

$$e(x) = \frac{e^*(x) - [1 - \lambda_0(x)]}{\lambda_0(x) + \lambda_1(x) - 1}, \quad (6)$$

from which we see that $e(x)$ will take on values at the boundary according to:

$$e(x) = 1 \Leftrightarrow \lambda_1(x) = e^*(x).$$

It follows that the parameter (3) estimated from the triple (Y, D, X) could in general refer to a different population than the one implied by (Y, D^*, X) . To avoid this, we ensure that $e(x)$ is strictly below one by assuming that misclassification is such that the following condition holds for all values of X :

Assumption 5 (*Restriction on the extent of misclassification*): $\lambda_1(x) > e^*(x)$.

This assumption allows us to treat the common support in the presence of measurement error as the true common support and simply implies that all true participants are used in the

estimation of the ATT from raw data. If this were not the case, the quantities in the following would be defined over a different subset of the truly treated.

5.3 Bias on the treatment effect on the treated

If one were interested in the average treatment effect (ATE), that is the average return for an individual irrespective of whether the qualification of interest has been acquired or not:

$$E(Y_1 - Y_0) = \int \Delta^*(x)f(x)dx,$$

the discussion could stop here.⁹ In particular, one would only need to integrate the conditional average treatment effect $\Delta^*(x)$ over the distribution of X in the population, the latter being observed in the data. Note also that the attenuation-bias result from Proposition 1 would keep holding unconditional on X , so that ignoring measurement error in treatment status would lead to a downward-biased estimate of the ATE. The correspondence between a zero raw average effect and a zero true average effect, however, no longer holds, unless the misclassification probabilities are assumed not to depend on X .

By contrast, if interest lies in recovering the average return to education for those who invested in that qualification (ATT), the conditional effect $\Delta^*(x)$ needs to be integrated over the distribution of X in the (truly) treated group, $f(x|D^* = 1)$, which is not observed.¹⁰ The following proposition provides a characterization of the bias introduced by measurement error for the estimation of (1), that is the relationship between Δ^* and Δ . The proof is reported in the Appendix.

Proposition 2 (*Bias on the Treatment Effect for the Treated*) *If Assumptions 1 to 5 are satisfied, the relationship between the true ATT and the effect estimated from raw data can be written as follows*

$$\begin{aligned} \Delta^* &= \int \omega(x)\Delta(x)f(x|D = 1)dx, \\ &= \Delta + \int [\omega(x) - 1]\Delta(x)f(x|D = 1)dx, \quad \text{where} \\ \omega(x) &= \frac{Pr(D = 1)}{Pr(D^* = 1)} \left[1 + \frac{1}{e(x)} \frac{1 - \lambda_0(x)}{\lambda_0(x) + \lambda_1(x) - 1} \right] \quad \text{and} \end{aligned} \tag{7}$$

⁹Note that identification of ATE requires a strengthened Assumption 1, implying in particular homogeneous returns (given X) or the absence of selection into education based on unobserved returns.

¹⁰By contrast, this distribution could be directly inferred if information on the true treatment status D^* and X were available from validation data. In this case, it can be easily shown that Δ would underestimate Δ^* .

$$Pr(D^* = 1) = \int [1 - \lambda_0(x)]f(x)dx + \int [\lambda_0(x) + \lambda_1(x) - 1]e(x)f(x)dx.$$

This result shows that if the two λ 's were known, the ATT could be estimated by appropriately re-weighting the conditional differences in outcomes based on recorded treatment data, $\Delta(x)$, with weights defined by $\omega(x)$. Note that, as it should be, $\omega(x) = 1$ for all individuals if there is no measurement error. Moreover, weights cannot be signed in general, implying that Δ^* can be over- or under-estimated depending on the unknown probabilities $\lambda_1(x)$ and $\lambda_0(x)$. A notable exception under the most likely condition $\lambda_0(x) + \lambda_1(x) > 1$ is when the true incidence of treatment in the population, $P(D^* = 1)$, is smaller than the one observed from raw data, $P(D = 1)$. This is a sufficient condition for Δ to provide a downward-biased estimate of Δ^* , as all weights would be larger than one. Although $P(D^* = 1)$ is in general unobserved, one could gauge the relative size of the two probabilities if external validation data (e.g. government statistics on educational attainment in our application) were available.

Furthermore, note that Δ being zero no longer implies the absence of a treatment effect, as was the case when conditioning on X .

5.4 Special cases

In what follows we discuss two sets of restrictions that can be imposed on the misclassification probabilities to sign the bias induced by misreporting.

Special Case 1 (*Only over-reporting of qualifications*): $\lambda_0(x) = 1 \forall x$.

To see why this condition represents a situation where only over-reporting of qualifications can occur, note that it corresponds to $P(D^* = 1|D = 0) = 0$, which rules out that true graduates may be found among those reporting not to have a degree, in other words, ruling out under-reporting. By setting $\lambda_0(x)$ equal to one in (5) we get that the conditional treatment effect is always right-signed for all X , but biased towards zero. Furthermore, it follows from Proposition 2 that

$$\omega(x) = \frac{Pr(D = 1)}{Pr(D^* = 1)} = \frac{\int e(x)f(x|D = 1)dx}{\int \lambda_1(x)e(x)f(x|D = 1)dx} \geq 1$$

for all X , so that the estimated effect Δ is always biased towards zero for Δ^* .¹¹

¹¹The assumption that individuals never under-report qualifications they have obtained can be weakened by assuming that over-reporting is just more likely than under-reporting. This case of monotone misclassification

Special Case 2 (*Misclassification independent of X*): $\lambda_1(x) = \lambda_1$ and $\lambda_0(x) = \lambda_0 \forall x$.

Although the assumption that the percentage of correct classification is independent of respondents' characteristics is clearly only made here for convenience, it could be weakened by assuming constant probabilities within cells defined by X . Alternatively, the same arguments made below would still apply if one allowed misclassification to vary with X but only through the propensity score index, $e(X)$. We can thus allow misreporting to depend on X , provided that individuals who have the same conditional probability of claiming to have the qualification of interest also have the same conditional probabilities of overreporting or of underreporting their attainment.

Using Proposition 2 it follows that

$$\omega(x) = \frac{1 + \frac{1}{e(x)} \frac{1-\lambda_0}{\lambda_0+\lambda_1-1}}{\frac{1-\lambda_0}{Pr(D=1)} + (\lambda_0 + \lambda_1 - 1)}.$$

Under the likely scenario $\lambda_0(x) + \lambda_1(x) > 1$, all the weights are positive and a first-order approximation to $\omega(x)$ around $(\lambda_0 = 1, \lambda_1 = 1)$ yields

$$\omega(x) \simeq 1 + (1 - \lambda_0) + (1 - \lambda_1) \left[\frac{1}{e(x)} - \frac{1 - Pr(D = 1)}{Pr(D = 1)} \right],$$

from which it can be seen that a sufficient (and testable) condition for $\omega(x)$ to be larger than one is that the propensity score at x be smaller than the odds ratio, i.e. $e(x) \leq \frac{Pr(D=1)}{1-Pr(D=1)}$. From a study of $\omega(x)$ as a function of the λ 's, it can be shown that only for values of the parameter $P(D = 1)$ smaller than 0.3 is there the possibility that, depending on the value of $e(x)$, the corresponding weight at x is positive but smaller than one. However we found that even in this case the distribution of weights is skewed towards values (often much) larger than one, so that in most empirical applications the 'raw' estimate is most likely to be a lower bound.¹²

6 Extension to multiple treatments

So far we have considered treatments within the binary treatment framework, and in fact treatments where the specific educational level of interest cuts right through the entire educational

imposes that $\lambda_1(x) < \lambda_0(x)$ for all values X , or, in a more intuitive form, $P(D^* = 0|D = 1) > P(D^* = 1|D = 0)$. Monotone misclassification reflects the idea supported by cognitive studies that when respondents are asked questions about socially and personally sensitive topics, they tend to under-report undesirable behaviours and attitudes, and over-report desirable ones.

¹²More detailed results are available upon request.

spectrum (e.g. any qualification versus none, or degree versus non-degree). This does not of course rule out interest in the incremental returns to sequential multiple treatments, or in the returns to binary treatments for a more narrowly defined educational split, such as the return to college vis-à-vis stopping with high school diploma, or the return to finishing school with some qualifications vis-à-vis nothing. Even for considering such types of binary treatments, the analysis needs to be extended to a multiple-treatment framework, since account needs to be taken of the potential misclassification in the reporting of all educational levels, not just in the two ones being considered.¹³

Moving to a multiple treatment framework not only is often policy-relevant, but in the presence of misclassification it may often become a necessity for justifying the widely invoked non-differential misclassification assumption. As anticipated in Section 4, this assumption would by construction be violated if in defining the treatment indicator one were to lump across features of the treatment that affect both its effect and the precision of its recall.¹⁴

More generally, a need to consider an extended framework arises in any situation where underlying the binary treatment indicator is a dose-response framework. In our application we consider educational categories, which are inherently ordered and sequential, but such a set-up can occur much more generally. For instance, when considering completion of college for those who enrolled, or participation in a programme for the unemployed, the underlying treatment – college or the programme – has itself a duration, which is likely to affect both recall of the event and outcomes. Another example relates to treatments taken more or less recently; recall is likely to depend on how long ago the treatment was received, and the treatment effects themselves might evolve, in particular depreciate, over time.

Assumption 3 would thus appear to be most defensible when the treatment is disaggregated into multiple treatments that fully embody the feature that if ignored would cause the violation (sequential categories, duration, how long ago taken, etc.). Such an extended framework would consider each treatment component/feature as a separate treatment.

With our application in mind, in the following we extend our framework to consider three

¹³For instance, even if one only wanted to compare college to high school, the other categories would still need to be considered, since, first, individuals reporting no qualifications might in reality have a high-school diploma or a college degree, and, second, individuals reporting college or high school diploma might in reality have neither of the two qualifications of interest.

¹⁴To see this, consider the treatment being defined as having any qualification as opposed to none. In such a situation, an individual with a degree will be more likely both to correctly report to have any qualification *and* to have higher earnings than an observationally-equivalent individual who has only completed high school.

levels of qualifications (or more generally, of exposure), which we assume to be of increasing intensity. Let these levels be defined by $D^* = 0$, $D^* = 1$ and $D^* = 2$, denoting, for example, high-school drop-outs, high-school graduates and college graduates, respectively. We are interested in the estimation of pairwise incremental returns, that is the wage return of obtaining a qualification of interest (e.g. college) relative to a lower qualification (e.g. high-school), when the only available measure of educational attainment D is potentially affected by error. We focus on the following three ATT's:

$$\begin{aligned}\Delta_{10}^* &\equiv E(Y_1 - Y_0|D^* = 1), \\ \Delta_{21}^* &\equiv E(Y_2 - Y_1|D^* = 2), \\ \Delta_{20}^* &\equiv E(Y_2 - Y_0|D^* = 2).\end{aligned}$$

The conditional ATT's based on true and observed attainment levels are defined as:

$$\begin{aligned}\Delta_{ij}^*(x) &= E(Y|D^* = i, x) - E(Y|D^* = j, x), \\ \Delta_{ij}(x) &= E(Y|D = i, x) - E(Y|D = j, x),\end{aligned}$$

with $i > j$, $(i, j) \in \{0, 1, 2\}$, and $\Delta_{20}^*(x) = \Delta_{10}^*(x) + \Delta_{21}^*(x)$ and $\Delta_{20}(x) = \Delta_{10}(x) + \Delta_{21}(x)$ (due to the adding-up conditions). Along the lines of what discussed in Section 4, the relationship between true quantities and quantities observed from raw data depends on the 3×3 matrix of misclassification probabilities $\Pi(x)$. If this matrix is invertible, each $\Delta_{ij}^*(x)$ can be written as a function of the $\Delta_{ij}(x)$'s, thus providing an extension of the result in Proposition 1.

The ATT's of interest can then be written as:

$$\begin{aligned}\Delta_{10}^* &= \int \Delta_{10}^*(x) f(x|D^* = 1) dx, \\ \Delta_{21}^* &= \int \Delta_{21}^*(x) f(x|D^* = 2) dx, \\ \Delta_{20}^* &= \int \Delta_{20}^*(x) f(x|D^* = 2) dx = \Delta_{21}^* + \int \Delta_{10}^*(x) f(x|D^* = 2) dx,\end{aligned}$$

which depend on the conditional distributions of X given $D^* = 1$ and $D^* = 2$.

For the above quantities to have a causal interpretation we need suitably extended versions of Assumptions 1 and 2 (see Imbens, 2000, and Lechner, 2001):

Assumption 6 (*Extended Conditional Independence and Common Support*)

$$\begin{aligned}E(Y_j|D^* = i, X) &= E(Y_j|D^* = j, X), \quad (i, j) \in \{0, 1, 2\}, \quad i > j \\ e_i^*(x) &\equiv Pr(D^* = i|X = x) < 1 \quad i \in \{1, 2\}, \quad \forall x.\end{aligned}$$

Restrictions imposed on the misclassification probabilities can help simplify the relationship between moments involving D^* and moments involving D , and therefore the analytical tractability of the problem. With our dose-response application in mind we thus impose that misclassification can occur only for *adjacent* categories of education, that is

$$\Pi(x) = \begin{bmatrix} \lambda_{00}(x) & \lambda_{10}(x) & 0 \\ \lambda_{01}(x) & \lambda_{11}(x) & \lambda_{21}(x) \\ 0 & \lambda_{12}(x) & \lambda_{22}(x) \end{bmatrix}, \quad (8)$$

which is a function of the four unknown probabilities (because of three adding up conditions): $\lambda_{00}(x)$, $\lambda_{11}(x)$, $\lambda_{22}(x)$ and $\lambda_{21}(x)$. The following proposition extends Proposition 1 to the case of multiple treatments.

Proposition 3 (*Bias on Conditional Treatment Effects*) *Provided that the determinant of (8) is different from zero:*

$$\delta(x) \equiv \lambda_{00}(x)[\lambda_{22}(x) - \lambda_{21}(x)] - \lambda_{22}(x)[1 - \lambda_{11}(x) - \lambda_{21}(x)] \neq 0$$

and if Assumptions 3 and 6 hold, we have that:

$$\begin{aligned} \Delta_{10}^*(x) &= \frac{\lambda_{22}(x) - \lambda_{21}(x)}{\delta(x)} \Delta_{10}(x) - \frac{\lambda_{21}(x)}{\delta(x)} \Delta_{21}(x), \\ \Delta_{21}^*(x) &= \frac{\lambda_{11}(x) + \lambda_{21}(x) - 1}{\delta(x)} \Delta_{10}(x) + \frac{\lambda_{00}(x) + \lambda_{11}(x) + \lambda_{21}(x) - 1}{\delta(x)} \Delta_{21}(x), \\ \Delta_{20}^*(x) &= \frac{\lambda_{11}(x) + \lambda_{22}(x) - 1}{\delta(x)} \Delta_{10}(x) + \frac{\lambda_{00}(x) + \lambda_{11}(x) - 1}{\delta(x)} \Delta_{21}(x), \end{aligned}$$

the last equation following from the adding-up condition.

The true conditional effects can thus be expressed as weighted sums of the raw effects $\Delta_{10}(x)$ and $\Delta_{21}(x)$. Weights are such that in the absence of misclassification raw effects coincide with true effects. Most importantly, in sharp contrast to the binary treatment case (see Proposition 1), the effects of misclassification on the relationship between $\Delta_{ij}^*(x)$ and $\Delta_{ij}(x)$ is not easily pinned down; depending on the extent and nature of misreporting across all categories, as well as on the sign and magnitude of both $\Delta_{10}(x)$ and $\Delta_{21}(x)$, $\Delta_{ij}(x)$ could be upward or downward biased for $\Delta_{ij}^*(x)$.¹⁵

¹⁵Nonetheless, by assuming that the mean response is monotonically increasing with D^* (which corresponds to assuming non-negative wage returns) and that the misclassification error is non-differential, one can derive conditions on the extent of misclassification under which $\Delta_{ij}^*(x)$ and $\Delta_{ij}(x)$ have at least the same sign. Battistin and Sianesi (2006a) show that if $\lambda_{ii}(x) > 0.5$ for all $i \in \{0, 1, 2\}$ sign reversal is avoided.

The following proposition extends Proposition 2 to the case of multiple treatments, showing that the true incremental ATT's can be represented as the sum of two components, each of which in the form of a weighted average. Given the results in Proposition 3, it should come to no surprise that the sign of the bias remains in general indeterminate.

Proposition 4 (*Bias on Treatment Effects for the Treated*) *If Assumptions 3 and 6 are satisfied and provided that $\delta(x) \neq 0$, the relationship between the true incremental ATT's and the effects estimated from raw data can be written as follows*

$$\Delta_{ij}^* = \int \omega_{1,ij}(x)\Delta_{10}(x)f(x|D = 1)dx + \int \omega_{2,ij}(x)\Delta_{21}(x)f(x|D = 2)dx,$$

where $i > j$ and $(i, j) \in \{0, 1, 2\}$ and the weights are reported in the Appendix.

7 Data and educational qualifications of interest

7.1 Data

In this paper we only consider methods relying on the conditional independence assumption, 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. There are extensive and commonly administered ability tests at early ages (mathematics and reading ability at ages 7 and 11), as well as accurately measured family background (parental education and social class) and school type variables, all ideal for methods relying on the conditional independence assumption. In fact, Blundell, Dearden and Sianesi (2005) 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 this 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, and our measure of educational qualifications is the one self-reported by respondents at age 33. Our sample of 3,642 is obtained by focusing on males only and restricting attention to those in work in 1991 with non-missing wage and education information.

7.2 Educational qualifications of interest

To put into context the educational qualifications to which we estimate the returns, we briefly outline the educational system in Britain (for more details, see Battistin and Sianesi, 2006a). 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 Ordinary Levels (O level) at 16 and then possibly move on to attain Advanced Levels (A levels) at the end of secondary school at 18. A levels still represent the primary route into higher education (HE). The academic and wide range of vocational qualifications have been classified into equivalent National Vocational Qualification (NVQ) levels, ranging from level 1 to level 5. The British system is thus quite distinct from the one in the US; nevertheless, forcing some comparisons, one could regard the no-qualifications group as akin to the group of high-school drop-outs, A levels to High School, and Higher Education to College.¹⁶

In our binary framework we consider the following two parameters in turn:

1. the return to achieving **any academic qualification** (level 2) compared to none¹⁷;
2. the return from undertaking **some form of higher education** (level 4 or 5) compared to anything less. This considers both the academic route and its vocational equivalent.

When we extend our framework to multiple treatments we consider incremental returns to the following three broad and sequential education levels:

1. **no qualifications** (level 1 or 0) - This treatment level basically reflects dropping out of school with no academic qualifications without later undertaking any vocational studies or formally recognized practice.

¹⁶In such a comparison the group with O levels as highest qualification is quite atypical, being made up of individuals who stop at the minimum leaving age with formal qualifications.

¹⁷This amounts to acquiring at least O levels compared to leaving school at the minimum age of 16 without any formal qualification, the counterfactual being thus akin to high-school drop-out status in the US. This parameter reflects a very well defined and homogenous qualification, and it captures all the channels in which the attainment of O levels 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 HE. Additional policy relevance of the returns to O levels arises from the finding that reforms raising the minimum school leaving age in the UK have impacted on individuals achieving low academic qualifications, in particular O levels (Chevalier *et al.*, 2003, Del Bono and Galindo-Rueda, 2004).

2. **intermediate qualifications** (level 2) - In addition to the academic O level exams held at age 16, this category includes their vocational equivalent.
3. **advanced qualifications** (level 3 or above) - This level requires at least high-school diploma (A levels) or their vocational equivalent; it thus includes all the more advanced qualifications up to university and postgraduate studies, and their vocational equivalent (e.g. professional degrees).

Table 1 shows the sample educational split corresponding to our parameters of interest.

8 Partial identification of causal effects in the presence of misclassification

8.1 Estimation issues

In this section we discuss how we derived estimates of the ATT for known values of the misclassification probabilities and confidence intervals for the partially identified ATT. In our empirical application we implement this approach to provide bounds to the returns to the educational qualifications outlined above.

Such bounds can be derived by exploiting the *observed* propensity score in a non/semi-parametric way, in particular allowing for arbitrarily heterogeneous individual returns and leaving the no-treatment outcome equation unspecified. Furthermore, both for the binary and the multiple treatment case, we allow misreporting to depend on X , albeit only through the propensity score index. To the best of our knowledge, we are the first ones to use the observed propensity score as a solution to deal with the curse of dimensionality arising from all these sources. In fact, the applications we are aware of which consider the estimation of the returns to educational attainment in the presence of misreporting either ignore the presence of X (Black, Berger and Scott, 2000); assume linearity, homogeneity of returns and misclassification independent of X (Kane, Rouse and Staiger, 1999); or impose a parametric structure to ease estimation (Lewbel, 2006). The issue is further considered in our companion paper (see Battistin and Sianesi, 2006b).

The idea underlying our estimation strategy is most simply put across by considering the case of binary treatments, though it can be trivially extended to the case of multiple treatments. If the misclassification probabilities are constant with respect to X , we have shown that the

weights in Proposition 2 vary with X only through the observed score $e(x)$. By applying the law of iterated expectations to the term on the right-hand-side of (7) we get

$$\begin{aligned} E\{\omega(x)\Delta(x)|D = 1\} &= E\{\omega(x)E\{\Delta(x)|D = 1, e(x)\}|D = 1\}, \\ &= E\{\omega(x)\Delta[e(x)]|D = 1\}, \end{aligned} \tag{9}$$

the last expression following since x is finer than $e(x)$ and from the observed propensity score being a balancing score for the distribution of X for individuals with $D = 1$ and $D = 0$. Known values of the probabilities of correct classification uniquely define the weights $\omega(x)$, and alternative estimators of the ATT result from considering the empirical analogue of (9).

Note that semi-parametric estimation is also feasible if weights are constant within cells defined by the propensity score $e(x)$ or, alternatively, if the misclassification probabilities are left to vary with X through $e(x)$. In this case we have

$$P(D^* = 1) = E\{1 - \lambda_0[e(x)]\} + E\{(\lambda_0[e(x)] + \lambda_1[e(x)] - 1)e(x)\}$$

and weights in Proposition 2 can be used.

By using Proposition 2, bounds on the true ATT can be derived by taking the maximum and the minimum value of the estimate of Δ^* when the probabilities $\lambda_0(x)$ and $\lambda_1(x)$ vary over the unit interval, or on a suitably chosen subset of $[0, 1] \times [0, 1]$. Indeed, leaving the misclassification probabilities to vary between zero and one is most likely to imply unreasonably high misclassification rates for the problem under consideration. One possibility is to use *a priori* restrictions on these probabilities derived from previous studies or from knowledge of the economic context under investigation. For example, results from validation studies and behavioral theories developed in the social sciences often suggest restrictions on misclassification.

In our application, we consider the empirical analogue of (9) by stratifying observations on the value of the propensity score and by allowing the λ 's be stratum-specific. This amounts to assuming that the misclassification probabilities are heterogeneous across individuals depending on values of the observed propensity score. Partial identification of the relevant ATT is then obtained through Proposition 2 (in the binary case) or Proposition 4 (in the multiple-treatment case) by considering the maximum and the minimum values of the Δ^* 's over the set defined by the sum of the exact classification probabilities exceeding a given value k and by imposing that overreporting is more likely than underreporting.

More specifically, in the binary case we first account for the high dimensionality of X by using stratification matching (see e.g. Dehejia and Wahba, 1999) and regression adjustment within each stratum.¹⁸ We then specify a grid for $\lambda_0(x)$ and $\lambda_1(x)$, compute the values of Δ^* using Proposition 2 for all combinations with $\lambda_0(x) \geq \lambda_1(x)$ (see Section 5.4), and take the maximum and the minimum over the sets defined by $\lambda_0(x) + \lambda_1(x) \geq k$, for $k \in [1.6, 2]$.¹⁹

In the multiple treatment case, we deal with the high dimensionality of X by defining strata on the two scores $e_1(x)$ and $e_2(x)$, again performing regression adjustment within each stratum to account for residual imbalance.²⁰ For given values of λ_{22} and λ_{21} , we then construct a grid for $\lambda_{00}(x)$ and $\lambda_{11}(x)$, and impose that $\lambda_{11}(x) \leq 1 - \lambda_{21}(x)$, which is needed to ensure that the resulting $\Pi(x)$ is indeed a probability matrix.²¹ We finally further ensure that overreporting is more likely than underreporting, which we translate into the condition that all the elements of $\Pi(x)$ above the diagonal are smaller than those below.²² We finally derive bounds on the incremental ATT's using Proposition 4 and by proceeding as in the binary case.

A final issue concerns the significance of our estimates. A growing body of research in the last years has looked into the problem of constructing confidence intervals for partially identified parameters (see, for example, Romano and Shaikh, 2006, and Chernozhukov, Hong, and Tamer, 2007). In our application, we follow Horowitz and Manski (2000) and derive confidence intervals for bounds that cover the entire identification region with 95 percent probability. By denoting with \hat{L} and \hat{U} the lower and the upper bounds, we report confidence intervals of the form $[\hat{L} - \zeta, \hat{U} + \zeta]$, where ζ is a positive constant obtained by bootstrapping the distribution of bounds so as to ensure the required probability of coverage. As stated in Imbens and Manski (2004; see Lemma 1), the probability that the interval considered covers the true ATT is at

¹⁸Given our relatively small sample size, we performed the latter to account for residual imbalance of X within stratum (both in the case of any academic qualification and HE we could reject joint balancing of the observables at 10% for only one stratum). Within each stratum s on the estimated propensity score, we run a regression of Y on $e(x)$, separately for the $D = 1$ and $D = 0$ groups, and calculate the corresponding raw conditional treatment effect $\Delta(e_s)$ as the difference in predicted outcomes at the mean stratum propensity score. The propensity score is estimated from a logit regression.

¹⁹Note incidentally that because of (6) and the fact that $e(x) \in [0, 1]$, the following restrictions apply (under case $\lambda_0(x) + \lambda_1(x) > 1$): $\lambda_0(x) \geq 1 - e^*(x)$ and $\lambda_1(x) \geq e^*(x)$. Thus assigning values to $\lambda_0(x)$ and $\lambda_1(x)$ has implications on the range of values taken on by the true (but unobserved) propensity score.

²⁰Within each stratum, we regress Y on $e_1(x)$, $e_2(x)$ and $e_1(x) \cdot e_2(x)$ separately for each group defined by D .

²¹Note that, as in the binary case, the assumption we maintain that observations on (each value of) D are more accurate than pure guesses, $\lambda_{ii}(x) \geq 0.5$ for $i = 0, 1, 2$, implies that sign reversal is avoided and that misclassification is limited, in the sense that the probability of picking someone who truly has a given qualification is higher when drawing randomly from the group that claims to have that qualification rather than from the group claiming to have a different one (see Battistin and Sianesi, 2006a).

²²Algebraically, this amounts to imposing $\lambda_{00}(x) \geq \max\{\lambda_{11}(x) + \lambda_{21}(x), \lambda_{22}(x)\}$.

least 95 percent (thus leading to conservative inference).

8.2 Results

8.2.1 Binary levels of attainment

The return Δ from attaining any academic qualification at 16 for those who did so is estimated at 23.4 percent in the raw data (using the full set of controls outlined in Section 7.1). The average return to higher education for graduates is estimated at 23.1 percent (see Table 2).

We investigated the sensitivity of these estimates to the presence of misclassification by performing several types of analyses. We first used Proposition 2 to bound the true returns Δ^* by considering the case of misclassification independent of X when λ_0 and λ_1 are left to vary between 70 (very severe misclassification) and 100 percent (exact reporting). The lhs panel in Figure 1 plots how the value of the true return to any academic qualification varies as a function of the extent of misclassification. The minimum and the maximum values of Δ^* are 23.4 percent and 55.7 percent and are achieved for $\lambda_0 = 1$ and $\lambda_1 = 1$, and for $\lambda_0 = 0.7$ and $\lambda_1 = 0.7$. We thus find that, in line with our previous discussion, returns estimated from raw data yield the lower bound for the true return. A similar result holds for the return to higher education, for which the identification region corresponds to (23.1, 60.1).

The rhs panel in Figure 1 considers the identifying power of assuming that over-reporting is more likely than under-reporting. This restriction embodies the finding of cognitive studies that respondents tend to over-report desirable features and behaviours, and corresponds to finding the maximum and the minimum in the region for which $\lambda_0 \geq \lambda_1$. It is evident from the figure that such a restriction does not help improve identification power, as the lower and upper bounds coincide with those defined from the unconstrained region.

A further restriction to obtain tighter bounds is $\lambda_0 + \lambda_1 \geq k$ for increasing values of k . This corresponds to consider the intersection of the rhs panel in Figure 1 with two-dimensional planes that increasingly shift to the right for higher values of k . We implemented this idea for the case of misclassification probabilities dependent on X as described in the previous section. Table 2 reports the lower and upper bounds for the returns to attaining any academic qualification and to completing higher education for a number of values of k . Each set of lower and upper bounds thus relates to a different value of the sum of the misclassification probabilities. Moreover, we report the 95 percent confidence intervals for the identification region.

As expected, the bounds become more informative as the sum of λ 's approaches 2. An interesting result is that even when we allow for a non-negligible extent of misclassification - up to 10 percent of individuals misreporting their attainment in either direction - the point estimates of the lower and upper bounds for the returns are quite close, around 23-26 percent for both any academic qualification and higher education.

By using the uniquely rich NCDS data we can assess the plausibility that the biases from measurement error and from omitted variables cancel out in the estimation of returns in the UK, which would allow policy-makers to obtain up-to-date estimates of returns to qualifications using Labour Force Survey-type datasets, which rely on recall about individuals' educational attainment and do not contain any information on individual's ability and background.

To perform this exercise, we have estimated the return from the raw data controlling only for the Labour Force Survey-type variables of gender, age, ethnicity and region (gender and age implicitly via sample choice), Δ_{LFS} . For academic qualifications, Δ_{LFS} is 32.9 percent and for HE 35.9 percent. As to returns to any academic qualification, for the sum of the λ 's roughly larger than 1.8, ignoring both measurement error and ability biases yields an upper bound. More generally, we find that there is a chance for the two biases to cancel out only if misclassification is rather severe, in particular when at least 20 percent of individuals misreport their qualifications in either direction. In fact, for the case where $\lambda_0 + \lambda_1 \geq 1.9$, which is in line with the little available evidence so far (see in particular Kane, Rouse and Staiger, 1999), even the conservative confidence intervals do not contain the point estimate of Δ_{LFS} . As to the returns to higher education, it is even harder to appeal to the cancelling of the two biases.²³

We also derived the bounds for the case of constant λ 's. As expected, they were tighter; they however led to the same inferential conclusions about Δ^* and its relationship with Δ_{LFS} .

8.2.2 Multiple levels of attainment

Turning to a more disaggregated analysis in a multiple treatment framework, we focus on the incremental returns to three sequential broad education levels: no qualifications, O levels or vocational equivalent ("intermediate") qualifications and at least A levels or vocational equivalent ("advanced") qualifications. Ignoring potential misclassification, the incremental

²³This discussion is heuristic in that it ignores uncertainty in the estimated Δ_{LFS} . A more formal indicator would be obtained by considering the proportion of bootstrapped values of Δ_{LFS} that fall within the corresponding bootstrapped upper and lower bounds.

ATT's to acquire intermediate qualifications is 10.6 percent, to move from intermediate to advanced qualifications is 18.4 percent, and to acquire advanced qualifications compared to remaining with none is 28.3 percent.

We then proceeded to assess the robustness of these estimates to the presence of misclassification in recorded educational attainment. Starting with the case of misclassification independent of X , we considered the bounds which arise when letting the exact classification probabilities (λ_{00} , λ_{11} and λ_{22}) vary between 70 (very severe misclassification) and 100 percent (exact classification), and the probability of ‘forgetting’ their advanced qualifications by those stating to have only intermediate ones (λ_{21}) between 0 and 15 percent.

The ensuing identification regions are quite wide: (5.8, 27.1) for Δ_{10}^* , (11.3, 33.0) for Δ_{21}^* , and (27.9, 39.9) for Δ_{20}^* . As opposed to our binary-treatment application, imposing the restriction that over-reporting is more likely than underreporting does improve identification power, albeit only for Δ_{10}^* (a 5.6 percentage reduction in width by lowering the upper bound) and Δ_{21}^* (a 14.1 percentage reduction by raising the lower bound).

To gain some further insight we move to a graphical analysis. This is similar to what done for the binary treatment; however to overcome the additional difficulty that we now are working in a 5-dimensional space, we fix two dimensions, λ_{22} and λ_{21} . Specifically, we consider two rather extreme but still plausible profiles: one of severe misclassification in these two dimensions ($\lambda_{22} = 0.90$, $\lambda_{21} = 0.05$) and one of little misclassification ($\lambda_{22} = 0.95$, $\lambda_{21} = 0.01$).

In contrast to the binary case, one can visually appreciate from Figure 2 that the restriction in terms of over/under-reporting significantly increases identifying power (the graphs for the little misclassification $\{\lambda_{22}, \lambda_{21}\}$ profile is reported in Battistin and Sianesi, 2006a). For Δ_{10}^* and Δ_{20}^* , the gain from the restriction arises from ‘cropping’ the top, i.e. lowering the upper bound; for Δ_{21}^* , from raising the lower bound instead. Comparing the three ATT's, the percentage reduction in bounds width is largest for Δ_{20}^* for either $\{\lambda_{22}, \lambda_{21}\}$ profile. The percentage gain is in fact larger in the case of little as opposed to severe misclassification in terms of λ_{22} and λ_{21} . The identification regions for the three parameters under the restriction are in fact quite similar for the two profiles, and are roughly 10 to 15 percent for for Δ_{10}^* , 16 to 20 percent for Δ_{21}^* , and 29 to 30 percent for Δ_{20}^* . Indeed, under our restriction, Δ_{20}^* is very tightly bound under either profile.

Next, we allow for heterogeneity in $\lambda_{00}(x)$ and $\lambda_{11}(x)$, and investigate the identifying power

of considering bounds that in addition to keeping the values of $\{\lambda_{22}, \lambda_{21}\}$ constant to those of either of the two profiles and meeting the over/under-reporting restriction, also satisfy $\lambda_{00}(x) + \lambda_{11}(x) \geq k$, for increasing values of k . The results are reported in Table 3, together with the 95 percent confidence intervals for the identification region.

As expected, for given $\{\lambda_{22}, \lambda_{21}\}$ profile, the bounds monotonically narrow as $\lambda_{00}(x) + \lambda_{11}(x)$ approaches 2. For Δ_{10}^* , bounds become more informative because of a lowering of the upper bound, for Δ_{21}^* because of a raising of the lower bound. For Δ_{20}^* , however, the bounds remain unchanged for $k = 1.6$ and $k = 1.7$, and for higher levels of k both the lower and upper bounds move closer. Interestingly, the result that bounds either stay the same or become narrower as we get closer to exact reporting in terms of λ_{00} and λ_{11} no longer applies in terms of λ_{22} and λ_{21} . Specifically, bounds do not always become more informative when moving from severe to little misclassification in terms of λ_{22} and λ_{21} ; quite to the contrary they become wider for Δ_{10}^* and Δ_{21}^* in all cases except $k = 1.6$ for Δ_{10}^* .

As to how the ‘raw’ incremental effects Δ_{ij} ’s relate to the bounds, in contrast to the binary case no clear pattern applies. In fact, although the Δ_{ij} ’s are mostly downward biased, in terms of point estimate at times the estimated lower bound proves sharper than the raw effect. Furthermore, the raw effects could at times be upward biased, even though in our application this could happen only for Δ_{21}^* and, interestingly, only for a situation of more severe misclassification in all dimensions (λ_{00} , λ_{11} , λ_{22} and λ_{21}).

Finally, we consider the relationship between the bounds for the true incremental ATT’s and the naive estimates based on the raw data and LFS-style controls. For the two sets of values of $\{\lambda_{22}, \lambda_{21}\}$ we have considered, ignoring misreporting and ability biases (Δ_{LFS}) yields an upward biased estimate, in the sense that the Δ_{LFS} ’s always lie well outside of the bounds. Whatever the sum of λ_{00} and λ_{11} , we thus find no evidence that there is a chance for the two biases to cancel out. For Δ_{20}^* , this conclusion keeps holding in terms of the (conservative) CI for the identification region - estimating returns to advanced qualifications compared to none ignoring misclassification and selection yields severely upward biased estimates. For Δ_{10}^* , and especially Δ_{21}^* , this conclusion is more likely to hold, as one would expect, in the case of less severe misclassification ($k = 1.8, 1.9$, and $\lambda_{00} = 0.95, \lambda_{21} = 0.01$).

In contrast to the ‘traditional’ binary treatment case, when account is taken of potential misclassification across multiple treatments, the results in this section have shown that the

patterns that emerge can be exceptionally varied. However, the conclusion that in our application we cannot expect measurement error to cancel out all ability bias keeps holding in our multiple-treatment extension.

9 Conclusions

We have characterized the bias introduced by misclassification of the treatment indicator on the ATT under the conditional independence assumption. In general, matching-type estimators computed from misclassified data are biased for the true ATT, without it being a priori possible to sign the bias. We have provided conditions under which the attenuation bias result is still most likely to hold. We have further extended the framework to multiple treatments.

Results for partial identification of the ATT straightforwardly follow from our characterization of the bias. We propose an estimation strategy that allows one to derive bounds in a non/semi-parametric way.

Our empirical application has demonstrated the usefulness of the suggested approach by applying it to the estimation of returns to educational qualifications in the UK, both in a binary and multiple-treatment setting. Specifically, we have investigated the sensitivity of the raw estimates to the presence of misclassification, and explored the identification power of plausible restrictions on the nature and extent of misclassification. Lastly, we have assessed the plausibility for the two biases from measurement error and from omitted variables to cancel out. Under the most plausible scenarios as regards the extent of misreporting, it is indeed very hard to appeal to cancelling; ignoring both misreporting and omitted ability bias would generally lead to at times quite severely upward biased estimates of true returns.

More generally, our results have shown that under relatively mild restrictions we can obtain strong conclusions regarding our question of interest, although more assumptions are needed to obtain statistical significance.

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. (2006a), *Misreported Schooling and Returns to Education: Evidence from the UK*, Cemmap WP No. CWP07/06, London.
- [3] Battistin, E. and Chesher, A. (2007), *Treatment Effect Estimation with Covariate Measurement Error*, unpublished manuscript, Institute for Fiscal Studies.
- [4] Battistin, E. and Sianesi, B. (2006b), *Misreported Schooling, Multiple Measures and Returns to Educational Qualifications*, unpublished manuscript, Institute for Fiscal Studies.
- [5] Black, D., M. Berger, and F. Scott (2000), *Bounding Parameter Estimates with Non-Classical Measurement Error*, Journal of the American Statistical Association, 95, 451, 739-48.
- [6] 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.
- [7] Blundell, R., Dearden, L., Goodman, A. and Reed, H. (2000), *The Returns to Higher Education in Britain: Evidence from a British Cohort*, Economic Journal, 110, F82–F99.
- [8] Blundell, R. Dearden, L. Sianesi, B. (2004), *Measuring the Returns to Education*, chapter 6 in Machin, S. and Vignoles, A. (eds.), *The Economics of Education in the UK*, Princeton University Press, forthcoming.
- [9] Blundell, R. Dearden, L. Sianesi, B. (2005), *Evaluating the Impact of Education on Earnings in the UK: Models, Methods and Results from the NCDS*, forthcoming, Journal of the Royal Statistical Society A.
- [10] Bonjour, D., Cherkas, L., Haskel, J., Hawkes, D., and Spector, T., (2003), *Education and Earnings: Evidence from UK Twins*, American Economic Review, December, 1799-1812.

- [11] 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.
- [12] Card, D. (1996), *The Effect of Unions on the Structure of Wages: A Longitudinal Analysis*, *Econometrica*, Vol.64, No.4, pp.957-979.
- [13] 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.
- [14] Chevalier, A., Harmon, C., Walker, I. and Zhu, Y. (2003), *Does education raise productivity?*, University College Dublin Working Paper, ISSC, WP2003/01, Dublin.
- [15] Chernozhukov, V., Hong, H. and Tamer, E. (2007), *Estimation and Inference on Identified Parameter Sets*, *Econometrica*, forthcoming.
- [16] Dearden, L. (1999a), *The Effects of Families and Ability on Men's Education and Earnings in Britain*, *Labour Economics*, 6, 551-67.
- [17] Dearden, L. (1999b), *Qualifications and earnings in Britain: how reliable are conventional OLS estimates of the returns to education?*, IFS working paper W99/7.
- [18] Dearden, L., McIntosh, S., Myck, M. and Vignoles, A. (2000), *The Returns to Academic and Vocational Qualifications in Britain*, Centre for the Economics of Education Discussion Paper No. 04.
- [19] 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.
- [20] Dehejia, R.H and Wahba, S. (1999), *Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programmes*, *Journal of the American Statistical Association* 94, 1053-1062.
- [21] 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.

- [22] Gosling, A., Machin, S. and Meghir, C. (2000), *The changing distribution of male wages, 1966–93*, Review of Economic Studies, 67, 635-666.
- [23] Griliches, Z. (1977), *Estimating the returns to schooling: some econometric problems*, Econometrica, 45, 1–22.
- [24] Heckman, J.J. and Robb, R. (1985), *Alternative Methods for Evaluating the Impact of Interventions*, in Heckman, J.J. and Singer, B. (eds.), Longitudinal Analysis of Labour Market Data, Cambridge University Press, 156-246.
- [25] 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.
- [26] Horowitz, J.L. and Manski, C.F. (2000), *Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data*, Journal of the American Statistical Association, Vol. 95, No. 449, pp. 77-84.
- [27] Hu, Y. (2007), *Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution*, mimeo, University of Texas at Austin
- [28] Imbens, G.W. (2000), *The Role of the Propensity Score in Estimating Dose-Response Functions*, Biometrika, 87, 3, 706-710.
- [29] Imbens, G.W. (2004), *Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review*, Review of Economics and Statistics, 86, 4-29.
- [30] Imbens, G.W. and Manski, C.F. (2004), *Confidence Intervals for Partially Identified Parameters*, Econometrica, Vol. 72, No. 6, pp. 1845-1857.
- [31] Ives, R. (1984), *School reports and self-reports of examination results*, Survey Methods Newsletter, Winter 1984/85, 4-5.
- [32] Kane, T.J., Rouse, C. and Staiger, D. (1999), *Estimating Returns to Schooling when Schooling is Mismeasured*, National Bureau of Economic Research Working Paper No. 7235.

- [33] Lechner, M. (2001), *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*, in Lechner, M., Pfeiffer, F. (eds), *Econometric Evaluation of Labour Market Policies*, Physica/Springer, Heidelberg, 43-58.
- [34] Lewbel, A., (2006), *Estimation of Average Treatment Effects With Misclassification - Addendum*, mimeo.
- [35] Lewbel, A. (2007), *Estimation of Average Treatment Effects With Misclassification*, *Econometrica*, 75, 537-551.
- [36] Mahajan, A. (2006), *Identification and Estimation of Regression Models with Misclassification*, *Econometrica*, 74, 3, 631-665.
- [37] McIntosh, S. (2004), *Further Analysis of the Returns to Academic and Vocational Qualifications*, Centre for the Economics of Education Discussion Paper No. 35.
- [38] Molinari, F. (2007), *Partial Identification of Probability Distributions with Misclassified Data*, unpublished manuscript, Department of Economics, Cornell University.
- [39] Quandt, R. (1972), *Methods for Estimating Switching Regressions*, *Journal of the American Statistical Association*, 67, 306-310.
- [40] Romano, J. and Shaikh, A. (2006) Inference for identifiable parameters in partially identified econometric models. Technical Report 2006(9), Department of Statistics, Stanford University.
- [41] Roy, A. (1951), *Some Thoughts on the Distribution of Earnings*, *Oxford Economic Papers*, 3, 135-146.
- [42] 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.
- [43] Rubin, D.B. (1974), *Estimating Causal Effects of Treatments in Randomised and Non-randomised Studies*, *Journal of Educational Psychology*, 66, 688-701.
- [44] 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.

- [45] Sianesi, B. (2003), *Returns to Education: A Non-Technical Summary of CEE Work and Policy Discussion*, mimeo, June.

Appendix

Proof of Proposition 2

Using Bayes' theorem we get

$$\begin{aligned}f(x|D = 1) &= \frac{e(x)f(x)}{Pr(D = 1)}, \\f(x|D^* = 1) &= \frac{e^*(x)f(x)}{Pr(D^* = 1)},\end{aligned}$$

where $e(x)$ is the propensity score calculated from D . Since by the law of iterated expectations we have

$$e^*(x) = [1 - \lambda_0(x)] + e(x)[\lambda_0(x) + \lambda_1(x) - 1],$$

it also follows that

$$\begin{aligned}Pr(D^* = 1) &= \int e^*(x)f(x)dx, \\&= \int [1 - \lambda_0(x)]f(x)dx + \int e(x)[\lambda_0(x) + \lambda_1(x) - 1]f(x)dx.\end{aligned}$$

Since

$$\begin{aligned}f(x|D^* = 1) &= \frac{f(x|D^* = 1)}{f(x|D = 1)} f(x|D = 1) \\&= \frac{Pr(D = 1)}{Pr(D^* = 1)} \frac{e^*(x)}{e(x)} f(x|D = 1) \\&= \frac{Pr(D = 1)}{Pr(D^* = 1)} \left[\frac{1 - \lambda_0(x)}{e(x)} + \lambda_0(x) + \lambda_1(x) - 1 \right] f(x|D = 1),\end{aligned}$$

we can use (5) to write

$$\begin{aligned}\Delta^* &= \int \Delta^*(x)f(x|D^* = 1)dx, \\&= \frac{Pr(D = 1)}{Pr(D^* = 1)} \int \Delta(x) \left[1 + \frac{1}{e(x)} \frac{1 - \lambda_0(x)}{\lambda_0(x) + \lambda_1(x) - 1} \right] f(x|D = 1)dx, \\&= \int \omega(x)\Delta(x)f(x|D = 1)dx,\end{aligned}$$

where

$$\omega(x) = \frac{Pr(D = 1)}{Pr(D^* = 1)} \left[1 + \frac{1}{e(x)} \frac{1 - \lambda_0(x)}{\lambda_0(x) + \lambda_1(x) - 1} \right],$$

and

$$\frac{Pr(D^* = 1)}{Pr(D = 1)} = \frac{\int [1 - \lambda_0(x)]f(x)dx}{\int e(x)f(x)dx} + \frac{\int [\lambda_0(x) + \lambda_1(x) - 1]e(x)f(x)dx}{\int e(x)f(x)dx}. \blacksquare$$

Weights of Proposition 4

Weights for the case of misclassification independent of X are defined as follows:

$$\begin{aligned} \omega_{1,10}(x) &= \frac{P[D = 1](\lambda_{22} - \lambda_{21})(\lambda_{11} + \lambda_{00} - 1 + \frac{P[D=2|x](\lambda_{00}-\lambda_{22})+1-\lambda_{00}}{P[D=1|x]})}{\delta(x)(1 - \lambda_{00} + (\lambda_{11} + \lambda_{00} - 1)P[D = 1] + (\lambda_{00} - \lambda_{22})P[D = 2])}, \\ \omega_{2,10}(x) &= -\frac{P[D = 2]\lambda_{21}(\lambda_{00} - \lambda_{22} + \frac{P[D=1|x](\lambda_{00}+\lambda_{11}-1)+1-\lambda_{00}}{P[D=2|x]})}{\delta(x)(1 - \lambda_{00} + (\lambda_{11} + \lambda_{00} - 1)P[D = 1] + (\lambda_{00} - \lambda_{22})P[D = 2])}, \\ \omega_{1,21}(x) &= \frac{P[D = 1](\lambda_{11} + \lambda_{21} - 1)(\lambda_{21} + \lambda_{22} \frac{P[D=2|x]}{P[D=1|x]})}{\delta(x)(\lambda_{21}P[D = 1] + \lambda_{22}P[D = 2])}, \\ \omega_{2,21}(x) &= \frac{P[D = 2](\lambda_{00} + \lambda_{11} + \lambda_{21} - 1)(\lambda_{22} + \lambda_{21} \frac{P[D=1|x]}{P[D=2|x]})}{\delta(x)(\lambda_{21}P[D = 1] + \lambda_{22}P[D = 2])}, \\ \omega_{1,20}(x) &= \frac{P[D = 1](\lambda_{11} + \lambda_{22} - 1)(\lambda_{21} + \lambda_{22} \frac{P[D=2|x]}{P[D=1|x]})}{\delta(x)(\lambda_{21}P[D = 1] + \lambda_{22}P[D = 2])}, \\ \omega_{2,20}(x) &= \frac{P[D = 2](\lambda_{00} + \lambda_{11} - 1)(\lambda_{22} + \lambda_{21} \frac{P[D=1|x]}{P[D=2|x]})}{\delta(x)(\lambda_{21}P[D = 1] + \lambda_{22}P[D = 2])}. \end{aligned}$$

The expressions for the general case can be found in Battistin and Sianesi (2006a).■

Figure 1: Identification region for the return to any academic qualification, assuming constant misclassification probabilities and, in the right-hand-side panel, that overreporting is more likely than underreporting

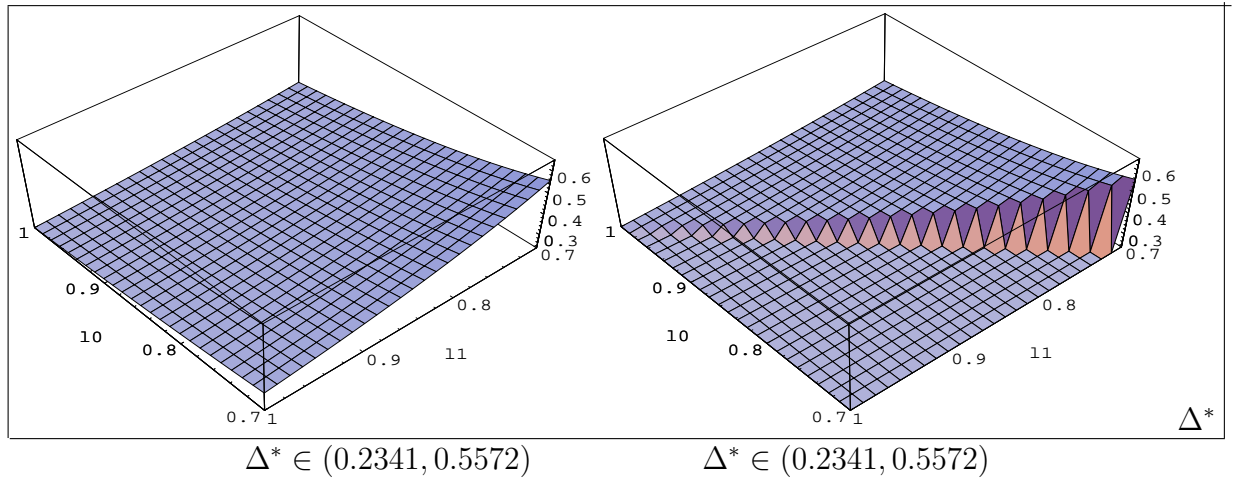


Table 1: Educational sample split ($N=3,642$)

	Multiple treatment	HE vs Less	Any Academic vs None
None	895 (25%)		
O/eq.	941 (26%)		
A/eq.+	1,806 (50%)		
No Acad			1,243 (34%)
Any Acad			2,399 (66%)
Below HE		2,472 (68%)	
HE		1,170 (32%)	

Table 2: Base case bounds on wage returns from Any Academic Qualification and from Higher Education

Δ^*	Any Qualification				Higher Education			
	Estimate		95% CI		Estimate		95% CI	
	lower	upper	lower	upper	lower	upper	lower	upper
$\lambda_0(x) + \lambda_1(x) \geq 1.6$	0.2336	0.3975	0.1421	0.4890	0.2310	0.4032	0.1926	0.4416
$\lambda_0(x) + \lambda_1(x) \geq 1.7$	0.2336	0.3340	0.1517	0.4159	0.2310	0.3378	0.1923	0.3765
$\lambda_0(x) + \lambda_1(x) \geq 1.8$	0.2336	0.2947	0.1595	0.3688	0.2310	0.2956	0.1947	0.3319
$\lambda_0(x) + \lambda_1(x) \geq 1.9$	0.2336	0.2607	0.1646	0.3297	0.2310	0.2597	0.1941	0.2966
Raw data:								
Full controls:	$\Delta = 0.2336$				$\Delta = 0.2310$			
LFS controls:	$\Delta_{LFS} = 0.3286$				$\Delta_{LFS} = 0.3588$			

Bounds are derived from Proposition 2 by allowing the misclassification probabilities $\lambda_0(x)$ and $\lambda_1(x)$ to depend on x through the propensity score $e(x)$ and by imposing that over-reporting is more likely than under-reporting, i.e. $\lambda_1(x) \leq \lambda_0(x)$ (see Section 5.4). Confidence intervals covering the identification region with 95 percent probability have been derived from 500 bootstrap replications following Horowitz and Manski (2000).

Table 3: Base case bounds on wage returns from incremental levels of attainment

Δ_{10}^*	$\lambda_{22}(x) = 0.95$ and $\lambda_{21}(x) = 0.01$				$\lambda_{22}(x) = 0.90$ and $\lambda_{21}(x) = 0.05$			
	Estimate		95% CI		Estimate		95% CI	
	lower	upper	lower	upper	lower	upper	lower	upper
$\lambda_0(x) + \lambda_1(x) \geq 1.6$	0.1057	0.1591	0.0331	0.2317	0.0981	0.1566	0.0249	0.2298
$\lambda_0(x) + \lambda_1(x) \geq 1.7$	0.1057	0.1504	0.0352	0.2209	0.0981	0.1392	0.0261	0.2112
$\lambda_0(x) + \lambda_1(x) \geq 1.8$	0.1057	0.1314	0.0358	0.2013	0.0981	0.1200	0.0285	0.1896
$\lambda_0(x) + \lambda_1(x) \geq 1.9$	0.1057	0.1156	0.0361	0.1852	0.0981	0.1046	0.0285	0.1742

Raw data:

Full controls: $\Delta_{10} = 0.1060$, LFS controls: $\Delta_{10,LFS} = 0.1922$

Δ_{21}^*	$\lambda_{22}(x) = 0.95$ and $\lambda_{21}(x) = 0.01$				$\lambda_{22}(x) = 0.90$ and $\lambda_{21}(x) = 0.05$			
	Estimate		95% CI		Estimate		95% CI	
	lower	upper	lower	upper	lower	upper	lower	upper
$\lambda_0(x) + \lambda_1(x) \geq 1.6$	0.1519	0.1944	0.0886	0.2577	0.1766	0.2159	0.1100	0.2825
$\lambda_0(x) + \lambda_1(x) \geq 1.7$	0.1550	0.1944	0.0926	0.2568	0.1820	0.2159	0.1166	0.2813
$\lambda_0(x) + \lambda_1(x) \geq 1.8$	0.1725	0.1944	0.1176	0.2493	0.1981	0.2159	0.1399	0.2741
$\lambda_0(x) + \lambda_1(x) \geq 1.9$	0.1860	0.1944	0.1341	0.2463	0.2107	0.2159	0.1549	0.2717

Raw data:

Full controls: $\Delta_{21} = 0.1843$, LFS controls: $\Delta_{21,LFS} = 0.2432$

Δ_{20}^*	$\lambda_{22}(x) = 0.95$ and $\lambda_{21}(x) = 0.01$				$\lambda_{22}(x) = 0.90$ and $\lambda_{21}(x) = 0.05$			
	Estimate		95% CI		Estimate		95% CI	
	lower	upper	lower	upper	lower	upper	lower	upper
$\lambda_0(x) + \lambda_1(x) \geq 1.6$	0.2874	0.2944	0.1899	0.3919	0.2975	0.3104	0.2015	0.4064
$\lambda_0(x) + \lambda_1(x) \geq 1.7$	0.2874	0.2944	0.1905	0.3913	0.2975	0.3104	0.2015	0.4064
$\lambda_0(x) + \lambda_1(x) \geq 1.8$	0.2883	0.2944	0.1908	0.3919	0.2991	0.3076	0.2016	0.4051
$\lambda_0(x) + \lambda_1(x) \geq 1.9$	0.2889	0.2917	0.1908	0.3898	0.3004	0.3029	0.2017	0.4016

Raw data:

Full controls: $\Delta_{20} = 0.2825$, LFS controls: $\Delta_{20,LFS} = 0.4339$

Bounds are derived from Proposition 4 by allowing the misclassification probabilities $\lambda_0(x)$ and $\lambda_1(x)$ to depend on x through the propensity scores $e_1(x) \equiv P(D = 1|x)$ and $e_2(x) \equiv P(D = 2|x)$ and by imposing that over-reporting is more likely than under-reporting (see Section 8.1 for details). Confidence intervals covering the identification region with 95 percent probability have been derived from 500 bootstrap replications following Horowitz and Manski (2000).

Figure 2: Identification region for returns to incremental levels of attainment, assuming constant misclassification probabilities and by imposing $\lambda_{22} = 0.90$ and $\lambda_{21} = 0.05$; the assumption that overreporting is more likely than underreporting is superimposed in the right-hand-side panels

