A New Instrumental Method for Dealing with Endogenous Selection*

Xavier d'Haultfoeuille
Université Paris I-Panthéon-Sorbonne,
CREST - INSEE

First version: July 2008
This version: June 2009

Abstract

This paper develops a new method for dealing with endogenous selection. The usual instrumental strategy based on the independence between the outcome and the instrument is likely to fail when selection is directly driven by the dependent variable. Instead, we suggest to rely on the independence between the instrument and the selection variable, conditional on the outcome. This approach may be particularly suitable for nonignorable nonresponse, binary models with missing covariates or Roy models with an unobserved sector. The nonparametric identification of the joint distribution of the variables is obtained under a completeness assumption, which has been used recently in several nonparametric instrumental problems. Even if the conditional independence between the instrument and the selection variable fails to hold, the approach provides sharp bounds on parameters of interest under weaker monotonicity conditions. Apart from identification, nonparametric and parametric estimations are also considered. Finally, the method is applied to estimate the effect of grade retention in French primary schools.

Keywords: endogenous selection, instrumental variables, nonparametric identification, completeness, inverse problems.

JEL classification numbers: C14, C31, C35.

*I am particularly grateful to Jean-Claude Deville for inspiring me to start this project and to Stéphane Bonhomme for his fruitful suggestions. I also wish to thank three anonymous referees, Romain Aeberhardt, Marine Carrasco, Elise Coudin, Bruno Crépon, Laurent Davezies, Philippe Février, Jean-Pierre Florens, Edwin Leuven, Thierry Magnac, Charles Manski, Arnaud Maurel, Jean-Marc Robin and the participants of the ESEM and of the CEMMAP seminar for their helpful comments.

1 Introduction

Missing observations are very common in micro data, either because of selection, nonresponse or simply because counterfactual variables cannot be observed. Ignoring this issue by making inference on the observed population generally leads to inconsistent estimates. Moreover, without additional assumptions, only bounds on the parameters of interest can be identified (see, e.g., Manski, 2003). Several approaches have been followed to achieve point identification. The first one is to assume that the selection variable and the outcome are independent conditional on the observed covariates. This is the so-called missing-at-random assumption (see, e.g., Little and Rubin, 1987), also known as the unconfoundedness assumption in the treatment effect literature (see, for instance, Imbens, 2004). This assumption is often considered too stringent because it rules out any correlation between the selection and the outcome variables. The second approach is to rely on instruments that determine selection but not outcomes (see, e.g., Heckman, 1974, on sample selection models, Angrist et al., 1996, or Heckman and Vytlacil, 2005, on treatment effects). This assumption does not, however, point identify the distribution of the outcome in general (see Manski, 2003). Moreover, it may be difficult to find such instruments in practice, in particular when selection depends heavily on the dependent variable. The third approach relies on functional restrictions rather than exclusion restrictions. For instance, Chamberlain (1986) obtains identification at infinity by imposing a linear structure. Finally, Lewbel (2007) obtains identification under the existence of a special regressor that is strongly exogenous (i.e., conditionally independent of the errors of the selection model), a large support condition and restrictions on the probability of selection.[1]

In this paper, another instrumental strategy for solving endogenous selection is considered. Nonparametric identification is based on independence between the instrument and the selection variable, conditional on the outcome and possibly on other explanatory variables. This assumption has been also used in the framework of nonignorable nonresponse by Chen (2001), Tang et al. (2003), Hemvanich (2004) and Ramalho and Smith (2007).[2] Apart from nonresponse, this assumption may be particularly suitable when selection is directly driven by the dependent variable. Consider for instance a variable that is observed only conditionally on an unobserved truncation. Finding an instrument that only affects

[1] This probability has to tend to zero or one when the special regressor tends to infinity.

[2] The difference with these papers is that they focus mainly on parametric and semiparametric estimation issues, whereas the emphasis is put on nonparametric identification here. In particular, we generalize the identification results of Hemvanich (2004), obtained when the support of the outcome is finite, to any kind of outcome.

selection is impossible if this truncation variable is purely random. Instead, any variable that affects the dependent variable will satisfy the exclusion restriction considered here. Other examples where this assumption can be useful include Roy models with unobserved sector, one stratum response based samples or truncated count data models. As in usual instrumental regressions, a rank condition between the instrument and the outcome is also required to achieve identification. This condition is stated in terms of completeness and is already considered in several nonparametric instrumental problems (see, among others, Newey and Powell, 2003, Blundell et al., 2007, and Hu and Schennach, 2008). Under completeness and conditional independence, the joint distribution of the data is identified nonparametrically.[3] The key point is that it is enough to recover the probability of selection conditional on the outcome. This is similar to the unconfoundedness situation where the problem consists of identifying the propensity score. The difference between the two settings is that the identification of the propensity score is trivial under unconfoundedness, whereas the conditional probability we consider is more difficult to retrieve. We show that this function satisfies an integral inverse problem whose solution is unique under the completeness condition.

The joint distribution of the data can still be recovered under a parametric restriction on the selection model if only some moments of the instrument and not its full distribution are used. This result may be useful for estimation or when only aggregated information on the instrument is available. The idea of using moments of the instrument to deal with nonresponse has also been applied in survey sampling (see Deville, 2002). It is also related to the literature on auxiliary information, which has been developed either for efficiency reasons (see Imbens and Lancaster, 1994, Hellerstein and Imbens, 1999) or, as here, to provide identification (see Hellerstein and Imbens, 1999, and Nevo, 2002). Our parametric framework extends Nevo's result to the case of endogenous selection.

The fact that the identification strategy relies on an exclusion restriction may seem restrictive in some applications,[4] but contrary to the missing-at-random assumption, for instance, this condition is testable. Furthermore, the method can be informative even if the exclusion restriction fails, but selection depends monotonically on both the outcome and the

[3] In particular, the marginal effect of the instrument on the outcome, or the effect of the selection variable on the outcome, are identified.

[4] It is not needed in Lewbel's framework, for instance. On the other hand, the existence of a special regressor, which may be difficult to find in practice, is not needed here. The instrument may be continuous or discrete; the completeness condition only requires that their support has at least as many elements as the support of the outcome. Moreover, almost no restriction is imposed on the conditional probability of selection.

instrument. In this case, we provide sharp and finite bounds on some parameters of the outcome. Thus, even if the dependent variable is unbounded, one can obtain compact intervals on parameters of interest. This result is similar to the one of Manski and Pepper (2000, see their Proposition 2, Corollary 2) but within a slightly different framework and under other assumptions. Instead of their monotone treatment response condition, which states that outcomes increase with the treatment, the result relies on the existence of an instrument that affects selection in a monotonic way. Such a condition is weak and is likely to be satisfied in many contexts, including the use of data with nonignorable nonresponse and treatment effects estimation. In this latter case in particular, the result should be of practical importance, as it allows one to go beyond the standard routine of computing matching estimators as point estimates of these effects.

In addition to identification issues, we also consider estimation of the conditional probability of selection. Standard GMM can be used in the parametric case or in the nonparametric one with a discrete outcome. In a nonparametric setting with a continuous dependent variable, the parameter is functional and solves a linear inverse problem. We propose an estimator based on Tikhonov regularization, as Hall and Horowitz (2005) or Carrasco et al. (2006), and show its consistency. Then valid inference on the whole population is based on an inverse probability weighting procedure, in a similar fashion to Horvitz and Thompson (1952), Hellerstein and Imbens (1999), Hirano et al. (2003) or Wooldridge (2007). Finite sample properties of these estimators are investigated through Monte Carlo simulations.

Finally, the method is used to estimate the effect of grade retention in the 5th grade in France on test achievement. Aside from the usual counterfactual problem, the identification of this effect is complicated by the fact that French students only take standardized tests at the beginning of the 3rd and 6th grades. Thus, ability at the end of the 5th grade, which is one of the main factors of grade retention, is observed for promoted students, thanks to the 6th grade test, but not for retained students. Consequently, the problem fits within our framework. Sharp bounds on the effect of grade retention are computed using the 3rd grade test score as an instrument. Overall, the short-term impact of grade retention seems more likely to be positive, a result in line with that of Jacob and Lefgren (2004) for third-graders in Chicago.

The rest of the paper is structured as follows. Section 2 is devoted to identification issues. Estimation methods are described in Section 3. Monte Carlo results are presented in Section 4, and the application to grade retention is developed in Section 5. The appendix contains all proofs.

2 Identification

2.1 The setting and main result

Let D, Y and Z denote respectively the selection dummy variable, the dependent variable and the instrument. The first assumptions define the selection problem.

Assumption 1 *We observe D and (Y, Z) when $D = 1$. Y is not observed when $D = 0$.*

Assumption 2 *The distribution of Z is identified.*

Assumptions 1 and 2 are satisfied when only Y is missing, as in selection or item nonresponse problems. It also encompasses the case of unit nonresponse, in which (Y, Z) are missing when $D = 0$. In this latter situation, auxiliary information on Z is needed to satisfy Assumption 2. This information typically comes from refreshment samples, censuses or administrative data. In these two latter cases, assuming the identifiability of the whole distribution of Z may be overly strong, and we will see in Subsection 2.4 that it may be replaced by the knowledge of some moments of Z, at the price of imposing parametric restrictions.

Assumptions 1 and 2 alone are not sufficient to point identify the distribution of (D, Y, Z). More structure on the dependence between these variables is needed. If selection directly depends on Y, the usual assumption of exogenous selection fails, and it may be difficult to find an instrument that affects the selection variable but not the outcome. On the other hand, a variable Z related to Y but not to D may be available in this case. More precisely, we assume the following:[5]

Assumption 3 $D \perp\!\!\!\perp Z \,|\, Y$.

This assumption was also considered by Chen (2001), Tang et al. (2003), Hemvanich (2004) and Ramalho and Smith (2007) in a nonresponse framework. It is also a special case of Assumption (41) of Manski (1994). It can be interpreted as follows. The selection equation depends on Y, which is missing when $D = 0$ and thus cannot be identified with the data alone. On the other hand, if an instrument that affects Y but not directly D is available,

[5] We could refine this assumption by supposing that $D \perp\!\!\!\perp Z \,|\, Y, X$ where X denote covariates whose distribution is identified. The subsequent analysis would then be conditional on X. We do not introduce such covariates until Subsection 2.4 in order to ease exposition.

one can identify this selection equation in a similar way to usual instrumental regressions. For instance, suppose that (D, Y, Z) satisfy the nonparametric system

$$\begin{cases} Y &= \varphi(Z, \varepsilon) \\ D &= \psi(Y, \eta). \end{cases} \tag{2.1}$$

Then Assumption 3 holds under an independence condition, as the following result shows.

Proposition 2.1 *Suppose that system (2.1) holds with* $\eta \perp\!\!\!\perp (Z, \varepsilon)$. *Then Assumption 3 holds.*

By letting $\psi(y, \eta) = \mathbb{1}\{\eta \leq P(D = 1|Y = y)\}$, we can suppose without loss of generality that η is independent of Y.[6] The exclusion restriction amounts to reinforcing this into a conditional independence between η and (Y, Z).

As indicated previously, a dependence condition between Y and Z is required to achieve the identification of the model. We rely afterwards on a completeness condition. For any random variable T and $q > 0$, let L_T^q denote the space of real functions g satisfying $E(|g(T)|^q) < +\infty$. Let us also denote $\mathcal{B}$ the set of real functions g such that $g(Y)$ is bounded below almost surely and $g \in L_Y^1$.

Assumption 4 *Y is $\mathcal{B}$-complete for Z, i.e., for all $g \in \mathcal{B}$,*

$$\Big(E(g(Y)|Z) = 0 \quad a.s.\Big) \Longrightarrow \Big(g(Y) = 0 \quad a.s.\Big). \tag{2.2}$$

Assumption 4 is weaker than the usual completeness condition, which requires that (2.2) holds for any $g \in L_Y^1$, but stronger than bounded completeness, which is equivalent to (2.2) for bounded functions only (see, e.g., Mattner, 1993, for a discussion on the difference between completeness and bounded completeness). The standard completeness condition is used in nonparametric instrumental regression settings under additive separability (see Newey and Powell, 2003, Darolles et al., 2006) and in nonclassical measurement error problems (see Chen and Hu, 2006 and Hu and Schennach, 2008),[7] while the bounded completeness condition is used for instance by Blundell et al. (2007).

Completeness can be easily characterized when Y and Z have finite supports. Indeed, letting $(y_1, ..., y_s)$ and $(z_1, ..., z_t)$ denote these supports, this condition amounts to $\text{rank}(M) =$

[6] In this case, ψ is not necessarily structural.

[7] Indeed, Assumption 2.4 of Chen and Hu (2006) and Assumption 2 of Hu and Schennach (2008) are equivalent, under technical conditions, to a completeness condition.

s, where M is the matrix of typical element $P(Y = y_i|Z = z_j)$ (see Newey and Powell, 2003). Hence, the support of Z must be at least as rich as the one of Y ($t \geq s$), and the dependence between the two variables must be strong enough for s linearly independent conditional distributions to exist. In this case, completeness is equivalent to bounded completeness. Completeness or bounded completeness are much more difficult to characterize when the support of Y or Z is infinite and only sufficient conditions have been obtained so far. Both hold when the density of Y conditional on Z belongs to an exponential family (see Newey and Powell, 2003). Assumption 4 is also satisfied under an additive decomposition, a large support assumption and technical restrictions on ε in system (2.1), as shown in the following proposition.

Proposition 2.2 *Consider system (2.1) with $Y \in \mathbb{R}$ and suppose that:*

a) (Additive decomposition) $\varphi(Z, \varepsilon) = \mu(\nu(Z) + \varepsilon)$ and $Z \perp\!\!\!\perp \varepsilon$.

b) (Large support) The measure of $\nu(Z)$ is continuous with respect to the Lebesgue measure and the support of $\nu(Z)$ is $\mathbb{R}$ almost surely.

c) (Regularity conditions on ε) The distribution of ε admits a continuous density f_ε with respect to the Lebesgue measure. Moreover, $f_\varepsilon(0) > 0$ and there exists $\alpha > 2$ such that $t \mapsto t^\alpha f_\varepsilon(t)$ is bounded. Lastly, the characteristic function of ε does not vanish and is infinitely often differentiable in $\mathbb{R}\backslash A$ for some finite set A.

Then Y is $\mathcal{B}-$complete for Z.

The additive decomposition and the large support condition are identical to Assumptions A1 and A2 made by D'Haultfœuille (2008) to study completeness and bounded completeness.[8] The regularity conditions on ε are satisfied for many distributions such as the normal ones, the student distributions with degrees of freedom greater than one[9] and the stable distributions with characteristic exponent greater than one. Interestingly, these regularity conditions are hardly stronger than the one needed to achieve bounded completeness, namely, the zero freeness of the characteristic function of ε (see D'Haultfœuille, 2008,

[8]The additive decomposition considered here encompasses many nonlinear models, beyond the nonparametric additive models for which $\mu(x) = x$. Usual ordered choice models correspond to $\mu(x) = \sum_{k=1}^{K} k\mathbb{1}_{]\alpha_{k-1};\alpha_k]}(x)$ (where $\mathbb{1}_A(x) = 1$ if $x \in A$, 0 otherwise) for some given thresholds $\alpha_0 = -\infty < \alpha_1 < ... < \alpha_K = +\infty$. Simple Tobit models correspond to $\mu(x) = \max(0, x)$. Duration models like the accelerated failure time model (for which $\mu(x) = \exp(x)$) or the proportional hazard model (for which μ is an unknown increasing function and $-\varepsilon$ is distributed according to a Gompertz distribution) also fit in this framework.

[9]See, e.g., Mattner (1992) for a proof that the conditions on the characteristic function of student distributions are satisfied.

Theorem 2.1). Hence, in this framework at least, $\mathcal{B}$-completeness appears to be almost equivalent to bounded completeness.

Because identification is based on inverse probability-weighted moment conditions, we make the following assumption.

Assumption 5 *$P(Y) > 0$ almost surely.*

This assumption is similar to the common support condition in the treatment effect literature. It does not hold if D is a deterministic function of Y, as in simple truncation models where $D = \mathbb{1}\{Y \geq y_0\}$, y_0 denoting a fixed threshold. It also fails for random truncation models of the form $D = \mathbb{1}\{Y \geq \eta\}$ if η is strictly greater than the infimum of Y. In Example 2 below, this would be the case if the reservation wage η of individuals were always greater than the lowest potential wage Y.

Theorem 2.3 *If Assumptions 1-5 hold, then the distribution of (D, Y, Z) is identified.*

Basically, the idea of the proof is the following. Under Assumption 3 and 4, the equation in Q

$$E\left(\frac{D}{Q(Y)}\middle| Z\right) = 1 \tag{2.3}$$

has a unique solution, P. Identification of P follows because the left-hand side is identified for any given Q. Once P is known, it is easy to show that the distribution of (D, Y, Z) is identified. We now present several potential applications of this framework.

Example 1: nonignorable nonresponse

In this case, the outcome Y is observed only if the individual responds to the survey or a given question in the questionnaire ($D = 1$). One aims at recovering the full distribution of Y, given that nonresponse directly depends on Y. For instance, accepting to respond to the question, "Have you taken drugs at least once during the last month?" is likely to depend on the answer Y (1=Yes, 0=No) itself. The method can be applied if an instrument affects Y but not directly D. In the drugs example, local drug prices affect drug use but are unlikely to directly influence survey response. Note that in this example where Y is binary, the completeness condition is easy to check, since it is equivalent to a nonzero correlation between Y and the instrument.

Example 2: Roy model with an unobserved sector

In this example, let Y denote the wage an individual can obtain in sector 1, and η be the corresponding wage in sector 0. The individual chooses the sector that offers him the higher wage. Y is observed if sector 1 is chosen but η is never observed. Thus, in this case $D = \mathbb{1}\{Y \geq \eta\}$.[10] For instance, Y may represent the potential wage of an individual, which is observed only if the person enters the labor market, while η denotes his reservation wage. The aim is to recover the distribution of Y, or the effect of covariates X on Y. The usual exclusion restriction requires the existence of a variable that affects η but not Y. On the other hand, the strategy above can be applied if an instrument Z, which affects the potential wage but not directly the reservation wage, is available, so that η is independent of Z conditional on Y (or conditional on (X, Y) if one adds covariates). A possible example of such an instrument is the local unemployment rate (see Haurin and Sridhar, 2003, for evidence that the local unemployment rate does not affect the reservation wage).[11]

Example 3: Sample from one response stratum

In this example, a researcher seeks to study the effect of Y on a binary variable D but observes Y only for the stratum $D = 1$.[12] Our instrumental strategy relies on the existence of an instrument Z that affects Y but not D directly and whose distribution is identified. Suppose for instance that one wants to study the efficiency of vaccination in a developing country, but only data on ill people are available, and the vaccination rate in the population is unknown. In this case, D is the dummy variable of being ill, while Y is the dummy of being vaccinated. If there has been an important vaccination campaign after a given date, one can use the dummy of being born after this date as an instrument.[13] Once more, the completeness condition is satisfied as soon as the correlation between Y and the instrument is not zero.

This example also includes truncated count data models. In this case, the aim is to recover the effect of Y on an integer-valued variable N, given that Y is observed only when $N > 0$.[14]

[10] Following the previous discussion, Assumption 5 will be satisfied if η can be lower than any value of Y, with a positive probability.

[11] No statistical test for completeness conditions has been developed yet in the case where Y is continuous. Thus, Assumption 4 has to be maintained in this example. However, one can test the implications of Assumption 4 by checking, for instance, that $E(Y|Z)$ is not a constant function.

[12] In this case, Y is a covariate rather than an outcome. The notation Y is maintained for notational consistency with Assumption 1.

[13] If the risk of disease is related to age as well, one can use only the individuals born just before and just after the beginning of the campaign, as in regression discontinuity designs.

[14] Hence, $D = \mathbb{1}\{N > 0\}$ here and recovering $P(N = k|Y)$ for all $k \in \mathbb{N}$ amounts to identifying

Consider for instance the estimation of the price elasticity of a good through the use of retail data.[15] If we observe the quantities sold N and the sales $N \times Y$, but not directly the prices Y, then these prices can be deduced only when the quantities sold are positive. The framework can be applied if there is an instrument that affects the prices but not directly the demand, and whose distribution is identified. Production cost shifters such as the prices of the inputs may be good candidates.

2.2 Testability

In some contexts, the conditional independence assumption may seem overly strong. An interesting feature of this assumption, yet, is that it is refutable, contrary to the usual missing-at-random assumption. First, equation (2.3) may have no solution. This is especially clear when Y and Z have finite supports. If, indeed, Y and Z take respectively s and t distinct values, with $t > s$, then (2.3) can be written as a system of t equations with s unknown parameters, so that the model is overidentified.

But even if $s = t$, the model is testable since the solution Q of equation (2.3) has to be a positive probability, i.e., $Q(y) \in (0, 1]$ for all y.[16] Consider the illustrative case where $(Y, Z) \in \{0, 1\}^2$. Let $p(y, z) = P(D = 1, Y = y|Z = z)$, $\alpha = 1/Q(0)$ and $\beta = 1/Q(1)$. Then, as soon as $p(0, 0)p(1, 1) \neq p(0, 1)p(1, 0)$, that is to say, under the completeness condition, equation (2.3) is equivalent to

$$\begin{aligned} \alpha &= \frac{p(1,1) - p(1,0)}{p(0,0)p(1,1) - p(0,1)p(1,0)}, \\ \beta &= \frac{p(0,0) - p(0,1)}{p(0,0)p(1,1) - p(0,1)p(1,0)}. \end{aligned}$$

Hence, when $p(1, 1) - p(1, 0)$ and $p(0, 0) - p(0, 1)$ have opposite signs, for instance, Assumption 3 is rejected. Basically, this happens when $z \mapsto P(D = 1|Y = y, Z = z)$ varies too much compared to $z \mapsto P(Y = y|Z = z)$.

Now, when a solution $Q \in (0, 1]$ of equation (2.3) does exist, one can expect that Assumption 3 cannot be rejected, since intuitively, this equation makes use of all the available

$P(D = 1|Y)$. Note that this example differs from the simple truncation model $D = \mathbb{1}\{Y \geq s\}$ described above. In particular, Assumption 5 will hold as long as $P(N = 0|Y) < 1$ almost surely.

[15] As discussed by Grogger and Carson (1991), truncated counts arise more generally with data from surveys that ask participants about their number of participations, or administrative records in which inclusion in the database depends on having engaged in the activity of interest.

[16] If the completeness condition does not hold, Q may not be unique. Then at least one of the solutions has to belong to $(0, 1]$.

information. Theorem 2.4 formalizes this idea.

Theorem 2.4 *Suppose that Assumptions 1, 2 and 5 hold. Then Assumption 3 can be rejected if and only if there exists no solution to equation (2.3) that belongs to* $(0, 1]$.

When Y is discrete and takes values in $\{y_1, ..., y_s\}$, a statistical test of Assumption 3, under the maintained assumption of completeness, amounts to testing the multiple inequality constraints $f(y_j) \geq 1$ for $j = 1, ..., s$, with $f = 1/P$. Such a test can be implemented with the GMM estimator of P presented in Subsection 3.1 below.[17] The situation is more involved when Y is continuous. Under Assumptions 1-5 and additional technical conditions, a consistent nonparametric estimator $\widehat{f}$ of f is developed in Subsection 3.2. This estimator is constrained to belong to $[1, M]$ with $M > 1$. It should be possible to develop a consistent, unconstrained estimator $\widetilde{f}$ of f. Then a test of Assumption 3 could be based on the distance between $\widehat{f}$ and $\widetilde{f}$ since, under Assumption 3, $\widetilde{f}(y)$ should be greater than one for most values of y, and the distance between the two should therefore be close to zero.[18]

2.3 Set identification without conditional independence

A second interesting feature of equation (2.3) is that it provides an informative bound on some parameters of interest under monotonicity conditions, which are far weaker than the conditional independence condition of Assumption 3. Because monotonicity conditions are only meaningful in ordered sets, we focus on the case where $(Y, Z) \in \mathbb{R}^2$. Let $\widetilde{Z}$ denote a variable that may differ from Z and whose distribution is also identified. Assumption 3 is replaced by the following ones.

Assumption 3' $z \mapsto P(D = 1|Y, Z = z)$ *is increasing almost surely.*

Assumption 6 $y \mapsto P(D = 1|Y = y, \widetilde{Z})$ *is increasing almost surely.*

Assumption 3' weakens the conditional independence between the selection variable and the instrument set in Assumption 3 to a monotone dependence. It is also a variant of the usual instrumental condition which assumes that the instrument affects the probability of selection but is independent of the outcome. Here, the effect on the probability of

[17] See, e.g., Gouriéroux and Monfort (1995, Section 21.4) for details on multiple inequality tests.

[18] The critical region of such a test would depend on the asymptotic distribution of $(\widehat{f}, \widetilde{f})$, whose derivation is beyond the scope of the paper.

selection is restricted to be monotonic, but no independence condition between Y and Z is needed. Assumption 6 weakens the missing-at-random assumption to a monotone dependence between the selection variable and the outcome.

Theorem 2.5 below provides bounds on parameters of the form $E(h(Y))$ for $h \in H_Y$ or $h \in H_{YZ}$, with

$$\begin{aligned} H_T &= \{h \in L_T^1 \text{ and } h \text{ is increasing}\} \ (T = Y \text{ or } Z), \\ H_{YZ} &= \{h \in L_Y^1 / \exists \widetilde{h} \in H_Z / h(Y) = E(\widetilde{h}(Z)|D = 1, Y)\}. \end{aligned}$$

The set H_Y includes, among others, functions of the form $h(y) = \lambda y$ with $\lambda > 0$ and indicator functions $h_u(y) = \mathbb{1}\{y \geq u\}$, so that parameters of the form $E(h(Y)), h \in H_Y$, include the survival function of Y taken at any point. The set H_{YZ} is more abstract. In an informal way, H_{YZ} increases as the dependence between Y and Z becomes stronger. As a simple illustration, this set only includes constant functions when Y and Z are independent (conditional on $D = 1$) but is equal to H_Y when $Y = Z$. More formally, H_{YZ} is a subset of the range of the conditional expectation operator $g \mapsto (y \mapsto E(g(Z)|D = 1, Y = y))$, which itself is linked to the null space of this operator. When (Y, Z) has finite support, the dimension of the range increases as the dimension of the null space decreases. Thus, at least in finite dimensions, H_{YZ} will be maximal if the conditional expectation operator is injective, that is to say under a completeness condition on Y and Z.

It seems difficult to test formally that $h \in H_{YZ}$ for a given, increasing, function h. On the other hand, we can test the stronger condition:

$$E(Z|D = 1, Y) = \alpha + \beta h(Y), \quad \beta > 0. \tag{2.4}$$

Tests of such functional forms are described for instance by Yatchew (1998, Subsection 4.2).

We assume afterwards that equation (2.3) has at least one solution.[19] If the constant function $P(D = 1)$ is a solution, we let $Q(Y) = P(D = 1)$; otherwise, Q is any of the solutions.

Theorem 2.5 *Suppose that $P(D = 1) > 0$ and Assumptions 1 and 2 hold for Z and $\widetilde{Z}$. Then:*
a) Under Assumption 6, $E[h(Y)] \leq E\left[E(h(Y)|\widetilde{Z}, D = 1)\right]$ for all $h \in H_Y$. Moreover, this upper bound is sharp.

[19] On the other hand, the solution need not be unique, so that the completeness condition is not required here.

b) Under Assumptions 3', $E\left[Dh(Y)/Q(Y)\right] \leq E\left[h(Y)\right]$ for all functions $h \in H_{YZ}$. Moreover, this lower bound is sharp provided that at least one solution Q belongs to $(0;\, 1]$.
c) For all functions $h \in L^1_Y$, these three expectations are equal when $D \perp\!\!\!\perp (Y, Z, \widetilde{Z})$ or when $Z = \widetilde{Z} = Y$.

Part a) of Theorem 2.5 is not specific to the methodology developed here and is rather intuitive. Part b), on the other hand, entails that the moment condition used here leads to a sharp lower bound on $E[h(Y)]$. This lower bound does not depend on the choice of Q so that no completeness condition is required. The bound also holds even if no solution Q belongs to $(0;\, 1]$. In this case however, the bound may not be sharp because one could exploit the fact that the conditional independence assumption is rejected by the data.

An important consequence of Theorem 2.5 is that for all functions $h \in H_Y \cap H_{YZ}$, we can obtain a bounded identification interval for $E(h(Y))$. This occurs even if $h(Y)$ is unbounded. In this sense, the result is similar to Proposition 2, Corollary 2 of Manski and Pepper (2000), under a different set of assumptions. In particular, we do not rely on the monotone treatment response condition, which is difficult to adapt to selection models or nonresponse problems. Moreover, the monotone treatment response assumption can be strong in the context of treatment effects. In the Roy model with an unobserved sector developed in Example 2, this assumption implies that $Y \geq \eta$ (or $\eta \geq Y$) almost surely, so that in equilibrium only one sector is chosen, which seems a rather unrealistic situation. Assumption 3' is different in that it assumes that the probability of selection increases with the instrument. This assumption is rather weak and should be satisfied in many contexts including treatment effects estimation, or estimation of parameters under nonignorable missing data. In Example 2 above, one may use standard instruments such as non-wage income or the number of children.

As part c) shows, the interval is reduced to a point if D is fully missing at random. Hence, the length of the interval can be interpreted as a measure of the severity of the selection problem. Because the interval is also reduced to a point when $Z = \widetilde{Z} = Y$, its length also reflects the quality of the chosen instruments. As the dependence between $(Z, \widetilde{Z})$ and Y increases, parameters of the distribution of Y can be better predicted from the distribution of the instruments. Moreover, the upper (resp. lower) inequality turns into an equality whenever $Y \perp\!\!\!\perp D|\widetilde{Z}$ (resp. $Z \perp\!\!\!\perp D|Y$). Hence, Z and $\widetilde{Z}$ must be chosen differently since $\widetilde{Z}$ intends to reduce selection on unobservables correlated with the outcome, whereas Z should be as independent of the selection variable as possible, given Y.

As noted before, H_{YZ} increases as the dependence between Y and Z becomes stronger. Hence, the choice of the instrument also matters for the range of applications of the lower bound. If it seems difficult, without further restrictions, to describe the set $H_Y \cap H_{YZ}$ of functions h such that an interval can be built on $E[h(Y)]$, this set will contain at least all functions $h(y) = \lambda y$ with $\lambda > 0$ under the testable linear condition that $E(Z|D = 1, Y) = \alpha + \beta Y$ (with $\beta > 0$). In this case in particular, $E[Y]$ can be bounded below and above. If Y and Z exhibit a positive dependence, the following proposition states that the set $H_Y \cap H_{YZ}$ will be equal to H_{YZ}.

Proposition 2.6 *Suppose that for all z, $y \mapsto F_{Z|Y=y,D=1}(z)$ is decreasing. Then $H_{YZ} \subset H_Y$.*

2.4 Parametric identification

Nonparametric identification relies on the unicity of a functional equation. However, one may be reluctant to use nonparametric estimators in practice, because of the curse of dimensionality for instance. Furthermore, Assumption 2 may be too strong in some circumstances. Suppose for instance that the instrument is observed only when $D = 1$ (as in the cases of unit nonresponse or panel attrition) but that some auxiliary information on this instrument is available. This auxiliary information may not be sufficient to identify the full distribution of Z. If Z is multivariate and its different components are observed in different sources that cannot be matched, only the marginal distributions are identified. If the instrument is measured with a zero mean error in these auxiliary data, only $E(Z)$ can be recovered. When Assumption 2 fails to hold but $E(Z)$ is known, Theorem 2.7 below shows that identification can still be obtained, under parametric restrictions on P. It generalizes a result of Nevo (2002) to the case where $Y \neq Z$. The ideas behind are also very similar to the method of generalized calibration developed by Deville (2002) in a survey sampling framework to handle nonignorable nonresponse with instruments. Deville (2002), however, does not address the issue of identification of P.

Since we consider a parametric framework, we explicitly add covariates X. In the following, we suppose that $V = (X', Y')' \in \mathbb{R}^p$ and $W = (X', Z')' \in \mathbb{R}^q$. The identification result is based on the following assumptions.

Assumption 2' *a) $E(W)$ is known. b) $P(D = 1|V) = F(V'\beta_0)$ where F is a known, differentiable and strictly increasing function from $\mathbb{R}$ to $(0, 1)$. c) For all $\delta \in \mathbb{R}^p$, $P(V'\delta = 0|D = 1) = 1$ implies that $\delta = 0$.*

Assumption 3' $D \perp\!\!\!\perp Z|V$.

Assumption 4' $rank(E(DWV'F'(V'\beta_0)/F^2(V'\beta_0))) = p$.

Assumption 4" $E(Z|D=1,V) = \Gamma_1 X + \Gamma_2 Y$ *where* Γ_2 *is full rank.*

Assumption 2' weakens Assumption 2 on data availability, at the price of imposing a parametric restriction on P. Such a restriction is satisfied for instance if the selection equation is a logit or probit model. Like Assumption 4 in the nonparametric setting, Assumption 4' is the rank condition. As usual, this condition implies that $q \geq p$. Finally, Assumption 4" is a special case of Assumption 4', which restricts the regression of Z on V to be linear.

Theorem 2.7 *If Assumptions 1, 2' and 3' are satisfied, then:*
a) β_0 *is locally identified if and only if Assumption 4' holds.*
b) If Assumption 4" holds, β_0 *is globally identified.*

Local identification is obtained under a condition very similar to the rank condition in linear instrumental regressions. Theorem 2.7 also provides a sufficient and testable condition that ensures the global identification of β_0.

3 Estimation

We now turn to the parametric and nonparametric estimation of P. The first assumption describes the sampling process.

Assumption 7 *We observe a sample* $((D_1, X_1, Y_1^*, Z_1), ..., (D_n, X_n, Y_n^*, Z_n))$ *of independent copies of* (D, X, Y^*, Z)*, with* $Y^* = DY$.

The i.i.d. assumption is standard but can be weakened without affecting the consistency or the rate of convergence of the estimators. We also suppose, for the sake of simplicity, that Z is observed for both $D = 1$ and $D = 0$.

3.1 Parametric estimation

It follows from Theorem 2.3 that when Y has a finite support $\{y_1, ..., y_s\}$, the equation

$$E\left(\frac{D}{\sum_{k=1}^{s} P(y_k)1\{Y = y_k\}} - 1 \middle| Z\right) = 0$$

provides the identification of the parameters $(P(y_k))_{1\leq k\leq s}$ if Assumptions 3, 4 and 5 hold. Hence, consistent and asymptotically normal estimators can be obtained by GMM. Similarly, if P satisfies the restrictions of Assumption 2', then

$$E\left[\left(\frac{D}{F(V'\beta_0)}-1\right)W\right]=0. \tag{3.1}$$

Moreover, the proof of Theorem 2.7 (see equation (6.11)) ensures that under Assumption 4", β_0 is globally identified by (3.1). Thus GMM estimators can also be used in this context.

3.2 Nonparametric estimation

If Y has continuous components and one is reluctant to rely on parametric restrictions on P, then the situation is more involved because a function, and not only parameters, must be estimated by conditional moment conditions. The same issue arises in nonparametric instrumental regression (see, e.g., Newey and Powell, 2003, Hall and Horowitz, 2005, Darolles et al., 2006 and Horowitz and Lee, 2007). For the sake of simplicity, we assume that there is no covariate X and that $(Y,Z)\in[0,1]^2$. Moreover, we only prove consistency, since the paper is mainly focused on identification. The rate of convergence could be obtained by adapting the arguments of Hall and Horowitz (2005).

Let us denote $f=1/P$ and let T be the linear operator defined by

$$T\,\phi(z)=E(D\phi(Y^*)|Z=z).$$

Then equation (2.3) may be written as

$$T\,f=1.$$

We rely on this equation for estimating f. Because the problem is ill-posed, regularization is needed to ensure consistency of the estimator. We adopt here a Tikhonov regularization, as Hall and Horowitz (2005), Darolles et al. (2006) and Horowitz and Lee (2007). First, let us consider the kernel estimator of T:

$$\widehat{T}\,\phi(z)=\frac{\sum_{i=1}^n D_i\phi(Y_i^*)K_{h_n}(z-Z_i)}{\sum_{i=1}^n K_{h_n}(z-Z_i)}.$$

For any $1<M<\infty$, let us define D_M as the subset of real measurable functions ϕ defined on $[0,1]$ and such that $M\geq\phi(Y)\geq 1$ almost surely. For any square integrable function ϕ defined on $[0,1]$, let $||\phi||^2=\int_0^1\phi(u)^2du$. Our estimator of f satisfies

$$\widehat{f}\in\arg\min_{\phi\in D_M}\left\|\widehat{T}\,\phi-1\right\|^2+\alpha_n\left\|\phi\right\|^2,$$

where α_n is a regularization parameter which prevents the solution from being unstable (see, e.g., Carrasco et al., 2006, for a discussion on regularization in ill-posed inverse problems). Under the assumptions below, the minimization problem always admits a solution, but it may not be unique (see Bissantz et al., 2004). In this case, $\widehat{f}$ can be any solution. The consistency result relies on the following assumptions.

Assumption 8 *a)* $f \in D_M$. *b) The distribution of* (Y, Z) *is continuous with respect to the Lebesgue measure and the marginal densities* f_Y *and* f_Z *satisfy* $\sup_{y\in[0,1]} f_Y(y) < +\infty$ *and* $\inf_{z\in[0,1]} f_Z(z) > 0$.

Assumption 9 *For all* $h > 0$ *and* $u \in \mathbb{R}$, $K_h(u) = K_1(u/h)$ *where* K_1 *is positive,* $\int K_1(u)du = 1$ *and* $\int uK_1(u)du = 0$.

Assumption 10 $\alpha_n \to 0$, $h_n^2 + 1/(nh_n) \to 0$ *and* $(h_n^2 + 1/(nh_n))/\alpha_n \to 0$.

Assumption 8-a) strengthens Assumption 5. Assumption 9 is weak and standard in non-parametric estimation. Assumption 10, which is identical to Assumption 3 of Horowitz and Lee (2007), is also standard. It implies that the bandwidth h_n tends to zero at a slower rate than $1/n$, and that the regularization parameter α_n tends to zero at a slower rate than h_n^2.[20]

Theorem 3.1 *Under Assumptions 3-4 and 7-10,*

$$\lim_{n\to\infty} E\left(\left\|\widehat{f} - f\right\|^2\right) = 0.$$

Theorem 3.1 implies that $\left\|\widehat{f} - f\right\|^2$ converges in probability to zero. Inverse probability weighting procedures can now be used to estimate parameters on the whole population. Let $\widehat{f}^{-i}$ denote the estimator of f obtained with the sample $(D_j, Y_j^*, Z_j)_{j\neq i}$. For any $g \in L^2_{Y,Z}$ and $\theta = E(g(Y, Z))$, define

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} D_i \widehat{f}^{-i}(Y_i^*) g(Y_i^*, Z_i).$$

Corollary 3.2 ensures that $\widehat{\theta}$ is consistent.

Corollary 3.2 *Suppose that Assumptions 3, 4 and 7-10 hold. Then*

$$\lim_{n\to\infty} E\left(|\widehat{\theta} - \theta|\right) = 0.$$

[20] We suppose here that α_n is a deterministic sequence. See, e.g., Gagliardini and Scaillet (2006) for a data-driven selection procedure.

4 Monte Carlo simulations

In this section, we investigate the finite sample properties of parametric and nonparametric estimators of P and inverse probability-weighted estimators of $E[g(Y)]$. Let us consider the following model:

$$\left\{ \begin{array}{rcl} Y & = & \Lambda(\Lambda^{-1}(Z) + \varepsilon) \\ D & = & \mathbb{1}\{P(Y) \geq \eta\}, \end{array} \right. \tag{4.1}$$

where $\Lambda(x) = 1/(1 + \exp(-x))$ is the logistic cumulative distribution function, $P(y) = 1 - 0.6/(1 + 19y^2)$, (Z, ε, η) are mutually independent, $Z \sim U[0,1]$, $\varepsilon \sim N(0,1)$ and $\eta \sim U[0,1]$.[21] The function P was chosen to match the estimate of P in the application (see Figure 2 below). Within this framework, Assumptions 3 and 4 are satisfied by Propositions 2.1 and 2.2. One can also show that Assumption 8 holds. In particular, $f(y) = 1/P(y) \leq 2.5$. Finally, $P(D = 1) \simeq 0.8$, so that approximately 20% of the Y's are missing.

Estimator	Statistic	$n = 100$	$n = 200$	$n = 500$	$n = 1,000$
$\widehat{f}_1$	MISE	0.1978	0.1532	0.1058	0.0791
$\widehat{f}_2$		0.2010	0.1478	0.1017	0.0758
$\widehat{f}_3$		1.9343	0.3697	0.0673	0.0286
$\widehat{\theta}_1$	RMSE	0.0330	0.0252	0.0155	0.0120
	(bias)	(-0.0081)	(-0.0066)	(-0.0061)	(-0.0058)
$\widehat{\theta}_2$		0.0373	0.0275	0.0174	0.0139
		(-0.0191)	(-0.0130)	(-0.0099)	(-0.0091)
$\widehat{\theta}_3$		0.0316	0.0238	0.0140	0.0104
		(0.0001)	(-0.0005)	(-0.0002)	(-0.0001)
$\widehat{\theta}_4$	RMSE	0.0338	0.0253	0.0154	0.0112

The results are obtained with 1,000 simulations for each sample size. The bias of $\widehat{\theta}_4$ is not indicated as this estimator is unbiased.

Table 1: Performances of the parametric and nonparametric estimators.

We consider two nonparametric estimators $\widehat{f}_1$ and $\widehat{f}_2$ of f that share the same regularization parameter, $\alpha_n = 0.05 \times n^{-1/5}$ but have different bandwidths, namely $h_{1n} = 0.03 \times n^{-1/5}$ and $h_{2n} = 0.02 \times n^{-1/5}$. We also consider a parametric estimator $\widehat{f}_3$ that belongs to the

[21] The model amounts to assuming a linear dependence between $\Lambda^{-1}(Y)$ and $\Lambda^{-1}(Z)$. We work with (Y, Z) rather than $(\Lambda^{-1}(Y), \Lambda^{-1}(Z))$ to be consistent with the previous assumption that $(Y, Z) \in [0,1]^2$.

following flexible parametric family:

$$f(y;\beta) = 1 + \exp\left(-\beta_0 - \sum_{k=1}^{4} y\mathbb{1}\{y \geq a_k\}\beta_k\right), \tag{4.2}$$

where $\beta = (\beta_0, ..., \beta_4)$, $a_1 = -\infty$ and (a_2, a_3, a_4) are the estimated quartiles of the distribution of Y conditional on $D = 1$. β is estimated by GMM, using the instrumental variables 1 and $(Z\mathbb{1}\{Z \geq c_i\})_{1\leq i\leq k}$, where $c_1 = -\infty$ and the (c_2, c_3, c_4) are the estimated quartiles of Z. We measure the accuracy of the three estimators of f by the usual mean integrated square error (MISE):

$$\text{MISE}(\widehat{f}) = E\left(\int_0^1 (\widehat{f}(u) - f(u))^2 du\right).$$

We also consider inverse probability-weighted estimators of $\theta = E(Y) = 1/2$. Let us define, for $k \in \{1, 2, 3\}$,

$$\widehat{\theta}_k = \frac{1}{n}\sum_{k=1}^{n} D_i \widehat{f}_k(Y_i^*) Y_i^*.$$

We also compute the infeasible estimator

$$\widehat{\theta}_4 = \frac{1}{n}\sum_{k=1}^{n} D_i f(Y_i^*) Y_i^*.$$

The accuracy of each estimator is described through its bias and root mean square error (RMSE). Results are displayed in Table 1. On average, $\widehat{f}_2$ outperforms $\widehat{f}_1$ and also the parametric estimator $\widehat{f}_3$ for small sample sizes. $\widehat{f}_3$ is, indeed, somewhat erratic for small n but is far more accurate than the nonparametric estimators for larger n. It seems, in this design, that the bias due to the parametric misspecification is negligible compared to the accuracy gains obtained by the parametric procedure. The corresponding estimator $\widehat{\theta}_3$ is also the most precise one, even in small samples. $\widehat{\theta}_1$ outperforms $\widehat{\theta}_2$, confirming the idea that a better first-step nonparametric estimator does not necessarily result in a better second-step estimator. Finally, $\widehat{\theta}_4$ is less accurate than $\widehat{\theta}_3$ and is comparable with $\widehat{\theta}_1$. Estimating f in a first step may actually yield a lower asymptotic variance than the one of $\widehat{\theta}_4$, similarly to what happens in the estimation of average treatment effects using the propensity score.[22]

[22]Indeed, Hirano et al. (2003) show that an inverse probability-weighted estimator of average treatment effects based on an estimated propensity score is asymptotically efficient, whereas the estimator based on the true propensity score may not.

5 Application

5.1 Introduction

In this section, we apply the strategy developed above to estimate bounds on the short-term effects of grade retention among 5th-grade students in France. Whereas most countries have almost completely given up grade retention as an educational policy,[23] the level of grade retention in France is still high. In 2002, for instance, a quarter of students had repeated grades at least once in primary school (see Troncin, 2004). Nonetheless, there has been no serious attempt to measure the impact of grade retention on student achievement in the French educational system.[24]

The study is based on a panel of the French Ministry of Education, which follows 9,641 children who entered the first grade of primary school in 1997. Data consist of schooling trajectories and standardized test scores at the beginning of the 3rd grade (variable Z) and the 6th grade (variable Y for the 2002 test and Y_1 for the 2003 test).[25] As the 6th grade test scores are reported only for pupils who reached this grade in 2002 or in 2003, the initial sample contains 7,175 students who were in the 5th grade in 2001 and in the 6th grade in 2002 or 2003.[26] 23.8% of these data are excluded because of missing test scores in the 3rd or 6th grades. The final sample consists of 5,467 children. Among them, 2.2% were retained in the 5th grade ($D = 0$), 6.7% in the 6th grade ($D = 1$ and $D_1 = 0$) while the others did not repeat ($D = 1$ and $D_1 = 1$). Table 2 displays the average test scores on

[23] A notable exception is the United States where several states have reintroduced this policy by linking promotion to state or district assessment (see Jacob and Lefgren, 2004).

[24] Troncin (2005) measures the effect of grade retention on achievement in the first grade of primary school using a propensity score matching approach, but he relies on data from only one school. Cosnefroy and Rocher (2004) study the effect of grade retention on achievement in the 3rd grade with the same data as here, using a linear regression approach. In other countries, the effects of grade retention are controversial. Jacob and Lefgren (2004) put forward the possibility for disadvantaged children to catch up, and Jacob (2005) shows that the threat of grade retention encourages all students to increase their efforts. On the other hand, many educational and sociological studies underline the harmful effects of grade retention on the motivation of children (see, e.g., Crahaye, 1996), drop-out rate (see Jimerson et al., 2002) and even academic performance (see, e.g., the meta-analyses of Holmes, 1989, or Jimerson, 2001). Nevertheless, most of these studies rely on very few controls (see Lorence, 2006), so that they probably underestimate the true effect of grade retention on achievement.

[25] Tests corresponding to a given grade differ partly from year to year. The scores considered here are built using only common items. The three scores are also standardized on the final sample.

[26] Other situations correspond to missing data on the trajectories, grade-advanced pupils, pupils retained before the 5th grade and students in special classrooms.

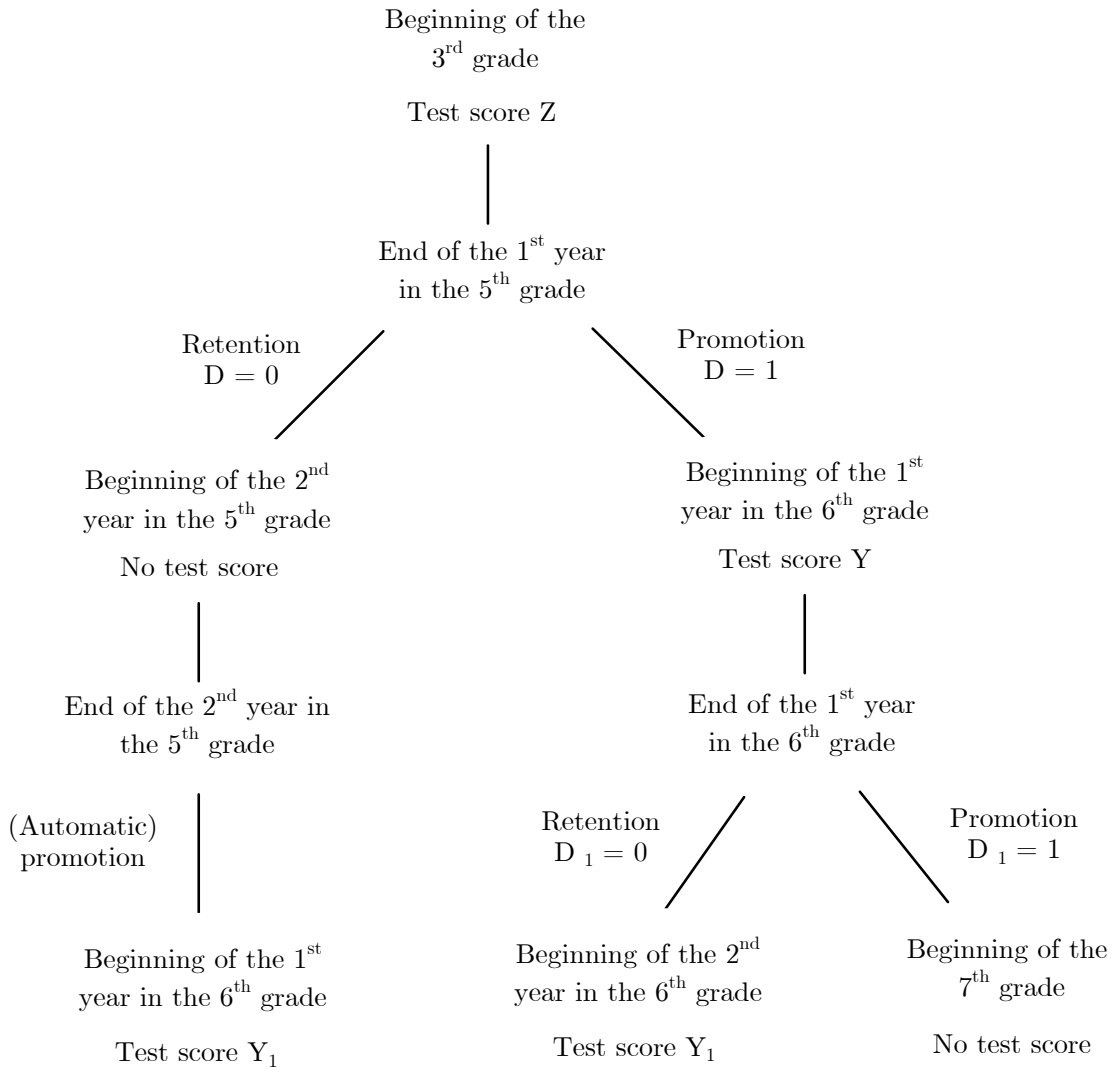

Figure 1: Promotion, retention and available test scores.

this sample. The 6th grade test score in 2002 is missing for children retained in the 5th grade because they only entered this grade in 2003. Similarly, in 2003 the 6th grade test score is not observed for children who did not repeat since they were in the 7th grade that year.

As expected, the differences between retained and promoted pupils in terms of test achievement are large. On average, the 3rd grade test scores of 5th- (resp. 6th-) grade repeaters are more than 1.5 (resp. more than 1) standard deviations below those of students who did not repeat. The table also displays the progression of students retained in the 6th grade during their first year in this grade. This progression is available because these students take the test twice, at the beginning of their first and second year in the 6th grade (see

Figure 1). This feature will be useful in the following.

	Retained in the 5th grade ($D=0$)	Retained in the 6th grade ($D=1, D_1=0$)	Promoted in both grades ($D=1, D_1=1$)
Number of observations	120	365	4982
3rd grade test score Z	-1.48 (0.91)	-1.02 (0.90)	0.11 (0.94)
2002 6th grade test score Y	-	-1.32 (0.81)	0.12 (0.93)
2003 6th grade test score Y_1	-0.90 (0.87)	-0.64 (0.79)	-

Table 2: Summary statistics.

We focus here on the average effect of retention in the 5th grade on test score achievement one year later. Let $Y_1(1)$ (resp. $Y_1(0)$) denote the 6th grade test score a student would have obtained in 2003 if he had been promoted to the 6th grade (resp. retained in the 5th grade). The parameter of interest is defined as

$$\Delta^{TT} = E(Y_1(0) - Y_1(1)|D=0). \tag{5.1}$$

When $D=0$, $Y_1(0)$ is observed by Y_1, but $Y_1(1)$ is unobserved. It is difficult to rely on an instrumental strategy to overcome this counterfactual issue because there is no exogenous rule driving grade retention decisions in France.[27] Therefore, we assume that the progression of the retained students had they been promoted in the 6th grade can be bounded in the following way:

$$0 \leq E(Y_1(1) - Y|D=0, Y) \leq E(Y_1(1) - Y|D=1, D_1=0, Y). \tag{5.2}$$

The lower bound simply states that, on average, retained students would not have done worse than what they did one year before, had they been promoted. The upper bound states that, on average, their progression would have been smaller than that of students with the same initial test score who were promoted in the 6th grade and retained one year later. The idea behind is that, on average, teachers do not make mistakes by retaining pupils who would have benefited more from the 6th grade than some of the promoted students. The two bounds somewhat represent two extreme situations. The lower bound corresponds to perfect decisions of retention, in that retained students would not have

[27] As an evidence of the discretionary nature of grade retention in France, an order of the Minister of Education in 2005 asserts that grade retention should be taken by teachers after discussion with parents, according to the student's ability and his progression during the year.

benefited at all from being promoted. The upper bound corresponds to a fully randomized choice among students who would have equally benefited from being promoted.

Under condition (5.2), we get

$$E(Y_1|D=0) - E\left[h(Y)|D=0\right] \leq \Delta^{TT} \leq E(Y_1|D=0) - E(Y|D=0), \qquad (5.3)$$

where $h(Y) = E(Y_1(1)|D=1, D_1=0, Y)$. Students retained in the 6th grade take the standardized test twice, so we observe both Y and $Y_1(1)$ for them ($Y_1(1) = Y_1$ in this case), and h is identified. On the other hand, Y is unobserved for students retained in the 5th grade. Hence, $E[h(Y)|D=0]$ and $E(Y|D=0)$ are not identified without further restrictions. However, we can use the method developed above to point or set identify them. First, Y, the main determinant of D, is unobserved when $D=0$. Second, the 3rd grade test score Z is observed for both values of D and is correlated with Y. We now discuss two possible strategies, based respectively on the independence assumption $D \perp\!\!\!\perp Z|Y$ and the monotonicity conditions considered in Subsection 2.3.

5.2 Empirical strategies

First strategy: conditional independence

First, let us suppose that grade retention in the 5th grade is independent of the 3rd grade test score conditional on Y, i.e., a model of the form:

$$\begin{cases} Y &= \varphi(Z, \varepsilon) \\ D &= \psi(Y, \eta), \end{cases}$$

where $\eta \perp\!\!\!\perp (Z, \varepsilon)$. The completeness condition is also assumed to hold. Informally, both will be satisfied if the 3rd grade score affects ability at the end of the 5th grade, measured by Y, but not directly grade retention. Under these assumptions, Theorem 2.3 can be applied and we can identify $E(h(Y)|D=0)$ by

$$\begin{aligned} E\left[h(Y)|D=0\right] &= \frac{1}{p}\left(E[h(Y)] - (1-p)E[h(Y)|D=1]\right) \\ &= \frac{1}{p}\left((1-p)E\left[\frac{h(Y)}{P(Y)}|D=1\right] - (1-p)E[h(Y)|D=1]\right) \\ &= \frac{1-p}{p}E\left[\frac{1-P(Y)}{P(Y)}h(Y)|D=1\right], \end{aligned}$$

where $p = P(D = 0)$. $E(Y|D = 0)$ can be identified similarly. Then, using (5.3), we obtain the following lower and upper bounds on Δ^{TT}:

$$\underline{\Delta}_1^{TT} = E[Y_1|D = 0] - \frac{1-p}{p} E\left[\frac{1-P(Y)}{P(Y)} h(Y)|D = 1\right], \tag{5.4}$$

$$\overline{\Delta}^{TT} = E[Y_1|D = 0] - \frac{1-p}{p} E\left[\frac{1-P(Y)}{P(Y)} Y|D = 1\right]. \tag{5.5}$$

To estimate these bounds, we first need estimates of h and P. To estimate h, we use a kernel estimator with a Gaussian kernel and a bandwidth estimated by cross-validation (see Figure 2). P is estimated using the flexible parametric form $P(y;\beta) = 1/f(y;\beta)$ with $f(y;\beta)$ defined by (4.2). The same instruments as in the Monte Carlo simulations are used, except that the thresholds (a_2, a_3, a_4) (resp. (c_2, c_3, c_4)) correspond to the estimated quantiles of order 8, 16 and 24 of the distribution of Y conditional on $D = 1$ (resp. of Z).[28] The resulting estimate $P(.;\widehat{\beta})$ is displayed in Figure 2.[29]

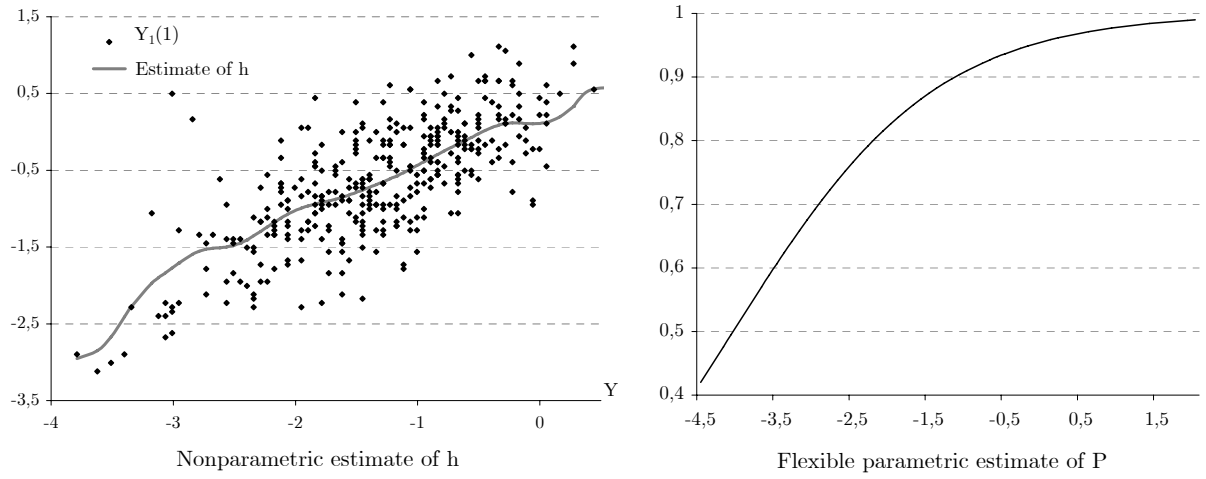

Figure 2: Estimates of h and P.

The estimators of $\underline{\Delta}_1^{TT}$ and $\overline{\Delta}^{TT}$ are the empirical analogs of (5.4) and (5.5):

$$\widehat{\underline{\Delta}_1^{TT}} = \frac{1}{n_0}\left[\sum_{i/D_i=0} Y_{1i} - \sum_{i/D_i=1} \frac{P(Y_i;\widehat{\beta})}{1-P(Y_i;\widehat{\beta})} \widehat{h}(Y_i)\right],$$

$$\widehat{\overline{\Delta}^{TT}} = \frac{1}{n_0}\left[\sum_{i/D_i=0} Y_{1i} - \sum_{i/D_i=1} \frac{P(Y_i;\widehat{\beta})}{1-P(Y_i;\widehat{\beta})} Y_i\right],$$

where n_0 denotes the number of pupils who repeated the 5th grade.

[28] We tried several specifications. Final results are not sensitive to the choice of the thresholds.

[29] This plot corresponds to $\widehat{\beta}_0 = 3.07$, $\widehat{\beta}_1 = 0.75$, $\widehat{\beta}_2 = 4.13$,$\widehat{\beta}_3 = 34.3$, $\widehat{\beta}_4 = 0.42$.

Second strategy: monotonicity

Basically, the conditional independence condition holds if Y is a perfect measure of ability at the end of the 5th grade and if teachers only take into account current ability when deciding whether to retain a student or not. Even if the latter statement is rather plausible, given that teachers in France usually do not observe children's ability before they enter their grade, the former statement seems too restrictive. Lagged test scores probably contain information on current ability and thus may explain part of grade retention. It also seems very plausible that both variables have a positive effect on promotion, so that Assumptions 3' and 6 hold. To provide empirical evidence that supports this claim, we estimate a logit model for D_1 on the sample of students who were promoted in the 6th grade. For these students, both Y and Z are known. The results, which are reported in Table 3, confirm the positive effect of both variables. As expected, we also observe a far smaller effect of the 3rd grade test score.

Variable	Estimate (std. err.)
2002 6th grade score Y	1.31 (0.08)
3rd grade score Z	0.23 (0.07)

Table 3: Logit estimation of the probability of promotion in the 6th grade.

To apply Theorem 2.5 and obtain bounds on $E(h(Y)|D=0)$, we also need to check that $h \in H_Y \cap H_{YZ}$. That h is increasing is apparent from Figure 2. To check that $h \in H_{YZ}$, we implement, as suggested in Subsection 2.3, the specification test of the form (2.4).[30] We obtain a positive and significant slope coefficient in (2.4) and do not reject the linear specification at the 1% level. Hence, we do not reject the assumption that $h \in H_{YZ}$.

Under Assumptions 3' and 6, and the condition $h \in H_Y \cap H_{YZ}$, we can apply Theorem 2.5 to obtain the following bounds on $E(h(Y)|D=0)$:

$$\frac{1-p}{p} E\left[\frac{1-Q(Y)}{Q(Y)} h(Y)|D=1\right] \leq E\left[h(Y)|D=0\right] \leq E\left[E(h(Y)|Z,D=1)|D=0\right],$$

where Q denotes a solution of $E(D/Q(Y)-1|Z)=0$.[31]

To get bounds on $E(Y|D=0)$, we also check that the identity function belongs to H_{YZ}. This is true if $E(Z|D=1,Y)=\gamma+\lambda Y$ with $\lambda>0$. The specification test does not reject

[30] More precisely, we implement the simple differencing test suggested by Yatchew (1998, p. 701) using the kernel estimator $\widehat{h}$ instead of h.

[31] We do not use P here to emphasize the fact that the solution of this equation is not $P(D=1|Y)$ anymore. However, both P and Q are estimated by $P(.;\widehat{\beta})$.

the null at the level of 5%, so we accept the hypothesis that the identity function belongs to $H_Y \cap H_{YZ}$. Under these assumptions, we get the same upper bound on Δ^{TT} as under conditional independence, but another lower bound, which satisfies

$$\underline{\Delta}_2^{TT} = E[Y_1|D=0] - E\left[E(h(Y)|Z, D=1)|D=0\right]. \tag{5.6}$$

Moreover, $\underline{\Delta}_2^{TT}$ and $\overline{\Delta}^{TT}$ are sharp by Theorem 2.5.

To estimate $\underline{\Delta}_2^{TT}$, a kernel estimator $\widehat{g}$ of $g(z) = E(h(Y)|Z=z, D=1)$ is first estimated and then plugged into the empirical analog of (5.6):

$$\widehat{\underline{\Delta}_2^{TT}} = \frac{1}{n_0}\left[\sum_{i/D_i=0} Y_{1i} - \sum_{i/D_i=1} \widehat{g}(Z_i)\right].$$

5.3 Results

The final results are displayed in Table 4. Under the assumption of a fully valid instrument, the estimated identification interval for Δ^{TT} only covers positive values, so grade retention results in a positive short-term effect even in the least favorable case.[32] The pattern is less clear if one only relies on monotonicity conditions. If grade retention only depended on the 3rd grade test score, repeating a grade would be harmful for test achievement. This assumption does not seem very credible, though. As emphasized previously, the effect of Y on D is probably much more important than the one of Z. Thus, even in a worst case scenario the true effect is more likely to be close to $\widehat{\underline{\Delta}_1^{TT}}$, that is to say around zero.

Estimator	Value	95% Confidence interval
$\widehat{\overline{\Delta}^{TT}}$	1.17 (0.24)	[0.75,1.67]
$\widehat{\underline{\Delta}_1^{TT}}$	0.29 (0.16)	[0.02,0.65]
$\widehat{\underline{\Delta}_2^{TT}}$	-0.43 (0.06)	[-0.53,-0.30]

Standard errors are obtained through bootstrap with 1,000 replicates. Effects are measured in standard deviations terms.

Table 4: Bounds on Δ^{TT} under different assumptions.

In conclusion, and even if the uncertainty in Δ^{TT} is rather large,[33] the short-term effect of grade retention seem more likely to be positive. This result is in line with those found by

[32] The null hypothesis that the lower bound is negative is rejected at 5%.

[33] This uncertainty is mainly due to endogenous selection in grade retention, which prevents us from recovering the counterfactual progression $Y_1(1) - Y$ of retained students. This issue accounts for 55% of

Jacob and Lefgren (2004) for third-graders in Chicago but is more optimistic than those found for sixth-graders. This difference might stem from the fact that the grade retention decision rules are not the same in the two countries. Letting teachers and parents decide on the basis of their observations during the whole year, and not only on two tests as in Chicago, may reduce the impact of measurement error on grade retention. On the other hand, a discretionary process as in France is likely to favor (or penalize) systematically some subpopulations of students, independently of their ability, and thus to decrease the efficiency of grade retention. The results suggest that the former effect dominate the latter.

6 Conclusion

This paper considers the issue of endogenous selection. The key assumption for identification, which contrasts with the usual ones in selection problems, is the independence between the instrument and the selection variable, conditional on the outcome. A general nonparametric identification result is obtained under a completeness condition. This framework can be applied to a broad class of selection models, including Roy models with unobserved sectors, nonignorable nonresponse or binary choice models when data are observed for only one response stratum. Set identification is also considered when the conditional independence condition fails. Sharp and finite bounds on a class of parameters of interest are obtained under weaker conditions of monotonicity. These results are applied to estimate bounds on the effect of grade retention in France.

The paper raises two challenging issues. First, one may wonder whether the ideas developed here could be adapted to a generalized Roy model in which selection depends on the predicted dependent variable rather than on the dependent variable itself. In this case, the conditional independence condition breaks down but the structure of the model may provide information for point or at least set identification. Second, the sharp upper bounds are obtained on an abstract set of parameters. Further characterizations of this set appear desirable, for both theoretic and practical reasons.

---

the width of the set $\left[\widehat{\underline{\Delta}_2^{TT}}, \widehat{\overline{\Delta}^{TT}}\right]$, while the uncertainty in the true effect of the instrument Z on grade retention only accounts for 45% of this width.

Appendix: proofs

Proposition 2.1

Let $\mathcal{A}_y = \{u/\psi(y,u) = 1\}$ and $\mathcal{C}_{y,z} = \{v/\varphi(z,v) = y\}$. We get, for all (y,z),

$$\begin{aligned} P(D=1|Y=y,Z=z) &= P(\eta \in \mathcal{A}_y|Y=y,Z=z) \\ &= P(\eta \in \mathcal{A}_y|\varepsilon \in \mathcal{C}_{y,z}(y,z),Z=z) \\ &= P(\eta \in \mathcal{A}_y) \\ &= P(\eta \in \mathcal{A}_y|Y=y) \\ &= P(D=1|Y=y), \end{aligned}$$

where the third and fourth equalities follow from the condition $\eta \perp\!\!\!\perp (Z,\varepsilon)$. Thus, Assumption 3 holds $\square$

Proposition 2.2

The proof proceeds in three steps.

1. First, we show that there exist positive c_1, c_2 and $0 < \alpha' < \alpha - 2$ such that

$$c_1 \leq (f_\varepsilon \star f_{\alpha'})(x) \times (1+|x|)^{\alpha'+1} \leq c_2, \tag{6.1}$$

where $f_{\alpha'}$ denotes the density of an α'-stable distribution of characteristic function $\exp(-|t|^{\alpha'})$, and $\star$ denotes the convolution product.
To prove (6.1), note that $f_{\alpha'}$ satisfies, for well-chosen $c < C$ (see e.g. Mattner 1992, p.146),

$$c \leq f_{\alpha'}(x) \times (1+|x|)^{\alpha'+1} \leq C. \tag{6.2}$$

Let $I = [a,\, b] \subset [-1,\, 1]$ denote an interval such that $\inf_{x\in I} f_\varepsilon(x) = m > 0$ (such an interval exists by the regularity conditions). For all $x \in \mathbb{R}$ and all $t \in I$,

$$\begin{aligned} 1+|x-t| &\leq 1+\max(|x-a|,|x-b|) \\ &\leq 1+|x|+\max(|a|,|b|) \\ &\leq 2(1+|x|). \end{aligned}$$

Thus,

$$\begin{aligned}(f_\varepsilon \star f_{\alpha'})(x) &\geq \int_I f_\varepsilon(t) f_{\alpha'}(x-t)dt \\ &\geq mc \int_I \frac{dt}{(1+|x-t|)^{\alpha'+1}} \\ &\geq \frac{mc(b-a)}{2^{\alpha'+1}(1+|x|)^{\alpha'+1}}.\end{aligned}$$

This proves the first inequality of (6.1). To prove the second, remark that by the regularity conditions, there exists M such that

$$(1+|t|)^\alpha f_\varepsilon(t) \leq M. \tag{6.3}$$

Moreover, for all $x \geq 0$ and $t < x/2$, we get $1+|x-t| \geq (1+x)/2$. Thus, using both (6.2) and (6.3), we get

$$\begin{aligned}\int_{-\infty}^{x/2} f_\varepsilon(t) f_{\alpha'}(x-t)dt &\leq \frac{2^{\alpha'+1}MC}{(1+x)^{\alpha'+1}} \int_{-\infty}^{x/2} \frac{dt}{(1+|t|)^\alpha} \\ &\leq \frac{2^{\alpha'+1}MC}{(1+x)^{\alpha'+1}} 2 \int_{-\infty}^{0} \frac{dt}{(1-t)^\alpha} \\ &\leq \frac{2^{\alpha'+2}MC}{(\alpha-1)(1+|x|)^{\alpha'+1}}.\end{aligned} \tag{6.4}$$

Besides, because $f_{\alpha'}(x-t) \leq C$ and $\alpha - 1 > \alpha' + 1$,

$$\begin{aligned}\int_{x/2}^{+\infty} f_\varepsilon(t) f_{\alpha'}(x-t)dt &\leq MC \int_{x/2}^{+\infty} \frac{dt}{(1+t)^\alpha} \\ &\leq \frac{2^{\alpha-1}MC}{(1+x)^{\alpha-1}} \\ &\leq \frac{2^{\alpha-1}MC}{(1+|x|)^{\alpha'+1}}.\end{aligned}$$

This, together with (6.4), shows that for all $x \geq 0$, there exists a constant C' such that $(f_\varepsilon \star f_{\alpha'})(x) \times (1+|x|)^{\alpha'+1} \leq C'$. The same reasoning can be applied to any $x < 0$, and the second inequality of (6.1) follows.

2. Now let us show that for any $g \in \mathcal{B}$ such that $E[g(Y)|Z] = 0$ a.s., we get, almost everywhere (a.e. for short),

$$(g \circ \mu) \star \phi = 0, \tag{6.5}$$

where $\phi = f_{-\varepsilon} \star f_{\alpha'}$.

By the definition of $\mathcal{B}$, there exists a K such that $g(Y) \geq K$ almost surely. Let $\widetilde{g}(u) = g(\mu(u)) - K$. Using the additive decomposition, we get

$$\begin{aligned}\mathbb{E}[g(Y) - K|Z] &= \mathbb{E}[\widetilde{g}(\nu(Z) + \varepsilon)|Z] \\ &= \int \widetilde{g}(\nu(Z) + u) f_\varepsilon(u) du \\ &= \int \widetilde{g}(u) f_{-\varepsilon}(\nu(Z) - u) dt.\end{aligned}$$

This implies, by the large support assumption, that

$$\mathbb{E}[g(Y)|Z] = 0 \text{ a.s.} \Leftrightarrow \int \widetilde{g}(u) f_{-\varepsilon}(t - u) dt = -K \text{ a.e.}$$

In other words, $\widetilde{g} \star f_{-\varepsilon} = -K$. Let α' and $f_{\alpha'}$ be defined as previously. We get, a.e.,

$$(\widetilde{g} \star f_{-\varepsilon}) \star f_{\alpha'} = -K.$$

Because $\widetilde{g}, f_{-\varepsilon}$ and $f_{\alpha'}$ are nonnegative functions, we can apply Fubini's theorem, so that $\widetilde{g} \star (f_{-\varepsilon} \star f_{\alpha'}) = -K$ a.e. Equation (6.5) follows.

3. Finally, let us prove that the location family generated by ϕ is complete. This proves the result because then, $g \circ \mu = 0$ a.e. and thus $g(Y) = 0$ almost surely. For this purpose, we check the conditions of Theorem 1.1 of Mattner (1992). First, ϕ satisfies condition (i) of this theorem by (6.1) and Proposition 1.2 of Mattner (1992). Second, the characteristic function Ψ_ϕ corresponding to the density ϕ is as follows:

$$\Psi_\phi(t) = \Psi_\varepsilon(-t) \times \exp(-|t|^{\alpha'}), \tag{6.6}$$

where Ψ_ε denotes the characteristic function of ε. Thus, by the regularity conditions, Ψ_ϕ is infinitely differentiable on $\mathbb{R}\backslash(A \cup \{0\})$ and condition (ii) of Mattner's theorem holds. Finally, by (6.6) and the regularity conditions, Ψ_ϕ does not vanish anywhere. Thus, Theorem 1.1 in Mattner (1992) can be applied, which concludes the proof □

Theorem 2.3

By Assumption 3 and the definition of P,

$$\begin{aligned}P(D = 1|Z) E\left[\frac{1}{P(Y)}|D = 1, Z\right] &= E\left(\frac{D}{P(Y)}\middle|Z\right) \\ &= E\left(\frac{E(D|Y, Z)}{P(Y)}\middle|Z\right) \\ &= E\left(\frac{E(D|Y)}{P(Y)}\middle|Z\right).\end{aligned}$$

Hence,

$$E\left(\frac{D}{P(Y)} - 1\middle| Z\right) = 0. \tag{6.7}$$

By Assumption 2, $P(D = 1|Z)$ can be identified from the data. Thus, for any function R, $E[D/R(Y) - 1|Z]$ can be computed from the data and any candidate for P must therefore satisfy equality (6.7). Now let Q be such a candidate and let $g = P/Q - 1$. g is bounded below by -1. Moreover, Q must satisfy $E[D/Q(Y)] = 1$, which can also be written as $E[P(Y)/Q(Y)] = 1$. This implies that

$$E\left[|g(Y|\right] \leq E\left[\frac{P(Y)}{Q(Y)}\right] + 1 < \infty.$$

Hence, $g \in \mathcal{B}$. Moreover,

$$\begin{aligned} 0 &= E\left(\frac{D}{Q(Y)} - 1\middle| Z\right) \\ &= E\left(\frac{P(Y)}{Q(Y)} - 1\middle| Z\right) \\ &= E\left(g(Y)|Z\right). \end{aligned}$$

This, together with Assumption 4, implies that $g(Y) = 0$ a.s., so that $Q(Y) = P(Y)$ a.s. Thus, P is identified.

To finish the proof, let $f_{D,Y,Z}$ denote the density of (D, Y, Z) with respect to an appropriate measure. $f_{D,Y,Z}(1, y, z)$ is identified by $f_{Y,Z|D=1}(y, z)P(D = 1)$. Moreover, by Assumption 3,

$$\begin{aligned} P(y) &= P(D = 1|Y = y, Z = z) \\ &= \frac{f_{D,Y,Z}(1, y, z)}{f_{Y,Z}(y, z)}. \end{aligned}$$

Similarly,

$$1 - P(y) = \frac{f_{D,Y,Z}(0, y, z)}{f_{Y,Z}(y, z)}.$$

Thus,

$$f_{D,Y,Z}(0, y, z) = \left[\frac{1 - P(y)}{P(y)}\right] f_{D,Y,Z}(1, y, z).$$

Hence, the joint distribution of the data is identified □

Theorem 2.4

Part "if" of the theorem is trivial. To prove the "only if" implication, let us consider a solution Q, which belongs to $(0, 1]$. Define also a function $g_{D,Y,Z}$ by

$$g_{D,Y,Z}(d, y, z) = \left[\frac{1 - Q(y)}{Q(y)}\right]^{1-d} f_{Y,Z|D=1}(y, z)P(D = 1).$$

$g_{D,Y,Z}$ is a density (with respect to a convenient measure λ), since it is nonnegative and integrates to one:

$$\begin{aligned} & \int \left[g_{D,Y,Z}(0,y,z) + g_{D,Y,Z}(1,y,z) \right] d\lambda(y,z) \\ = & \int \frac{f_{Y,Z|D=1}(y,z)P(D=1)}{Q(y)} d\lambda(y,z) \\ = & E\left\{ E\left[\frac{E(D|Y,Z)}{Q(Y)} | Z \right] \right\} \\ = & 1. \end{aligned}$$

Moreover,

$$g_{D,Y,Z}(1,y,z) = f_{Y,Z|D=1}(y,z)P(D=1) \tag{6.8}$$

and

$$\begin{aligned} g_Z(z) &= f_Z(z) \int \frac{f_{Y,Z|D=1}(y,z))P(D=1)}{Q(y)f_Z(z)} dy \\ &= f_Z(z) E\left[\frac{E(D|Y,Z)}{Q(Y)} | Z \right] \\ &= f_Z(z). \end{aligned}$$

This last equality, together with (6.8), ensures that $g_{D,Z}(d,z) = f_{D,Z}(d,z)$. Thus, $g_{D,Y,Z}$ is coherent with the observed data. Finally, because $g_Y(y) = f_{Y|D=1}P(D=1)/Q(y)$, we obtain the following after straightforward manipulations:

$$\begin{aligned} g_{D,Z|Y}(1,z,y) &= Q(y) f_{Z|Y,D=1}(z,y), \\ g_{D,Z|Y}(0,z,y) &= (1-Q(y)) f_{Z|Y,D=1}(z,y). \end{aligned}$$

In other words, the corresponding distribution of (D,Y,Z) satisfies the independence condition of Assumption 3. To conclude, if there exists a solution Q to equation (2.3) which lies in $(0,1]$, then one can rationalize the observed data by a distribution that satisfies the independence condition $\square$

Theorem 2.5

We rely on the following standard result, which is proved for the sake of completeness.

Lemma 6.1 *Let T denote a real random variable and $(h_1,h_2) \in (L^2_T)^2$ be increasing functions. Then $cov(h_1(T),h_2(T)) \geq 0$.*

Proof: let (T_1, T_2) denote two independent copies of T. Then, because both h_1 and h_2 are increasing,

$$(h_1(T_1) - h_1(T_2)) \times (h_2(T_1) - h_2(T_2)) \geq 0.$$

Thus, taking expectations and using the fact that (T_1, T_2) are i.i.d, we get

$$2\left\{E\left[h_1(T)h_2(T)\right] - E\left[h_1(T)\right]E\left[h_2(T)\right]\right\} \geq 0.$$

The result follows □

a) By Lemma 6.1 and Assumption 6,

$$\text{cov}(h(Y), P(D = 1|Y, \widetilde{Z})|\widetilde{Z}) \geq 0.$$

Thus,

$$E(h(Y)|\widetilde{Z})P(D = 1|\widetilde{Z}) \leq E\left(h(Y)D|\widetilde{Z}\right).$$

This implies that

$$E(h(Y)|\widetilde{Z}) \leq E\left(h(Y)|D = 1, \widetilde{Z}\right).$$

Hence, by integration,

$$E[h(Y)] \leq E\left[E\left(h(Y)|D = 1, \widetilde{Z}\right)\right].$$

Moreover, this upper bound is sharp because the two terms are identical under the untestable assumption that $D \perp\!\!\!\perp Y|\widetilde{Z}$.

b) Let $h \in H_{YZ}$ and $\widetilde{h} \in H_Z$ be such that $h(Y) = E[\widetilde{h}(Z)|D = 1, Y]$. We get

$$\begin{aligned}
E\left[\frac{Dh(Y)}{Q(Y)}\right] - E[h(Y)] &= E\left[\frac{DE(\widetilde{h}(Z)|D = 1, Y)}{Q(Y)}\right] - E[h(Y)] \\
&= E\left[\frac{D\widetilde{h}(Z)}{Q(Y)}\right] - E[h(Y)] \\
&= E\left[\widetilde{h}(Z)E\left(\frac{D}{Q(Y)}|Z\right)\right] - E[h(Y)] \\
&= E\left[\widetilde{h}(Z)\right] - E[h(Y)] \\
&= E\left[E\left(\widetilde{h}(Z)|Y\right) - E\left(\widetilde{h}(Z)|D = 1, Y\right)\right].
\end{aligned}$$

Now, because $\widetilde{h}$ and $z \mapsto P(D = 1|Y, Z = z)$ are increasing with probability one, we have, similarly to a),

$$E(\widetilde{h}(Z)|D = 1, Y) \geq E(\widetilde{h}(Z)|Y). \tag{6.9}$$

Thus,

$$E\left[\frac{Dh(Y)}{Q(Y)}\right] \leq E[h(Y)]. \tag{6.10}$$

Moreover, by Theorem 2.4, if there exists a solution Q to equation (2.3), which lies in $(0,1]$, one cannot reject that (6.9) and (6.10) are actually equalities. This implies that $E[Dh(Y)/Q(Y)]$ is a sharp lower bound of $E[h(Y)]$.

c) If $D \perp\!\!\!\perp (Y, \widetilde{Z})$, by independence,

$$E\left[E(h(Y)|D=1,\widetilde{Z})\right] = E\left[E(h(Y)|Z)\right] = E\left[h(Y)\right].$$

Moreover, because $P(D=1)$ is a solution to (2.3), $Q(Y) = P(D=1)$, so that

$$E\left[\frac{Dh(Y)}{Q(Y)}\right] = E(h(Y)).$$

Now, if $Y = \widetilde{Z}$,

$$E\left[E(h(Y)|D=1,\widetilde{Z})\right] = E\left[h(Y)\right].$$

Moreover, because $Y = Z$, equation (2.3) is equivalent to $Q(Y) = P(D=1|Y)$. Hence,

$$E\left[\frac{Dh(Y)}{Q(Y)}\right] = E\left[\frac{E(D|Y)h(Y)}{Q(Y)}\right] = E\left[h(Y)\right] \ \square$$

Proposition 2.6

Let $\widetilde{h}$ denote an increasing function. We have

$$E(\widetilde{h}(Z)|D=1,Y) = \int \widetilde{h}(z) dF_{Z|Y,D=1}(z).$$

Because $\widetilde{h}$ is increasing, there exists a positive measure μ such that for all $z \leq z_1$,

$$\widetilde{h}(z_1) - \widetilde{h}(z) = \int_z^{z_1} d\mu(u).$$

Thus, for all y and all $M \in \mathbb{R}$,

$$\begin{aligned} E(\widetilde{h}(Z)|D=1,Y=y) &= \int_M^{\infty}\int_M^{z} d\mu(u) dF_{Z|Y=y,D=1}(z) - \int_{-\infty}^{M}\int_z^{M} d\mu(u) dF_{Z|Y=y,D=1}(z) \\ &\quad +2\widetilde{h}(M). \end{aligned}$$

Hence, by Fubini's theorem on nonnegative functions,

$$E(\widetilde{h}(Z)|D=1,Y=y) = \int_M^{\infty}(1-F_{Z|Y=y,D=1}(u))d\mu(u) - \int_{-\infty}^{M} F_{Z|Y=y,D=1}(u)d\mu(u) + 2\widetilde{h}(M).$$

Consequently, we get, for all $y \leq y_1$,

$$\begin{aligned} & E(\widetilde{h}(Z)|D=1,Y=y_1) - E(\widetilde{h}(Z)|D=1,Y=y) \\ = & \int [F_{Z|D=1,Y=y}(u) - F_{Z|D=1,Y=y_1}(u)] d\mu(u). \end{aligned}$$

By assumption, the right-hand side is nonnegative, and the result follows $\square$

Theorem 2.7

a) β_0 satisfies

$$E\left(\frac{DW}{F(V'\beta_0)}\right) = E\left(\frac{W}{F(V'\beta_0)}E(D|Z,V)\right) = E\left(\frac{W}{F(V'\beta_0)}E(D|V)\right) = E(W),$$

where the second equality follows from Assumption 3. Local identification only requires that the differential of $\beta \to E(DW/F(V'\beta))$ is full rank at $\beta = \beta_0$. This differential is $-E(DWV'F'(V'\beta_0)/F^2(V'\beta_0))$, so the result follows from Assumption 4'.

b) Suppose that there exists β such that

$$E\left(\frac{DW}{F(V'\beta)}\right) = E(W) = E\left(\frac{DW}{F(V'\beta_0)}\right). \tag{6.11}$$

Then

$$E\left(\left(\frac{1}{F(V'\beta_0)} - \frac{1}{F(V'\beta)}\right)W(\beta_0 - \beta)\middle| D = 1\right) = 0.$$

Thus

$$E\left(\left(\frac{1}{F(V'\beta_0)} - \frac{1}{F(V'\beta)}\right)E(W|V, D = 1)(\beta_0 - \beta)\middle| D = 1\right) = 0.$$

Now, by Assumption 4",

$$E(W|V, D = 1) = \begin{pmatrix} I_r & 0 \\ \Gamma_1 & \Gamma_2 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} \equiv \Gamma V,$$

where I_r is the identity matrix of size r and r is the dimension of X. Moreover, because Γ_2 is full rank, Γ is also full rank. Hence

$$E\left(\left(\frac{1}{F(V'\beta_0)} - \frac{1}{F(V'\beta)}\right)(V'\beta_0 - V'\beta)\middle| D = 1\right) = 0.$$

Because F is strictly increasing, for any $x \neq y$, $(x - y)(1/F(x) - 1/F(y)) < 0$, so that

$$P(V'(\beta_0 - \beta) = 0|D = 1) = 1.$$

By Assumption 2'-c), this implies that $\beta = \beta_0$ □

Theorem 3.1.

As Horowitz and Lee (2007), we adapt the proof of Theorem 2 of Bissantz et al. (2004). By definition of $\widehat{f}$,

$$\max\left(\left\|\widehat{T}\,\widehat{f} - 1\right\|^2, \alpha_n\left\|\widehat{f}\right\|^2\right) \leq \left\|\widehat{T}\,\widehat{f} - 1\right\|^2 + \alpha_n\left\|\widehat{f}\right\|^2 \leq \left\|\widehat{T}\,f - 1\right\|^2 + \alpha_n\left\|f\right\|^2 \tag{6.12}$$

Let $\delta_n = h_n^2 + 1/nh_n$. Because $E(\left\|\widehat{T}\, f - 1\right\|^2) = O(\delta_n)$ (see e.g. Györfi et al., 2002) and $\delta_n/\alpha_n \to 0$, we get

$$\limsup E(\left\|\widehat{f}\right\|^2) \leq \left\|f\right\|.$$

Inequalities (6.12) and $\delta_n/\alpha_n \to 0$ also imply that $E(\left\|\widehat{T}\,\widehat{f} - 1\right\|^2) \to 0$. D_M is weakly closed as a closed and convex set (see Bissantz et al., 2004). Moreover, for all $\phi \in D_M$, by Jensen's inequality,

$$(T\,\phi)^2 \leq E(\phi(Y)^2\,|Z).$$

Hence,

$$\begin{aligned}
\left\|T\,\phi\right\|^2 &\leq \int \left[\int \phi(y)^2 f_{Y|Z}(y|z)dy\right] dz \\
&\leq \int \phi(y)^2 \left[\int \frac{f_{Z|Y}(z|y) f_Y(y)}{f_Z(z)} dz\right] dy \\
&\leq \frac{\sup_{y\in[0,1]} f_Y(y)}{\inf_{z\in[0,1]} f_Z(z)} \int \phi(y)^2 dy,
\end{aligned}$$

where the second inequality follows from Fubini's theorem and Bayes' theorem. By Assumption 8-b), there exists therefore a $A < +\infty$ such that

$$\left\|T\,\phi\right\|^2 \leq A\left\|\phi\right\|^2.$$

This inequality and the linearity of T proves that it is continuous. Hence, T is weakly continuous. This and the fact that D_M is weakly closed ensures that T is weakly sequentially closed (see Bissantz et al., 2004). Consequently, we can apply the end of the proof of Theorem 2 of Bissantz et al. (2004), and the result follows □

Corollary 3.2

By the triangular inequality,

$$|\widehat{\theta} - \theta| \leq \frac{1}{n}\sum_{i=1}^{n} D_i |g(Y_i^*, Z_i)||\widehat{f}^{-i}(Y_i^*) - f(Y_i^*)| + \left|\frac{1}{n}\sum_{i=1}^{n} D_i g(Y_i^*, Z_i) f(Y_i^*) - \theta\right|. \quad (6.13)$$

By Assumption 8, $|D_i f(Y_i^*)| \leq M$. Hence, $E[|D_i g(Y_i^*, Z_i) f(Y_i^*)|^2] < \infty$ and by the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} D_i g(Y_i^*, Z_i) f(Y_i^*) \xrightarrow{L^2} E(Dg(Y^*, Z) f(Y^*)).$$

Moreover,

$$\begin{aligned} E(Dg(Y^*, Z)f(Y^*)) &= E(Dg(Y, Z)f(Y)) \\ &= E(E(D|Y, Z)g(Y, Z)f(Y)) \\ &= \theta. \end{aligned}$$

Thus, the second term of the right-hand side of (6.13) tends to zero in quadratic mean.

Now, because the $(\widehat{f}^{-i}(Y_i^*))_i$ are identically distributed, the first term T_1 of the right-hand side of (6.13) satisfies

$$\begin{aligned} E\left(|T_1|\right) &= E\left(D_1|g(Y_1^*, Z_1)||\widehat{f}^{-1}(Y_1^*) - f(Y_1^*)|\right) \\ &\leq \sqrt{E\left(|g(Y_1^*, Z_1)|^2\right) E\left(|\widehat{f}^{-1}(Y_1^*) - f(Y_1^*)|^2\right)}, \end{aligned}$$

by the Cauchy-Schwartz inequality. Now, by independence between Y_1^* and $\widehat{f}^{-1}$,

$$E(|\widehat{f}^{-1}(Y_1^*) - f(Y_1^*)|^2) \leq \sup_{y\in[0,1]} f_Y(y) E\left(||\widehat{f}^{-1} - f||^2\right).$$

Thus, the left-hand side tends to zero by Theorem 3.1. As a consequence, $E(|T_1|)$ also tends to zero. This yields the announced result □

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**, 444–455.

Bissantz, N., Hohage, T. and Munk, A. (2004), 'Consistency and rates of convergence of nonlinear tikhonov regularization with random noise', *Inverse Problems* **20**, 1773–1789.

Blundell, R., Chen, X. and Kristensen, D. (2007), 'Nonparametric iv estimation of shape-invariant engel curves', *Econometrica* **75**, 1613–1669.

Carrasco, M., Florens, J. P. and Renault, E. (2006), Linear inverse problems and structural econometrics: Estimation based on spectral decomposition and regularization, *in* J. J. Heckman and E. E. Leamer, eds, 'Handbook of Econometrics', Vol. 6, North Holland.

Chamberlain, G. (1986), 'Asymptotic efficiency in semiparametric model with censoring', *Journal of Econometrics* **32**, 189–218.

Chen, C. (2001), 'Parametric models for response-biased sampling', *Journal of the Royal Statistical Society, Series B* **63**, 775–789.

Chen, X. and Hu, Y. (2006), Identification and inference of nonlinear models using two samples with arbitrary measurement errors. Cowles foundation discussion paper no. 1590.

Cosnefroy, O. and Rocher, T. (2004), 'Le redoublement au cours de la scolarité obligatoire : nouvelles analyses, mêmes constats', *Education et Formation* **70**, 73–82.

Crahaye, M. (1996), *Peut-on lutter contre l'échec scolaire ?*, De Boeck.

Darolles, S., Florens, J. P. and Renault, E. (2006), Nonparametric instrumental regression. Working Paper.

Deville, J. C. (2002), La correction de la non-réponse par calage généralisé, *in* 'Actes des Journées de Méthodologie Statistique 2002', INSEE, pp. 4–20.

D'Haultfœuille, X. (2008), 'On the completeness condition in nonparametric instrumental regression', *Econometric Theory, forthcoming* .

Gagliardini, P. and Scaillet, O. (2006), Tikhonov regularization for functional minimum distance estimators. Working Paper.

Gouriéroux, C. and Monfort, A. (1995), *Statistics and Econometric Models*, Cambridge University Press.

Grogger, J. T. and Carson, R. T. (1991), 'Models for truncated counts', *Journal of Applied Econometrics* **6**, 225–238.

Györfi, L., Kohler, M., Kryzak, A. and Walk, H. (2002), *A Distribution-Free Theory of Nonparametric Regression*, New York: Springer.

Hall, P. and Horowitz, J. L. (2005), 'Nonparametric methods for inference in the presence of instrumental variables', *Annals of Statistics* **33**, 2904–2929.

Haurin, D. R. and Sridhar, K. S. (2003), 'The impact of local unemployment rates on reservation wages and the duration of search for a job', *Applied Economics* **35**, 1469–1475.

Heckman, J. J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica* **42**, 679–694.

Heckman, J. J. and Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation', *Econometrica* **73**, 669–738.

Hellerstein, J. K. and Imbens, G. W. (1999), 'Imposing moment restrictions from auxiliary data by weighting', *The Review of Economics and Statistics* **81**, 1–14.

Hemvanich, S. (2004), The general missingness problems and estimation in discrete choice models. Working Paper.

Hirano, K., Imbens, G. W. and Ridder, G. (2003), 'Efficient estimation of average tratment effects using the estimated propensity score', *Econometrica* **71**, 1161–1189.

Holmes, T. (1989), Grade level retention effects : A meta-analysis of research studies, *in* L. A. Sheppard and M. L. Smith, eds, 'Flunking Grades. Research and Policies on Retention', New York, The Falmer Press, pp. 16–33.

Horowitz, J. L. and Lee, S. (2007), 'Nonparametric instrumental variables estimation of a quantile regression model', *Econometrica* **75**, 1191–1208.

Horvitz, D. G. and Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**, 663–685.

Hu, Y. and Schennach, S. (2008), 'Instrumental variable treatment of nonclassical measurement error models', *Econometrica* **76**, 195–216.

Imbens, G. (2004), 'Nonparametric estimation of average treatment effects under exogeneity: a review', *The Review of Economics and Statistics* **86**, 4–29.

Imbens, G. W. and Lancaster, T. (1994), 'Combining micro and macro data in microeconometric models', *Review of Economic Studies* **61**, 655–680.

Jacob, B. A. (2005), 'Accountability, incentives and behavior: Evidence from school reform in chicago', *Journal of Public Economics* **89**, 761–796.

Jacob, B. A. and Lefgren, L. (2004), 'Remedial education and student achievement: A regression-discontinuity analysis', *Review of Economics and Statistics* **86**, 226–244.

Jimerson, S. (2001), 'Meta-analysis of grade retention research: Implications for practice in the 21st century', *School Psychology Review* **30**, 420–437.

Jimerson, S., Anderson, G. E. and Whipple, A. D. (2002), 'Winning the battle and losing the war: Examining the relationship between grade retention and dropping out of high school', *Psychology in the Schools* **39**, 441–457.

Lewbel, A. (2007), 'Endogenous selection or treatment model estimation', *Journal of Econometrics* **141**, 777–806.

Little, R. and Rubin, D. B. (1987), *Statistical analysis with Missing Data*, John Wiley & Sons, New York.

Lorence, J. (2006), 'Retention and academic achievement research revisited from an united states perspective', *International Education Journal* **7**, 731–777.

Manski, C. F. (1994), The selection problem, *in* C. Sims, ed., 'Advances in Econometrics, Sixth World Congress', Cambridge University Press.

Manski, C. F. (2003), *Partial Identification of Probability Distribution*, Springer.

Manski, C. F. and Pepper, J. V. (2000), 'Monotone instrumental variables: With an application to the returns to schooling', *Econometrica* **68**, 997–1010.

Mattner, L. (1992), 'Completeness of location families, translated moments, and uniqueness of charges', *Probability Theory and Related Fields* **92**, 137–149.

Mattner, L. (1993), 'Some incomplete but boundedly complete location families', *Annals of Statistics* **21**, 2158–2162.

Nevo, A. (2002), 'Using weights to adjust for sample selection when auxiliary information is available', *Journal of Business and Economics Statistics* **21**, 43–52.

Newey, W. and Powell, J. (2003), 'Instrumental variable estimation of nonparametric models', *Econometrica* **71**, 1565–1578.

Ramalho, E. A. and Smith, R. J. (2007), Discrete choice nonresponse. CEMMAP working paper.

Tang, G., Little, R. J. A. and Raghunathan, T. E. (2003), 'Analysis of multivariate missing data with nonignorable nonresponse', *Biometrika* **90**, 747–764.

Troncin, T. (2005), Le redoublement : radiographie d'une décision à la recherche de sa légitimité. PhD Thesis, available at http://tel.archives-ouvertes.fr/docs /00/14/05/31/PDF/05076.pdf.

Wooldridge, J. (2007), 'Inverse probability weighted estimation for general missing data problems', *Journal of Econometrics* **141**, 1281–1301.

Yatchew, A. (1998), 'Nonparametric regression techniques in economics', *Journal of Economic Literature* **36**, 669–721.