Are Private Schools Better Than Public Schools?
Appraisal for Ireland by Methods for Observational Studies

Danny Pfeffermann,

Hebrew university of Jerusalem, *Israel*, *and University of Southampton, UK*;

Victoria Landsman,

National Cancer Institute, NIH, Rockville, MD, *USA.*

In observational studies the assignment of units to treatments is not under control. Consequently, the estimation and comparison of treatment effects based on the empirical distribution of the responses can be biased since the units exposed to the various treatments could differ in important unknown pretreatment characteristics, which are related to the response. An important example studied in this article is the question of whether private schools offer better quality of education than public schools. In order to address this question we use data collected in the year 2000 by OECD for the *Programme for International Student Assessment* (PISA). Focusing for illustration on scores in mathematics of 15-years old pupils in Ireland, we find that the raw average score of pupils in private schools is higher than of pupils in public schools. However, application of a newly proposed method for observational studies suggests that the less able pupils tend to enroll in public schools, such that their lower scores is not necessarily an indication of bad quality of the public schools. Indeed, when comparing the average score in the two types of schools after adjusting for the enrollment effects, we find quite surprisingly that public schools perform better on average. This outcome is supported by the methods of instrumental variables and latent variables, commonly used by econometricians for analyzing and evaluating social programs.

KEY WORDS: Average treatment effect; Goodness of fit; Identifiability; Instrumental variables; Private-dependent schools; Propensity scores; Sample distribution.

ACKNOWLEDGMENT: This paper is based on the PhD thesis of the second author written at the Hebrew University of Jerusalem, Israel, under the supervision of the first author. The research was supported by the Israel Science Foundation, Grant No. 1277/05. The authors are grateful to the Editor, Ass. Editor and two referees for very insightful and constructive comments that improved the quality of the paper.

1. INTRODUCTION

In observational studies the assignment of units to treatments often depends on latent variables that are related to the response variable even when conditioning on known covariates. Consequently, a direct comparison of the response distributions (given the model covariates) or moments of these distributions between treatment groups could be biased and result in wrong conclusions. An important example studied in this article is the question of whether private schools offer better quality of education than public schools. This question has important impact on educational policy and public finance (Hanushek, 2002). It is known that pupils enrolling to the two types of schools differ in their family background and other characteristics related to their scholastic achievements, such that a raw comparison of the scores of pupils attending the two types of schools can be misleading. In an attempt to deal with this question we use data collected in the year 2000 by OECD for the *Programme for International Student Assessment* (PISA). The purpose of this program is to study and compare the proficiency of pupils aged 15 from over than 30 countries in mathematics, science and reading. In this article we focus on scores in mathematics in Ireland and estimate the difference in the average score between the two types of schools by adjusting for the quality of pupils enrolling to them and for the effects of known covariates.

We start by applying several existing methods for observational studies to the data, which are described in Section 3, and find, similarly to Vanderberghe and Robins (2004), that some of these methods produce estimates of different magnitude and sign. We attempt to resolve this conflict by developing and applying a new method for inference from observational data, which extends recent methodology for analyzing sample survey data. The method derives the *sample distribution* of the observed response under a given treatment (score in mathematics in a given type of school in our application) as a function of the distribution that would be obtained under a strongly ignorable assignment of subjects to treatments, and the assignment probability, which is allowed to depend on the response value. The use of this approach is established by showing that the sample distribution is identifiable under some general conditions. The goodness of fit of the sample distribution can be tested by standard test statistics since it refers to the observed data.

By fitting the sample distribution to the observed data we can estimate the distribution under strongly ignorable assignment to treatments, and the assignment probabilities, which are then used for estimating the treatment effects or contrasts between them. Our approach permits also testing some of the assumptions underlying other methods for analyzing observational data, thus enabling us to understand better why different methods yield different answers in our application.

Section 2 describes the PISA data and defines the problem underlying this study more formally. Section 3 overviews some of the existing methods for observational studies and shows the results obtained when applying them to the PISA data for Ireland. We also consider 'probability weighted' versions of the estimators, which account for unequal sample selection probabilities that are possibly related to the response and may thus bias the inference. Computation of these estimators yields very similar estimates to the estimates obtained under the standard methods. Section 4 presents the proposed approach and shows the results obtained when applying it to the PISA data. The main outcome of this analysis is that after controlling for the effect of the enrollment process, the public schools actually outperform the private schools in the average math score, suggesting better quality of education. Here also we extend the method to account for unequal selection probabilities and obtain similar estimates. Section 5 overviews the main theoretical properties of the new approach. Theoretical derivations are presented in the Appendix. We conclude with a brief summary in Section 6.

2. DATA USED FOR APPLICATION AND FORMULATION OF THE PROBLEM

2.1 Sampling design and response values

In order to compare the private and public schools we use data collected in *Ireland* in the year 2000 by OECD for PISA.

Sampling design

PISA uses in most countries a stratified two-stage sampling design. The strata are defined by the size of the school, type of school and gender composition. In each stratum, a probability proportional to size (PPS) sample of schools was selected with the size defined by the number of 15 years old pupils enrolled in the school. A minimum of 150 schools has been selected in each country, or all the schools if there are less. In the second stage an equal probability sample of 35 pupils from the

corresponding age group was drawn from each of the sampled schools (or all the pupils in schools with less than 35 pupils aged 15). By this sampling design, pupils included in the sample do not have equal selection probabilities and each pupil is assigned therefore a sampling weight. The weight is the reciprocal of the pupil's sample inclusion probability, adjusted for non-participation of schools and nonresponse of pupils.

PISA distinguishes between two types of private schools: private-dependent schools where the government contributes 50% or more to the school core funding and private-independent schools with less than 50% government funding. The sample from Ireland consists of 54 public schools, 79 private-dependent schools and only 4 private-independent schools and hence in this paper we do not distinguish between the two types of schools and refer to them simply as private schools. For more information on the PISA sampling design and weighting, see PISA 2000, Technical Report (2002).

Computation of response values

The pupils' proficiencies (scores in mathematics in our case) are not observed directly in the PISA study and are viewed as missing data, which are imputed from the item responses $d_j = (d_{1j}, d_{2j}, ...d_{mj})$, where $d_{ij} = 1$ if pupil j answers correctly question i of the examination and $d_{ij} = 0$ otherwise, $i = 1,...,m$. PISA uses two approaches for imputing the scores: a maximum likelihood approach and a multiple imputation approach. In this paper we used the imputed values obtained under the second approach. See Appendix A for the imputation model. The PISA database contains five sets of imputed values. We standardized the imputed scores in each set by dividing them by their empirical standard deviation and then defined the response value to be the average of the five standardized values. After standardization and averaging, the range of the response values is approximately from 1 to 10. We compared the use of the average values to the results obtained when analyzing each of the five sets of standardized values separately and then combining the results using multiple imputation theory and obtained very similar results in all the analyses performed. Consequently, in this paper we restrict to the average response since it is convenient to have a single working model when simulating new observations, which is needed for the goodness-of-fit tests discussed later.

2.2 Formulation of the problem

The formulation of the problem for the PISA data follows what is known in the literature as the *counterfactual approach*. By this approach, every unit in the population is potentially exposed to every treatment. See, e.g., Rubin (1974), Rosenbaum and Rubin (1983), Smith and Sugden (1988) and Rosenbaum (2002).

Let U define the population of 15 years old pupils in Ireland. Every pupil $i \in U$ has two *potential* responses: Y_{1i}- the proficiency score if the student attends a private school, and Y_{0i} - the proficiency score if the student attends a public school. Let x denote a set of k known covariates (background characteristics) that affect the responses, with values x_i for pupil i. The (potential) population mean score in private schools (hereafter the treatment group) is defined as $\mu_1 = \frac{1}{N}\sum_{i=1}^{N} E_p[Y_{1i} \mid \mathrm{x}_i]$, where N is the population size and the expectation $E_p(\cdot)$ is with respect to the population model holding for the responses. The population mean score in public schools (hereafter the control group) is defined accordingly as $\mu_0 = \frac{1}{N}\sum_{i=1}^{N} E_p[Y_{0i} \mid \mathrm{x}_i]$. In many observational studies, contrasts between the parameters μ_1 and μ_0 are of primary interest. In this paper we focus on estimating the difference between the mean score in private and public schools, defined as,

(2.1) $$\tau = \mu_1 - \mu_0 = \frac{1}{N}\sum_{i=1}^{N} E_p[Y_{1i} \mid \mathrm{x}_i] - \frac{1}{N}\sum_{i=1}^{N} E_p[Y_{0i} \mid \mathrm{x}_i].$$

The contrast τ is known in the literature as the *average treatment effect* (ATE).

In practice, every unit in the population is only exposed to one treatment. Also, it is rarely the case that all the population units participate in the study. The observed data refer therefore to a sample S of size n, which in our application is divided into the two subsamples S_1 and S_0, where S_1 (S_0) is the subsample of pupils attending private (public) schools. For every pupil $i \in S$ we observe therefore y_{1i} if $i \in S_1$ or y_{0i} if $i \in S_0$.

Denote by π_i the probability that pupil i is included in the sample S and by $\tilde{p}_{ti}$ the probability that *sampled* pupil i is enrolled in school of type t. The sample inclusion probabilities π_i (or the sampling weights $w_i = 1/\pi_i$, with possible adjustments for nonresponse or calibration) are typically known for the sampled units,

as is the case in the PISA survey, but the treatment assignment probabilities, $\tilde{p}_{ti}$, are usually unknown and may depend on latent variables that are related to the response, Y_{ti}. As is well-known and illustrated later, if the effect of these latent variables on the response is not accounted for by the observed covariates, the resulting estimators of the population parameters can be highly biased.

Remark 1. The sample inclusion probabilities may likewise be related to the response values and thus bias the inference if not accounted for adequately. This is known in the survey sampling literature as *informative sampling*. Smith and Sugden (1988) define conditions on the sampling design and the treatment assignment process that warrant ignoring them in the inference process. As shown in subsequent sections, there is no evidence for informative sampling with the kind of models and inference methods applied to the PISA data from Ireland.

3. EXISTING METHODS, APPLICATION TO PISA DATA

In what follows we focus on the estimation of the ATE defined by (2.1), assuming that the sample selection probabilities are not related to the response variable and the covariates, and hence that there are no sampling effects. This is the common assumption in the literature even though seldom stated explicitly. After describing several methods in common use and applying them to the PISA data, we show the results obtained when extending the methods to the case where the sample is selected with known unequal probabilities that might be related to the response and compare the results with the results obtained when ignoring the sample selection. Let T define the indicator of the treatment group ($T=1$ for private schools, $T=0$ for public schools). Rosenbaum and Rubin (1983) establish two conditions that warrant ignoring the treatment assignment in the inference process:

- *SI(a)*: the assignment T and the response values (Y_1, Y_0) are independent given the covariates, x, for every unit (pupil),
- *SI(b)*: $0 < \Pr(T=1 \mid \mathrm{x}) < 1$ for every possible x.

Conditions *SI(a)* and *SI(b)* define a *strongly ignorable assignment process* given the covariates. When the assignment is strongly ignorable, it permits the application of a number of simple estimation techniques, which we review in Section 3.1. In Sections 3.2 and 3.3 we consider the latent variables method (LV) and the use of instrumental variables (IV), which do not assume strong ignorability assumptions.

3.1 Methods for strongly ignorable treatment assignments

Regression Estimator: Suppose that the true relationship between Y and x in the population has the general form $Y_t = r_t(\mathrm{x}) + u_t$, $E_p(u_t \mid \mathrm{x}) = 0$ for some functions $r_t(\mathrm{x})$, $t = 0,1$, where $E_p(\cdot)$ is the expectation under a strongly ignorable assignment. Then, the ATE is $(\overline{r}_1 - \overline{r}_0)$, where $\overline{r}_t = \frac{1}{N}\sum_{i=1}^{N} r_t(\mathrm{x}_i)$. When the expectations $r_t(\mathrm{x})$ are linear, $r_t(\mathrm{x}) = \mathrm{x}'\beta_t$, then under the assumptions *SI*(*a*), *SI*(*b*) the regression coefficients can be estimated by ordinary least squares (OLS) and the ATE estimator takes the form,

(3.1) $$\hat{\tau}_{\mathrm{OLS}} = \overline{\mathrm{x}}'(\hat{\beta}_{1,\mathrm{OLS}} - \hat{\beta}_{0,\mathrm{OLS}}),$$

where $\overline{\mathrm{x}}' = (\overline{\mathrm{x}}_1,...,\overline{\mathrm{x}}_K) = \sum_{i=1}^{n} \mathrm{x}_i / n$ and $\hat{\beta}_{t,\mathrm{OLS}}$ is the OLS estimator in subsample S_t.

Matching Estimator: Another procedure in common use is to match the units from the treatment and control groups based on the covariates and then compare the responses. Matching procedures are widely discussed in the literature; see, e.g., Rosenbaum (2002). They do not require specifying the form of the functions $r_t(\mathrm{x})$. Abadie and Imbens (2006) consider the following matching estimate with replacement. Denote by $J_{iM}(t)$ the indices of the M closest matches in S_t for unit $i \in S_{1-t}$, $t = 0,1$. Define for unit $i \in S_{1-t}$, $\hat{y}_{(1-t),i} = y_{(1-t),i}$ and $\hat{y}_{ti} = \frac{1}{M}\sum_{j \in J_{iM}(t)} y_{tj}$. Estimate,

(3.2) $$\hat{\tau}_{\mathrm{M}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{1i} - \hat{y}_{0i}).$$

Other methods use probability weighting with the weights defined by the inverse of the 'propensity score', $e(\mathrm{x}) = \Pr(T = 1 \mid \mathrm{x})$. Rosenbaum and Rubin (1983) show that the conditions *SI(a)*, *SI(b)* for strong ignorability imply the same conditions when x is replaced by $e(\mathrm{x})$, thus validating the use of propensity scores for ATE estimation. In practice, the propensity scores are unknown and are estimated by fitting logistic or probit models, or by use of non-parametric techniques (McCaffrey *et al*., 2004). Below we describe two ATE estimators that use the propensity scores for weighting.

Brewer-Hajek (B-H) estimator: This estimator resembles the familiar Brewer-Hajek (Brewer, 1963; Hajek, 1971) estimator in survey sampling. Let $T_i = 1(0)$ if unit $i \in S_1\,(i \in S_0)$ and define $Y_i = T_i Y_{1i} + (1 - T_i)Y_{0i}$.

(3.3) $$\hat{\tau}_{\text{B-H}} = \left(\sum_{i=1}^{n} \frac{T_i}{\hat{e}(\mathrm{x}_i)}\right)^{-1} \sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(\mathrm{x}_i)} - \left(\sum_{i=1}^{n} \frac{1-T_i}{1-\hat{e}(\mathrm{x}_i)}\right)^{-1} \sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{e}(\mathrm{x}_i)}.$$

Doubly-robust (DR) estimator: If the population expectation $E_p(Y_t \mid \mathrm{x})$ can be modeled by some function $r_t(\mathrm{x})$, then by *SI(a)*, $r_t(\mathrm{x})$ is also the sample expectation and the ATE can be estimated as,

(3.4) $$\hat{\tau}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i - [T_i - \hat{e}(\mathrm{x}_i)]\hat{r}_1(\mathrm{x}_i)}{\hat{e}(\mathrm{x}_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i + [T_i - \hat{e}(\mathrm{x}_i)]\hat{r}_0(\mathrm{x}_i)}{1-\hat{e}(\mathrm{x}_i)}.$$

The estimator (3.4) has the "double-robustness" property of being consistent even if only the model assumed for the propensity scores or the population expectations are correctly specified (Lunceford and Davidian, 2005). Qin and Zhang (2007) consider another estimator that has a somewhat stronger robustness property.

3.2 Latent variable models

This method specifies the joint distribution of the outcome and the treatment selection by use of latent variable (LV) models. The model assumes:

- LV(a)- a structural equation for the population outcomes of the form, $Y_t = r_t(\mathrm{x}) + u_t; E_p(u_t \mid \mathrm{x}) = 0$, $t = 0,1$, and
- LV(b)- a latent variable u_T and an assignment rule T satisfying $T = 1[l(\mathrm{v};\alpha) + u_T > 0]$, where $1[\cdot]$ is the indicator function and $l(\mathrm{v};\alpha)$ is a given function of known covariates v, governed by a vector parameter α. (The covariates in v may include some of the covariates in x but it is generally recommended that $\mathrm{v} \neq \mathrm{x}$ to avoid co-linearity problems in the estimation process (see below). The random variables (u_1, u_0, u_T) are dependent. Under these assumptions, $E_{S_t}(Y_t \mid \mathrm{x}) = E(Y_t \mid \mathrm{x}, T = t) = r_t(\mathrm{x}) + E(u_t \mid \mathrm{x}, T = t) \neq r_t(\mathrm{x}) = E_p(Y_t \mid \mathrm{x})$, since $E(u_t \mid \mathrm{x}, T = t) \neq 0$. However, assuming that $r_t(x) = \mathrm{x}'\beta_t$, (u_1, u_0, u_T) are jointly normal and $l(\mathrm{v};\alpha) = \mathrm{v}'\alpha$, β_t can be estimated by the two-stage Heckman's method (Maddala, 1983), yielding,

(3.5) $$\hat{\tau}_{\text{LV}} = \overline{\mathrm{x}}'(\hat{\beta}_{1,\text{LV}} - \hat{\beta}_{0,\text{LV}}),$$

where $\hat{\beta}_{t,\text{LV}}$ is the LV estimator of β_t, $t = 0,1$. Heckman and Vytlacil (2006) refer to the latter model as the Generalized Roy Model and discuss semi-parametric econometric models, which relax some of the assumptions of this Model.

3.3 Instrumental variables models

Let $Y_t = \mathrm{x}'\beta_t + u_t$; $E_p(u_t \mid \mathrm{x}) = 0$, $t = 0,1$. Then for unit $i \in S$,

(3.6) $$Y_i = T_i \mathrm{x}_i'\beta_1 + (1-T_i)\mathrm{x}_i'\beta_0 + u_i = \tilde{\mathrm{x}}_i'\theta + u_i,$$

where $Y_i = T_i Y_{1i} + (1-T_i)Y_{0i}$, $\tilde{\mathrm{x}}_i' = (T_i \mathrm{x}_i', (1-T_i)\mathrm{x}_i')$, $\theta = (\beta_1', \beta_0')'$ and $u_i = T_i u_{1i} + (1-T_i)u_{0i}$ $= u_{0i} + T_i(u_{1i} - u_{0i})$. In observational studies u_1 and u_0 are correlated with T and hence θ cannot be estimated consistently from (3.6) without additional assumptions. Below we define a set of plausible assumptions warranting that the ATE is estimated consistently. See Wooldridge (2002) for discussion of this and alternative sets of assumptions. Assume the availability of an instrument h satisfying:

- IV(a)- $E_p(Y_t \mid \mathrm{x}, h) = E_p(Y_t \mid \mathrm{x})$ (the population expectation does not depend on the instrument, given the covariates);
- IV(b)- $E_p[T(u_1 - u_0) \mid \mathrm{x}, h] = 0$ (the assignment and the counterfactual gain in the error terms are uncorrelated given the covariates and the instrument);
- IV(c)- $\Pr(T = 1 \mid \mathrm{x}, h) = g(\mathrm{x}, h) \neq \Pr(T = 1 \mid \mathrm{x})$ (the assignment probabilities depend on the instrument and possibly on x).

Multiplying both sides of (3.6) by the column vector $z_i = (g_i \mathrm{x}_i', (1-g_i)\mathrm{x}_i')'$, where $g_i = g(\mathrm{x}_i, h_i)$ and taking expectations yields $E_p(z_i Y_i \mid \mathrm{x}_i, h_i) = E_p(z_i \tilde{\mathrm{x}}_i' \mid \mathrm{x}_i, h_i)\theta$, since under the model and IV(b), $E_p(z_i u_i \mid \mathrm{x}_i, h_i) = 0$. The IV estimator of θ computed from all the observations is $\hat{\theta}_{IV} = (\sum_{i=1}^{n} \hat{z}_i \tilde{\mathrm{x}}_i')^{-1} \sum_{i=1}^{n} \hat{z}_i Y_i$, where $\hat{z}_i = (\hat{g}_i \mathrm{x}_i', (1-\hat{g}_i)\mathrm{x}_i')'$. The estimator $\hat{g}_i = \hat{g}(\mathrm{x}_i, h_i)$ is commonly obtained by fitting probit or logit models. The ATE estimator is,

(3.7) $$\hat{\tau}_{\mathrm{IV}} = \bar{x}'(\hat{\beta}_{1,\mathrm{IV}} - \hat{\beta}_{0,\mathrm{IV}}),$$

with $\hat{\beta}_{t,\mathrm{IV}}$ defined by $\hat{\theta}_{\mathrm{IV}}$, $t = 0,1$.

Remark 2. Condition IV(c) is testable from the data but Conditions IV(a) and IV(b) relate to unobservable quantities and cannot in general be tested directly. Imbens and Angrist (1994) show that for a binary instrument, if condition IV(b) is not satisfied, then under a weaker *monotonicity condition*, $\hat{\tau}_{IV}$ estimates the treatment effect for a

subpopulation consisting of units for which the treatment status would be altered by the instrument. This treatment effect is called *local average treatment effect* (LATE).

3.3 Application of the methods to PISA data for Ireland

We applied the methods reviewed so far to the PISA data for Ireland described in Section 2. The sample consists of 1244 pupils from private schools and 683 pupils from public schools. Six covariates were found to be significant in at least one of the models described in Section 4: gender (GEN, 1 for girls, 0 for boys), mother's education (M.E; 1 for high education, 0 otherwise), family socio-economic index (S.E.I), index of home educational resources (H.E.R), average socio-economic index of the pupil's schoolmates (S.E.S; proposed by Vandenberghe and Robin (2004) to account for potential peer effects), and school location (S.loc; 1 if the school is located in an urban area, 0 otherwise). The continuous covariates have been standardized. To warrant fair comparability between the various methods, we included for the first four methods (Equations 3.1-3.4) all the six covariates in both the regressions and the models used for computing the propensity scores. For the LV and IV methods we included all the covariates except for S.loc in the regressions and all the covariates including S.loc in the school selection models (see Remark 3). Vandenberghe and Robin (2004) considered additional covariates, but these were not found to be significant in our analysis.

Remark 3. The variable *school location* was used by Vandenberghe and Robin (2004) as an instrumental variable. The authors show that it has a significant effect on the probability of attending private schools, thus satisfying the condition IV(c) in Section 3.3. However, the approaches considered in the literature for observational studies do not permit testing that the school location is exogenous to the pupil's proficiency given the other covariates, as required by Condition IV(a), because this condition refers to the population models of the unobservable potential responses. The authors argue that this requirement is plausible, using similar arguments to Hoxby (2000). See Section 4.6 for how we can test this condition under the approach proposed in Section 4.

Table 1 presents the ATE estimates and their standard errors. The first estimate, $\hat{\tau}_{\mathrm{D}} = \overline{y}_1 - \overline{y}_0$ is the crude difference between the simple sample means in the two types of schools. The matching estimator is computed based on $M = 4$ matches. We considered several matching estimates as obtained under different metrics for finding

the matches, with and without adjustments for imperfect matching, and obtained very close results in all the cases. For the instrumental variables method we used the school location as the instrument.

The estimates $\hat{\tau}_{\mathrm{M}}, \hat{\tau}_{\mathrm{LV}}$ and $\hat{\tau}_{\mathrm{IV}}$ were computed by using the functions *nnmatch*, *treatreg* and *ivreg* of the Stata software (StataCorp, 2004). The remaining estimates were programmed using the R software (R Development Core Team, 2004). Estimation of the standard errors of the matching estimators and the LV and IV estimators is incorporated in the Stata functions. See Abadie and Imbens (2006), and Wooldridge (2002) for details. Estimation of the standard errors of the Brewer-Hajek estimator and the doubly robust estimator is developed by Lunceford and Davidian (2005). The estimated standard errors of all the estimators account for the sample distribution of the responses given the observed covariates.

Table 1. Estimates of ATE and standard errors under existing methods

Method	$\hat{\tau}_{\mathrm{D}}$	$\hat{\tau}_{\mathrm{OLS}}$	$\hat{\tau}_{\mathrm{M}}$	$\hat{\tau}_{\mathrm{B\text{-}H}}$	$\hat{\tau}_{\mathrm{DR}}$	$\hat{\tau}_{\mathrm{LV}}$	$\hat{\tau}_{\mathrm{IV}}$
Estimate	0.36	0.12	0.21	0.16	0.17	- 0.49	- 0.61
Std. Error	0.05	0.05	0.05	0.05	0.05	0.19	0.24

The first notable outcome in Table 1 is that the difference $\hat{\tau}_{\mathrm{D}}$ between the simple sample means in the two types of schools is positive, which we anticipated because the more able pupils tend to enroll in private schools. The next four methods from left, which assume strongly ignorable assignment given the covariates, likewise produce small positive ATE estimates. By contrary, the IV and LV methods, which account for treatment assignment effects not explained by the covariates produce negative estimates, with much larger absolute values, suggesting that the public schools actually perform better after accounting for the enrollment effects. A similar outcome is obtained under the approach proposed in Section 4. The use of this approach explains also why the LV and IV methods are more appropriate for this data.

Remark 4. Vandenberghe and Robin (2004) computed what is known in the econometric literature as “the average treatment effect for the treated (ATT)”, using the same data and some of the methods reviewed before, and obtained similar results to the results in Table 1. Dronkers and Avram (2010) computed the ATT for reading scores using PISA data for all the countries by applying several variants of propensity

scores matching. The ATT estimates for Ireland in this study are positive, same as the ATE estimates for the scores in Mathematics based on propensity scores presented in Table 1 ($\hat{\tau}_{\text{B-H}}$ and $\hat{\tau}_{\text{DR}}$).

3.5 Probability Weighted Estimators for PISA Data

So far we ignored the sample selection process when computing the estimators in Section 3.4. The question arising is whether this is justified in the present study. We emphasize again that if the distribution of the response in the sample is affected by the sample selection scheme, the sampling is informative and failing to account for the sampling effects may bias the inference. In fact, even if only the distribution of the covariates in the sample is different from their population distribution, some of the ATE estimators may already be biased. Pfeffermann and Sverchkov (2009) review several existing approaches to account for possible sampling effects in the inference process. In this study we applied what is known as probability weighting, which basically consists of inflating each sample observation proportionally to its sampling weight. The idea of probability weighting is to obtain estimators that are consistent under the randomization (repeated sampling) distribution for the corresponding 'census estimates' that would be computed if all the population values had been observed. The census estimates are free of sampling effects.

We computed the probability weighted estimators (PWE) for all the methods considered so far. See Appendix B for the derivation of these estimators. As a first step we computed the unweighted and probability weighted estimators (in parenthesis) of the population means of the covariates and obtained the following results (the covariates are defined in Section 3.4). GEN: 0.53 (0.52), M.E: 0.61 (0.61), S.E.I: 0.00 (-0.016), S.E.S: 0.00 (-0.016), H.E.R: 0.00 (0.002), S.loc: 0.40 (0.39). As can be seen, the two sets of estimators are very close.

Table 2 shows for each of the methods the unweighted (UNWEI) and probability weighted (WEI) estimators of the mean score in the private and public schools, and the corresponding ATE estimator. The results in Table 2 indicate that the PWE of the mean score as obtained under the various methods are very similar to the corresponding unweighted estimators. This is definitely true for the private schools but even for the public schools the largest difference between the weighted and unweighted estimator is less than 2%. The very small differences between the weighted and unweighted estimators in each type of schools translate into somewhat

larger differences in the estimators of the ATE, but not to an extent that affects the inference. Notice in this regard that when computing the conventional 95% confidence intervals for the true ATE based on the unweighted ATE estimators, all the intervals contain the corresponding weighted estimators. In fact, this would be the case even for confidence intervals with confidence level as low as 68%. Our general conclusion from Table 2 is therefore that the sampling process can be ignored when analyzing the PISA data from Ireland by use of the methods considered so far.

Table 2. Unweighted and weighted estimators of schools mean score and ATE

Method	**Private schools**		**Public schools**		**ATE**	
	UNWEI	WEI	UNWEI	WEI	UNWEI	WEI
Simple difference	6.28	6.29	5.92	5.92	0.36	0.37
Regression	6.21	6.26	6.09	6.12	0.12	0.14
Matching	6.25	6.26	6.04	6.03	0.21	0.23
Brewer-Hajek (B-H)	6.24	6.26	6.08	6.06	0.16	0.20
Doubly Robust (DR)	6.23	6.25	6.06	6.07	0.17	0.18
Instrumental variable	6.00	6.02	6.61	6.52	-0.61	-0.50
Latent variable	6.00	6.02	6.49	6.41	-0.49	-0.39

4. AN ALTERNATIVE APPROACH FOR OBSERVATIONAL STUDIES

In this section we propose an alternative approach for ATE estimation, which as illustrated in Section 4.6, allows also testing the appropriateness of candidate instrumental variables or the use of propensity scores under the assumed model. The approach resembles the LV approach in the sense that it assumes a population model and a model for the treatment selection and applies a combined likelihood resulting from the two models, but all the subsequent developments are very different. As with the IV and LV methods, the use of this approach does not require strong ignorability assumptions. In what follows we describe the method and apply it to the PISA data assuming noninformative sampling, but later we also consider probability weighted estimation.

4.1 The sample distribution

Denote by $f_p(y_{ti} \mid \mathrm{x}_i)$ the *population pdf* for units in treatment group t under a strongly ignorable assignment process. We allow the assignment process to depend on known covariates v, some or all of which may be included in x. Denoting $\mathrm{z}=\mathrm{x} \cup \mathrm{v}$, we assume $f_p(y_{ti} \mid \mathrm{z}_i) = f_p(y_{ti} \mid \mathrm{x}_i)$ and $\Pr(T_i = t \mid y_{ti}, \mathrm{z}_i, i \in S)$ $= \Pr(T_i = t \mid y_{ti}, \mathrm{v}_i, i \in S)$. The *sample pdf* for unit i exposed to treatment t, given the covariates z_i, is obtained by Bayes theorem as,

(4.1) $$f_{S_t}(y_{ti} \mid \mathrm{z}_i) = f(y_{ti} \mid \mathrm{z}_i; T_i = t) = \frac{\Pr(T_i = t \mid y_{ti}, \mathrm{v}_i, i \in S) f_p(y_{ti} \mid \mathrm{x}_i)}{\Pr(T_i = t \mid \mathrm{z}_i, i \in S)},$$

where $\Pr(T_i = t \mid \mathrm{z}_i, i \in S) = \int \Pr(T_i = t \mid y_{ti}, \mathrm{v}_i, i \in S) f_p(y_{ti} \mid \mathrm{x}_i) dy_{ti}$.

Remark 5. It follows from (4.1) that the sample *pdf* is generally different from the population *pdf*, unless $\Pr(T_i = t \mid y_{ti}, \mathrm{v}_i, i \in S) = \Pr(T_i = t \mid \mathrm{z}_i, i \in S)$, in which case the assignment to treatments can be ignored for inference when conditioning on z.

Remark 6. The probabilities $\Pr(T_i = t \mid \mathrm{z}_i, i \in S)$ are *propensity scores*.

The sample *pdf* defined by (4.1) was shown in recent years to provide a valuable modeling approach for inference from complex sample surveys; see Pfeffermann and Sverchkov (2009) for review of studies that utilize the sample *pdf* for inference on generalized linear models, testing of distribution functions and prediction of finite population and small area means. The obvious distinction between survey sampling and observational studies is that in survey sampling the sample inclusion probabilities $\pi_i = \Pr(i \in S)$ are usually known, which enables estimating the probabilities $\Pr(i \in S \mid y_i, \mathrm{v}_i) = E_p(\pi_i \mid y_i, \mathrm{v}_i)$ and testing the informativeness of the sampling process; Pfeffermann and Sverchkov (2003, 2009). This is generally not the case in observational studies, requiring therefore modeling the parametric form of the probabilities $p_{ti} = \Pr(T_i = t \mid y_{ti}, \mathrm{v}_i, i \in S)$ in (4.1). As discussed below, modeling the sample *pdf* (4.1) allows estimating the unknown parameters governing the *pdf* $f_p(y_{ti} \mid \mathrm{x}_i)$ and the probabilities p_{ti}, and using them for estimating the ATE.

4.2 Estimating the parameters of the sample distribution

So far we suppressed for convenience in the notation the parameters governing the sample *pdf*. Adding the unknown parameters to the notation and assuming that the inclusion in the sample and the assignment to treatments are independent between units, and that the responses are likewise independent, the sample likelihood for treatment *t*, based on the sample S_t of size n_t takes the form,

(4.2) $$L_{S_t}[\alpha_t,\theta_t;\{y_{ti},\mathrm{z}_i\}]=\prod_{i=1}^{n_t}\frac{\Pr(T_i=t\mid y_{ti},\mathrm{v}_i,i\in S;\alpha_t)f_p(y_{ti}\mid \mathrm{x}_i;\theta_t)}{\Pr(T_i=t\mid \mathrm{z}_i,i\in S;\alpha_t,\theta_t)}.$$

Alternatively, the likelihood (4.2) can be replaced by the joint ('full') likelihood of the treatment selection and the response measurements defined as,

(4.3) $$\begin{aligned}&L_S[\alpha_t,\theta_t;\{y_{ti},i\in S_t;\mathrm{z}_j,j\in S\}]=\\&=\prod_{i=1}^{n_t}\Pr(T_i=t\mid y_{ti},\mathrm{v}_i,i\in S;\alpha_t)f_p(y_{ti}\mid \mathrm{x}_i;\theta_t)\prod_{\substack{j\in S\\ j\notin S_t}}[1-\Pr(T_j=t\mid \mathrm{z}_j,j\in S;\alpha_t,\theta_t)]\end{aligned}.$$

The likelihood (4.3) has the advantage of comprising the unconditional treatment assignment probabilities for units outside the sample S_t, thus using more information for estimating the model parameters. The full likelihood is often applied for handling informative nonresponse, see, e.g., Greenlees *et al*. (1982), Rotnitzky and Robins (1997), Gelman (2003, Ch.7) and Little (2004).

Replacing the unknown model parameters by their maximum likelihood estimates (*mle*) yields the estimates,

(4.4) $$\hat{f}_p(y_{ti}\mid \mathrm{x}_i)=f_p(y_{ti}\mid \mathrm{x}_i;\hat{\theta}_t);\;\; \hat{p}_{ti}=\Pr(T_i=t\mid y_{ti},\mathrm{v}_i,i\in S;\hat{\alpha}_t).$$

4.3 Estimation of population means

When the covariates x_i are known for every unit $i\in U$, then by (4.4), the population means, $\mu_t=\frac{1}{N}\sum_{i=1}^{N}E_p(Y_{ti}\mid \mathrm{x}_i)$, $t=0,1$, (and consequently the ATE, $\tau=\mu_1-\mu_0$), can be estimated by,

(4.5) $$\hat{\mu}_{t,p}=\frac{1}{N}\sum_{i=1}^{N}\hat{E}_p(Y_{ti}\mid \mathrm{x}_i;\theta_t)=\frac{1}{N}\sum_{i=1}^{N}E_p(Y_{ti}\mid \mathrm{x}_i;\hat{\theta}_t).$$

Note that if $E_p(Y_{ti}\mid \mathrm{x}_i;\theta_t)$ is linear, the computation of (4.5) only requires knowledge of the population means $\bar{X}=(\bar{X}_1,...,\bar{X}_k)'$.

When the covariates are unknown for units outside the sample, or the expectation is not linear, then as long as the sampling design is noninformative with respect to the distribution of the covariates, one can use instead the estimator,

(4.6) $$\hat{\mu}_{t,S} = \frac{1}{n}\sum_{i\in S}\hat{E}_p(Y_{ti} \mid \mathrm{x}_i;\theta_t).$$

Alternatively, one can use in this case the 'combined' estimator,

(4.7) $$\hat{\mu}_{t,C} = \frac{\sum_{i=1}^{n} T_i[Y_{ti} - \hat{E}_p(Y_{ti} \mid \mathrm{x}_i;\theta_t)]/\hat{p}_{ti}}{\sum_{i=1}^{n}(T_i/\hat{p}_{ti})} + \hat{\mu}_{t,S}.$$

Remark 7. The estimator (4.7) resembles the familiar GREG estimator in survey sampling (Särndal *et al*., 1992), and it looks similar to the 'doubly-robust' estimator (3.4). Notice, however, that (4.7) accounts for an informative assignment process as reflected by the use of the probabilities $\hat{p}_{ti} = \hat{P}r(T_i = t \mid y_{ti}, \mathrm{v}_i, i \in S)$ instead of the propensity scores $\hat{e}_{ti} = \hat{P}r(T_i = t \mid \mathrm{z}_i, i \in S)$. On the other hand, the estimator (4.7) does not posses a 'double robustness' property, since the unknown model parameters are estimated jointly from the likelihood (4.3).

The estimators (4.5) and (4.6) are functions of the *mle* $\hat{\theta}_t$, and hence their large sample variance can be estimated by use of the inverse of the estimated information matrix. Large sample properties of the combined estimator (4.7) and a consistent estimator of its variance can be derived by application of M-estimation theory. See Landsman (2008) for details. The estimated variances of all the three estimators account for the sample distribution of the responses given the observed covariates.

4.4 Application of new method to PISA data for Ireland

We assume a normal distribution for the potential population responses under strongly ignorable assignment and a logistic model for the assignment probabilities. The goodness of fit of the model is tested in Section 4.6.

(4.8) $$f_p(y_{ti} \mid \mathrm{x}_i) = N(\beta_{0t} + \mathrm{x}_i'\beta_t, \sigma_t^2);\ \Pr(T_i = t \mid y_{ti}, \mathrm{v}_i, i \in S) = \frac{\exp(\gamma_{0t} + \delta_t y_{ti} + \mathrm{v}_i'\gamma_t)}{1 + \exp(\gamma_{0t} + \delta_t y_{ti} + \mathrm{v}_i'\gamma_t)},\ t = 0,1.$$

When fitting the models to the two types of schools we found that the x-variables contain all the variables listed in Section 3.4 except for school location (S.loc), which was found to be highly insignificant in both the public and private school models. The v-variables contain all the variables listed in Section 3.4 except for mother's education

(M.E), which was found to be highly insignificant in both the public and private school models. As discussed in Section 5, the sample *pdf* (4.1) obtained in the normal/logistic case is identifiable and accommodates consistent and asymptotically normal (CAN) estimators for all the model parameters, if x and v differ in at least one covariate.

Results

We computed the *mle* of the unknown parameters by maximizing the likelihood (4.3) with respect to $\theta_t = (\beta_{0t}, \beta_t, \sigma_t^2)$ and $\alpha_t = (\gamma_{0t}, \delta_t, \gamma_t)$ $t = 0,1$. See Appendix C for the maximization procedure. Table 3 shows the estimates and standard errors (S.E.) of the model coefficients.

Table 3: Estimates and S.E. (in parenthesis) of model coefficients

Variable	**Private Schools**	**Public Scholls**
Assignment (logistic) model		
Const	- 2.95 (1.30)	13.88 (2.90)
δ_t	0.49 (0.21)	- 2.02 (0.39)
Gen.	0.77 (0.13)	- 0.76 (0.18)
S.E.I	- 0.12 (0.07)	0.40 (0.12)
H.E.R	3.16 (0.20)	- 2.57 (0.30)
S.E.S	0.09 (0.07)	0.27 (0.11)
S.loc	1.13 (0.13)	- 1.63 (0.24)
Population (normal) model		
σ_t	0.83 (0.02)	1.10 (0.07)
Const	6.09 (0.07)	6.89 (0.14)
Gen.	- 0.20 (0.05)	0.17 (0.08)
M.E	0.18 (0.05)	0.11 (0.07)
S.E.I	0.16 (0.03)	0.16 (0.04)
H.E.R	0.39 (0.09)	1.35 (0.20)
S.E.S	0.21 (0.02)	0.30 (0.04)

As anticipated, $\hat{\delta}_1 > 0$ and $\hat{\delta}_0 < 0$, but $\hat{\delta}_1$ is close to zero, although significant at the 5% level using the conventional *t*-statistic. On the other hand, $\hat{\delta}_0$ is highly negative, indicating that for given values of the covariates, the probability to attend a public

school decreases very rapidly as the proficiency score increases. This finding suggests that pupils attending public schools are generally less able, and their lower scores and not necessarily because of poor quality of the public schools. Another important result emerging from Table 3 is that the variable school location (S.loc) is highly significant even when including the response in the assignment models. We return to this result in Section 4.7. The coefficient of S.loc is positive for private schools and negative for public schools, suggesting that pupils from urban areas tend to enroll in private schools.

Table 4 shows the estimates of the population means by type of school and the corresponding estimates of the ATE. We present the two estimates defined in Section 4.3: the estimate $\hat{\mu}_{t,S}$ (Eq. 4.6) and the combined estimate $\hat{\mu}_{t,C}$ (Eq. 4.7).

Table 4. Estimation of population means and ATE

	Private School		**Public School**		**ATE**	
	$\hat{\mu}_{1,S}$	$\hat{\mu}_{1,C}$	$\hat{\mu}_{0,S}$	$\hat{\mu}_{0,C}$	$\hat{\tau}_S = \hat{\mu}_{1,S} - \hat{\mu}_{0,S}$	$\hat{\tau}_C = \hat{\mu}_{1,C} - \hat{\mu}_{0,C}$
Estimate	6.10	6.09	7.05	6.91	- 0.95	- 0.82
Std. Error	0.05	0.06	0.15	0.12	0.16	0.13

The two methods of estimating the population means yield similar estimates and the ATE estimates are therefore likewise similar, negative and highly significant, indicating the very interesting and important result that after accounting for the school selection by pupils, the mean score in the public schools is actually higher than in the private schools. A similar conclusion was reached in Section 3.4 by use of the LV and IV methods, although with smaller ATE estimates.

4.5 Probability weighted estimation under the proposed approach

We computed also PWE of the population means and the ATE under the proposed approach, which account for possible sampling effects, similarly to the estimators computed under the previous methods shown in Section 3.5. See Appendix B for the corresponding expressions. Table 5 shows the unweighted (UNWEI) and probability weighted estimators (WEI) of the school mean scores and the ATE. The results in this table reaffirm the results in Section 3.5 that the PISA sampling design is not informative for the models used for estimation of the ATE. We ignore the sample selection process in subsequent analysis.

Table 5. Unweighted and weighted estimators of schools mean score and ATE

Method	Private schools		Public schools		ATE	
	UNWEI	WEI	UNWEI	WEI	UNWEI	WEI
Est. Pop. regression	6.10	6.09	7.05	6.92	-0.95	-0.83
Combined estimator	6.09	6.08	6.91	6.80	-0.82	-0.72

4.6 Goodness of fit of the model fitted to the PISA data for Ireland

Assessing the goodness of fit of the model (4.8), or more generally, any other model of the form (4.1) seems formidable on first sight since no observations are available from the population distribution under strong ignorability, and the assignment probabilities are generally unknown. Note, however, that once the identifiability of the sample *pdf* has been established (see Section 5), one faces the classical problem of having a random sample from a hypothesized *pdf* that needs to be tested. Below we consider three tests, which compare the theoretical and empirical sample distributions of the responses and apply them to the PISA data. The power of the tests is studied further in Appendix D by using simulated data sets.

Goodness of fit tests

Suppose first that the true model parameters $\{\alpha_t, \theta_t\}$ are known. Denote by $U_{ti}(y) = F_{ti}(y \mid \mathrm{z}_i) = \int_{-\infty}^{y} f_{S_t}(y_{ti} \mid \mathrm{z}_i; \alpha_t, \theta_t) dy_{ti}$ the hypothesized *cdf* of $Y_{ti} \mid \mathrm{z}_i$, $i = 1, 2, ..., n_t$. For continuous F_{ti}, the random variables $U_{ti}(y)$ evaluated at the outcomes y_{ti} are independent Uniform [0,1] variables since the responses Y_{ti} are independent given the covariates z_i. Let $u_{t1}, ..., u_{tn_t}$ denote the values of $U_{t1}, ..., U_{tn_t}$ computed at the sample values $y_{t1}, ..., y_{tn_t}$ and let $F_{t,\mathrm{EMP}}$ define the empirical distribution of $u_{t1}, ..., u_{tn_t}$. The following goodness of fit tests are in common use, where $u_{t(1)}, ..., u_{t(n_t)}$ are the ordered values of $u_{t1}, ..., u_{tn_t}$ (Stephens, 1986):

(4.9) Kolmogorov - Smirnov: $KS_t = \max_i | F_{t,\mathrm{EMP}}(u_{t(i)}) - u_{t(i)} |$,

(4.10) Cramer - von Misses: $CM_t = \frac{1}{12n_t} + \sum_{i=1}^{n_t} \left[u_{t(i)} - \frac{2i-1}{2n_t} \right]^2$,

(4.11) Anderson - Darling: $AD_t = -n_t - \frac{1}{n_t}\sum_{i=1}^{n_t}[(2i-1)\ln(u_{t(i)}) + (2n_t + 1 - 2i)\ln(1-u_{t(i)})]$.

It is known (e.g., Babu and Feigelson, 2006) that KS is sensitive to large-scale differences in location and shape between the model and the empirical distribution, CM is sensitive to small-scale differences in the shape and AD is sensitive to differences near the tails of the distribution. All the three test statistics are *distribution-free* as long as the hypothetical *cdf* is fully specified (known parameters).

When computed with estimated parameters, the asymptotic distribution of the three statistics depends in a complex way on the true underlying *cdf*, and possibly also on the method of estimation. Correct critical values can be obtained in this case by use of parametric bootstrap. The procedure consists of generating a large number of samples from the estimated hypothesized model, re-estimating the unknown model parameters from each bootstrap sample and then computing the corresponding test statistics. The bootstrap distribution of the test statistics approximates the true distributions under the hypothesized model with correct order of error (Babu and Rao, 2004).

Validating the model fitted to the PISA data for Ireland

As explained above, the critical values for the distribution of the test statistics can be computed from the bootstrap distribution. To this end we generated 250 bootstrap samples for each type of school $(t = 0, 1)$ by first generating new outcomes Y_{ti} from the estimated normal population model $f_p(y_{ti} \mid \mathrm{x}_i; \hat{\beta}_t, \hat{\sigma}_t^{\,2})$, using the same covariates as for the actual sample, and then selecting the units to the sample S_t using the estimated logistic probabilities $\Pr(T_i = t \mid \mathrm{v}_i, y_{ti}; \hat{\delta}_t, \hat{\gamma}_t)$. Notice that this way the sample sizes in the two groups are no longer constant. We found that for the treatment group the mean sample size is 1245 with standard deviation 17, and for the control group the mean sample size is 700 with standard deviation 18. (The sample sizes in the true data set are 1244 and 683 respectively). Next we computed the *mle* of the parameters $\beta_t, \sigma_t^{\,2}, \gamma_t, \delta_t$ for each bootstrap sample and the test statistics (4.9)- (4.11).

Table 6 shows the values of the three test statistics for the PISA samples and their p-values, as computed from the corresponding bootstrap distributions. As can be

seen, all the three statistics are non-significant with p-values higher than 10%, thus supporting the use of the selected models.

Table 6. Goodness of fit test statistics and p-values (in parenthesis)

Private schools			**Public schools**		
KS	**CM**	**AD**	**KS**	**CM**	**AD**
0.023 (0.12)	0.089 (0.18)	0.62 (0.11)	0.027 (0.17)	0.062 (0.32)	0.45 (0.15)

Remark 8. Other than testing the model, we also computed the empirical means and standard deviations over the 250 bootstrap samples of the estimates of the model coefficients and all the ATE estimates considered in Sections 3 and 4. To save in space we do not show the detailed results but the means are generally close (and in most cases very close) to the values computed for the PISA data, and likewise for the standard deviations. These results indicate that the model coefficients can be estimated almost unbiasedly and with acceptable standard error estimates, despite the complicated structure of the sample model. Obtaining similar ATE estimates under the different methods to the estimates computed from the PISA data is another indication of the goodness of fit of the models fitted to this data set.

4.7 Testing the assumptions underlying existing methods

As mentioned earlier, the use of the proposed approach enables testing some of the assumptions underlying existing methods under the sample model fitted to the observed data. Consider first the logistic models for the assignment probabilities. The coefficients $\hat{\delta}_t$ of the response are significant in both models, with $\hat{\delta}_0$ in particular being highly negative. This result suggests that the covariates available for the present study are not sufficient to explain the choice of school, and hence that methods that assume strong ignorability (Assumption SI(a)) and in particular, methods that employ propensity scores computed with these covariates are not adequate. Notice in Table 1 that the use of these methods yields positive ATE estimates, although of lower magnitude than the crude difference, $\bar{y}_1 - \bar{y}_0$.

Next consider the IV method. In its simple form it assumes the model $Y_t = \mathrm{x}'\beta_t + u_t$; $E_p(u_t \mid \mathrm{x}) = 0$, with three added conditions on the instrument *h*: IV(a)- $E_p(Y_t \mid \mathrm{x}, h) = E_p(Y_t \mid \mathrm{x})$, IV(b)- $E_p[T(u_1 - u_0) \mid \mathrm{x}, h] = 0$, and IV(c)- $\Pr(T = 1 \mid \mathrm{x}, h)$

$\neq \Pr(T=1\,|\,\mathrm{x})$. The population model with the covariates x listed in Section 3.4 is validated in Section 4.6. Furthermore, our analysis shows that the instrument S.loc is highly insignificant in the two population models, thus supporting the condition IV(a). Condition IV(b) cannot be verified empirically but this condition is generally considered as being mild and it can be relaxed further (Wooldridge, 2002). Finally, the condition IV(c) is verified as well since the coefficients of the instrument in the models fitted for the assignment probabilities are highly significant (Table 3), even when including the response y as an additional explanatory variable. Indeed, the use of the IV method yields an ATE estimate of -0.61 (Table 1), which is the closest to the estimate obtained under our approach among the other methods considered.

Remark 9. We emphasize again that all the above conclusions are under the model that we have fitted (and validated) to the data.

5. FOUNDATION OF PROPOSED APPROACH

5.1 Identification of the sample distribution

An important question underlying the use of the sample *pdf* (4.1) is model identification. In order to get some insight into this issue we restrict to a given treatment t and hence drop for convenience the subscript t everywhere, denoting by $p(y,\mathrm{v};\alpha)=\Pr(i\in S_t\,|\,y_t,\mathrm{v};\alpha)$ the probability assignment rule (*PAR*) to the sample S_t, and by $f_p(y\,|\,\mathrm{x};\theta)$ the corresponding population *pdf* of $Y_t\,|\,\mathrm{x}$ under strong ignorability. We assume that the response is continuous. Let $J\subseteq R$ define the common domain of the y-values for these functions. The sample *pdf* for units in S_t is therefore $f_{S_t}(y\,|\,\mathrm{x},\mathrm{v};\theta,\alpha)=\dfrac{f_p(y\,|\,\mathrm{x};\theta)p(y,\mathrm{v};\alpha)}{\int f_p(y\,|\,\mathrm{x};\theta)p(y,\mathrm{v};\alpha)dy}$ and the identifiability of the sample *pdf* is defined as follows:

<u>*Definition* 1</u>. The sample *pdf* $f_{S_t}(y\,|\,\mathrm{x},\mathrm{v};\theta,\alpha)$ is identifiable if no different (proper) densities $f_p^{(1)}(y\,|\,\mathrm{x};\theta^{(1)})$, $f_p^{(2)}(y\,|\,\mathrm{x};\theta^{(2)})$ and different *PAR*s $p^{(1)}(y,\mathrm{v};\alpha^{(1)})$, $p^{(2)}(y,\mathrm{v};\alpha^{(2)})$ exist such that the pairs $[f_p^{(1)}(y\,|\,\mathrm{x};\theta^{(1)}),p^{(1)}(y,\mathrm{v};\alpha^{(1)})]$ and $[f_p^{(2)}(y\,|\,\mathrm{x};\theta^{(2)}),p^{(2)}(y,\mathrm{v};\alpha^{(2)})]$ induce the same sample *pdf* for every $y\in J$ and every set of covariates (x,v).

Clearly, if different pairs $[f_p^{(1)}, p^{(1)}],[f_p^{(2)}, p^{(2)}]$ yield the same sample *pdf*, the model is not identifiable. At first thought, this would seem to be always the case since (4.1) is the same *pdf* for the pair $[f_p^{(1)}(y \mid \mathrm{x};\theta^{(1)}), p^{(1)}(y,\mathrm{v};\alpha^{(1)})]$, and when the population *pdf* is $f_p^{(2)}(y \mid \mathrm{x},\mathrm{v};\theta^{(1)},\alpha^{(1)}) = \dfrac{f_p^{(1)}(y \mid \mathrm{x};\theta^{(1)})p^{(1)}(y,\mathrm{v};\alpha^{(1)})}{\int f_p^{(1)}(y \mid \mathrm{x};\theta^{(1)})p^{(1)}(y,\mathrm{v};\alpha^{(1)})dy}$ and the units are assigned with probabilities that do not depend on y given v (ignorable assignment). The *pdf* $f_p^{(2)}(y \mid \mathrm{x},\mathrm{v};\theta^{(1)},\alpha^{(1)})$, however, does not generally belong to a conventional parametric family and is very different from $f_p^{(1)}(y \mid \mathrm{x};\theta^{(1)})$, especially when the assignment mechanism is strongly informative. Hence, field experts should be able to decide which of the two *pdf*s, $f_p^{(1)}(y \mid \mathrm{x};\theta^{(1)})$ or $f_p^{(2)}(y \mid \mathrm{x},\mathrm{v};\theta^{(1)},\alpha^{(1)})$ is a more sensible population *pdf* for the potential responses in a given problem.

Conditions for model identification

Lemma 1 defines different conditions under which the sample *pdf* is identifiable. We assume for convenience that there are no covariates, but all the results generally hold when covariates x,v exist. Define $R_p(y;\theta^{(1)},\theta^{(2)}) = \dfrac{f_p^{(2)}(y;\theta^{(2)})}{f_p^{(1)}(y;\theta^{(1)})}$. We assume throughout this section that the functions $f^{(j)}(y;\theta^{(j)})$ and $p^{(j)}(y;\alpha^{(j)})$, $j=1,2$ are strictly positive on $J^* \subseteq J$.

<u>*Lemma* 1</u>.

(a) Suppose that $J^* = [c,\infty)$ for some constant c. If $f_p^{(1)}(y;\theta^{(1)})$ and $f_p^{(2)}(y;\theta^{(2)})$ are two different *pdf*s satisfying for any given $\theta^{(1)},\theta^{(2)}$, $\lim_{y\to\infty} R_p(y;\theta^{(1)},\theta^{(2)}) = 0,\infty$ or does not exist, then there are no different PARs $p^{(1)}(y;\alpha^{(1)})$, $p^{(2)}(y;\alpha^{(2)})$ with finite positive limits as $y \to \infty$ yielding the same sample pdf for all $y \in J^*$.

(*b*) Suppose that $J^* = (-\infty,c]$ for some constant c. If $f_p^{(1)}(y;\theta^{(1)})$ and $f_p^{(2)}(y;\theta^{(2)})$ are two different *pdf*s satisfying for any given $\theta^{(1)},\theta^{(2)}$, $\lim_{y\to-\infty} R_p(y;\theta^{(1)},\theta^{(2)}) = 0,\infty$ or does not exist, then there are no different PARs $p^{(1)}(y;\alpha^{(1)})$, $p^{(2)}(y;\alpha^{(2)})$ with finite positive limits as $y \to -\infty$ yielding the same sample *pdf* for all $y \in J^*$.

(*c*). Let y_0 be a limit point of J^*. If $f_p^{(1)}(y;\theta^{(1)})$ and $f_p^{(2)}(y;\theta^{(2)})$ are two different *pdf*s satisfying for any given $\theta^{(1)},\theta^{(2)}$, $\lim_{y\to y_0^+\,(y\to y_0^-)} R_p(y;\theta^{(1)},\theta^{(2)}) = 0,\infty$ or does not exist, then there are no different PARs $p^{(1)}(y;\alpha^{(1)})$, $p^{(2)}(y;\alpha^{(2)})$ with finite positive limits at $y = y_0$ yielding the same sample *pdf* for all $y \in J^*$.

Proof. Part (*a*) is similar to Lee and Berger (2001) and is proved in Appendix E. The proofs of the other two parts are similar.

Lemma 1 enables verifying the identifiability of the sample *pdf* for many combinations of population *pdf*s and PARs, but there are cases that need to be studied separately. For example, the Lemma is not applicable to the case of normal population *pdfs* and logistic PARS if the coefficients of y in the two logistic distributions are allowed to have opposite signs. This is because in this case one of the PARS will have a limit of zero as $y \to \infty$ or $y \to -\infty$, and the other PAR will have a limit of 1 (but see Result 1 below).

Further model identification results

Result 1 states the identifiability of the sample *pdf* resulting from the combination of a normal population *pdf* and a logistic *PAR*. The proof is given in Appendix F. Landsman (2008) considers other combinations of population *pdf*s and *PAR*s.

Result 1. No different pairs $[f_p(y \mid \mathrm{x};\theta),\ p(y,\mathrm{v};\alpha)]$ of a normal *pdf* and logistic *PAR* yield the same sample *pdf*, if the vectors x and v differ in at least one covariate.

Remark 10. The condition on the covariates seems to impose a limitation on the model but in practice there is no reason why the covariates used to model the response under strong ignorability should be the same as the covariates used to model the treatment assignment probabilities. See the models in Section 4.4.

5.2 Practical identifiability

Section 5.1 and the additional results in Landsman (2008) establish the 'theoretical identifiability' of the sample model under a large number of plausible combinations of population *pdf*s and *PAR*s. It is important to mention, however, that identifiability problems may arise in practice, depending on the forms of $f_p(y \mid \mathrm{x};\theta)$ and $p(y,\mathrm{v};\alpha)$. For example, Lee and Berger (2001) consider the case $f_p^{(1)}(y) = N(0,1)$,

$p^{(1)}(y) = \Phi(y-1)$. The authors show graphically that in this case the sample density $f_S(y) = f_p^{(1)}(y)p^{(1)}(y) / \int f_p^{(1)}(y)p^{(1)}(y)dy$ can be closely approximated by the normal density $N(0.92, 0.79^2)$. This means that even though the sample *pdf* $f_S(y)$ is theoretically identifiable (Landsman, 2008), a problem may arise in practice when fitting the model, distinguishing between this density and the sample density obtained from $f_p^{(2)}(y) = N(0.92, 0.79^2)$, $p^{(2)}(y) = const$. Lee and Berger (2001) refer to this phenomenon as 'practical non-identifiability'. Another example is where the *PAR* is logistic. Suppose that $p^{(1)}(y) = \{1 + [\exp(1+y)]^{-1}\}^{-1}$, $p^{(2)}(y) = \{1 + [\exp(-1-y)]^{-1}\}^{-1}$. Then, the pairs $[f_p^{(1)}(y) = N(0,1),\ p^{(1)}(y)]$ and $[f_p^{(2)}(y) = N(1,1),\ p^{(2)}(y)]$ induce the same sample *pdf* ("theoretical non-identifiability"), and this *pdf* is closely approximated by $N(0.28; 0.93^2)$ ("practical non-identifiability").

We emphasize that in the presence of covariates, if the vectors x and v differ in at least one covariate, the problem of practical Identifiability will generally disappear. See Landsman (2008) for a detailed analysis.

5.3 Asymptotic properties of the maximum likelihood estimators

The *mle* $\{\hat{\alpha}_t, \hat{\theta}_t, t = 0,1\}$ obtained by maximization of (4.3) are the solutions of the estimating equations $\sum_{k=1}^{n} u_{t,k} = 0$, where $u_{t,k}$ is the vector of the first derivatives of the *k*-th component of the log-likelihood with respect to (α_t, θ_t). Rotnitzky and Robins (1997) show that no $\sqrt{n}$-consistent and asymptotically normal (CAN) estimator of θ_t exists if (and only if) the derivatives of the log-likelihood with respect to θ_t are collinear with the derivatives of the log-likelihood with respect to α_t with probability 1. The derivatives are evaluated at the true parameter values. The authors illustrate that if the population model is $Y_{ti} \sim N(\beta_t, 1)$ and $\Pr(T_i = t \mid y_{ti}) = \dfrac{\exp(\gamma_{0t} + \delta_t y_{ti})}{1 + \exp(\gamma_{0t} + \delta_t y_{ti})}$, then no CAN estimator for β_t exists if in truth $\delta_t = 0$ (ignorable assignment).

Landsman (2008) uses the result of Rotnitzky and Robins (1997) and shows that if $f_p(y_{ti} \mid \mathrm{x}_i) = N(\beta_{0t} + \mathrm{x}_i'\beta_t, \sigma_t^2)$, $\mathrm{logit}[\Pr(T_i = t \mid y_{ti}, \mathrm{v}_i)] = \gamma_{0t} + \delta_t y_{ti} + \mathrm{v}_i'\gamma_t$ and $\mathrm{x} = \mathrm{v}$, then there exists no CAN for β_{0t} if in fact the true assignment probabilities are equal

($\delta_t = 0, \gamma_t = 0$). However, if x and v differ in at least one covariate, the derivatives of the log-likelihood (4.3) with respect to $\theta_t = (\beta_{0t}, \beta_t, \sigma_t)$ are not collinear with the derivatives with respect to $\alpha_t = (\gamma_{0t}, \gamma_t, \delta_t)$ with probability 1, and hence CAN estimators for the parameters exist even when the true assignment process is ignorable. This property enables testing the ignorability of the assignment process, as illustrated in Sections 4.4 and 4.7.

6. SUMMARY REMARKS

In this article we study the use of an alternative approach for observational studies that recovers the treatment assignment model and the population model under strong ignorability from the sample data. It is shown that the sample model holding for the observed data, which incorporates the two models, is identifiable under some general conditions. Furthermore, the goodness of fit of the sample model can be tested successfully by standard test statistics because the sample model refers to the observed data. As illustrated in Section 4.7, the sample model enables also testing the appropriateness of the use of some of the existing methods for a particular data set.

We applied the new approach for comparing the proficiency in mathematics of children aged 15 between public and private schools in Ireland. Our analysis shows that although the average score of pupils in the sample from private schools is significantly higher than the average score of pupils from public schools, the picture is reversed once the effect of the school selection is accounted for properly. A similar conclusion is reached by the methods of latent variables and instrumental variables.

The approach proposed in this article is fully parametric, which raises questions of its robustness to departures from the models fitted to the data. We emphasize again that the models can be tested and as the empirical illustrations show, the test statistics that we have applied have good power properties. Nonetheless, it is certainly worth considering the use of semi-parametric or nonparametric models either for the population models under strong ignorability and/or for the assignment probabilities, thus further robustifying the inference.

APPENDIX A: IMPUTATION OF PROFICIENCIES IN THE PISA STUDY

The multiple imputation approach draws at random multiple values from the conditional distribution of the unobserved proficiency, ψ_j of pupil j, given the m observed responses $d_j = (d_{1j},...,d_{mj})$ and covariates x_j representing individual background characteristics. The conditional *pdf* of ψ_j is expressed as,

$$f(\psi_j \mid d_j, \mathrm{x}_j) \propto \prod_{i=1}^{m} [\Pr(D_{ij} = 1 \mid a_i, b_i, \psi_j)]^{d_{ij}} [\Pr(D_{ij} = 0 \mid a_i, b_i, \psi_j)]^{(1-d_{ij})} f(\psi_j \mid \mathrm{x}_j, \lambda, \sigma) \text{ ,(A1)}$$

where $f(\psi_j \mid \mathrm{x}_j, \lambda, \sigma^2)$ is the normal distribution with mean $\mathrm{x}_j'\lambda$ and variance σ^2, and $\Pr(D_{ij} = 1 \mid a_i, b_i, \psi_j) = [1 + \exp(-a_i(\psi_j - b_i)]^{-1}$. The parameter a_i measures how question i distinguishes between pupils and the parameter b_i represents the 'difficulty' of question i. The responses to the various questions are assumed to be independent given the parameters (a_i, b_i, ψ_j). Five imputed values of ψ_j are drawn for every student j in the sample and stored in the PISA database.

APPENDIX B: PROBABILITY WEIGHTED ESTIMATORS OF ATE

Consider first the crude difference $\hat{\tau}_D = \bar{y}_1 - \bar{y}_0$. A simple modification to account for possible sampling effects is to replace it by the Horvitz-Thompson (1952) estimate $\hat{\tau}_{\text{D,H-T}} = (\frac{1}{N_1}\sum_{i \in S_1} w_i y_{1i} - \frac{1}{N_0}\sum_{j \in S_0} w_j y_{0j})$, where the w_i's are the sampling weights. This estimator is randomization-unbiased and consistent for $\tau_{\text{D}} = \bar{Y}_{1,p} - \bar{Y}_{0,p}$ over repeated sampling, where $\bar{Y}_{t,p}$, $t = 0,1$ are the corresponding finite population means. Alternatively, one can use the Brewer-Hajek (B-H) estimator, $\hat{\tau}_{\text{D,B-H}} = (\sum_{i \in S_1} w_i y_{1i} / \sum_{i \in S_1} w_i) - (\sum_{j \in S_0} w_j y_{0j} / \sum_{j \in S_0} w_j)$. This estimator is only approximately randomization-unbiased but it is known to have in general a smaller variance than the H-T estimator and it does not require knowledge of the population sizes N_t, $t = (0,1)$.

<u>*Regression estimate*</u>: The regression estimator is $\hat{\tau}_{\text{OLS}} = \bar{\mathrm{x}}'(\hat{\beta}_{1,\text{OLS}} - \hat{\beta}_{0,\text{OLS}})$ (Eq. 3.1). This estimator assumes that the distribution of $Y_t \mid \mathrm{x}$ in the sample is the same as in the population, such that the OLS estimators are unbiased under the regression

model, and that the simple means of the covariates are randomization-unbiased for the corresponding population means. A simple modification is to replace the estimator by the PWE $\hat{\tau}_w = \hat{\bar{\mathrm{x}}}'_w(\hat{\beta}_{1,PW} - \hat{\beta}_{0,PW})$, where $\hat{\bar{\mathrm{x}}}'_w$ is the H-T or the B-H estimator of the population mean of the covariates and $\hat{\beta}_{t,PW} = [\sum_{i\in S^t} w_i \mathrm{x}_i \mathrm{x}'_i]^{-1} \sum_{i\in S^t} w_i \mathrm{x}_i y_{ti}$ is the PWE of the 'census' regression coefficients $B_{t,p} = [\sum_{j\in U_t} \mathrm{x}_j \mathrm{x}'_j]^{-1} \sum_{j\in U_t} \mathrm{x}_j y_{tj}$. Notice that even if the sampling process is noninformative for the distribution of $Y_t \mid \mathrm{x}$, thus validating the use of the OLS coefficients, the estimator $\hat{\tau}_{\mathrm{OLS}}$ will still be biased if the sample mean of the covariates is not randomization-unbiased for the true population means.

<u>*B-H, DR and matching estimates*</u>. Three other estimators that assume strongly ignorable assignment considered in Section 3 are the Brewer-Hajek (B-H) estimator, the doubly-robust (DR) estimator and the estimator based on matching. The first two estimators use the estimated propensity scores. Suppose that the estimates are obtained by fitting the logistic function. In order to account for possible sampling effects, we estimate the vector coefficients by solving the weighted likelihood equations, $\sum_{i\in S} w_i[T_i - e(\mathrm{x}_i)]\mathrm{x}_i = 0$, where $e(\mathrm{x}_i)$ is the logistic function. Denote the resulting estimated propensity scores by $\hat{e}_w(\mathrm{x}_i)$. The PWE of the ATE are,

$$\hat{\tau}_{\text{B-H},w} = \left(\sum_{i=1}^n w_i \frac{T_i}{\hat{e}_w(\mathrm{x}_i)}\right)^{-1} \sum_{i=1}^n \frac{w_i T_i y_i}{\hat{e}_w(\mathrm{x}_i)} - \left(\sum_{i=1}^n \frac{w_i(1-T_i)}{(1-\hat{e}_w(\mathrm{x}_i))}\right)^{-1} \sum_{i=1}^n \frac{w_i(1-T_i)y_i}{(1-\hat{e}_w(\mathrm{x}_i))},$$

$$\hat{\tau}_{\mathrm{DR},w} = (\sum_{i=1}^n w_i)^{-1}\{\sum_{i=1}^n \frac{w_i T_i y_i - w_i[T_i - \hat{e}_w(\mathrm{x}_i)]\hat{r}_{1,w}(\mathrm{x}_i)}{\hat{e}_w(\mathrm{x}_i)} - \sum_{i=1}^n \frac{w_i(1-T_i)y_i + w_i[T_i - \hat{e}_w(\mathrm{x}_i)]\hat{r}_{0,w}(\mathrm{x}_i)}{1-\hat{e}_w(\mathrm{x}_i)}\},$$

where $\hat{r}_{t,w}(\mathrm{x}_i)$ is the PWE of $r_t(\mathrm{x}_i) = E_p(Y_{ti} \mid \mathrm{x}_i)$.

The computation of a PWE based on matching is tricky because for a small sample size or a small number of matches, the matched values in the sample for a given sampled unit may not be representative of the potential matches in the population. More generally, the process of searching for the closest matches is not captured by the randomization distribution. The following PWE is computed by the Stata package *nnmatch* (Abadie *et al*. 2001). Denote as before by $J_{iM}(t)$ the set of

indices of the closest matches in S_t for unit $i \in S_{1-t}$ and define $\hat{y}_{(1-t),i} = y_{(1-t),i}$ and $\hat{y}_{ti} = \sum_{j \in J_{iM}(t)} w_j y_{tj} / \sum_{j \in J_{iM}(t)} w_j$. The PWE of the ATE is $\hat{\tau}_{\mathrm{M},w} = \sum_{i=1}^{n} w_i (\hat{y}_{1i} - \hat{y}_{0i}) / \sum_{i=1}^{n} w_i$.

Remark 11. The distance $d_M(i)$ from x_i to the M nearest matches is also modified to ensure that the sum of the weights of the matches adds to M.

<u>*LV and IV methods*</u>. A PWE under the LV method can be computed as follows. First, fit the probit model for the assignment variable by maximizing the PW log-likelihood $\sum_{i \in S} w_i \{T_i \log[p(\mathrm{v}_i)] + (1 - T_i)\log[1 - p(\mathrm{v}_i)]\}$, where $p(\mathrm{v}_i) = \Phi(-\mathrm{v}_i'\alpha)$ with $\Phi(\cdot)$ defining the standard normal cumulative distribution. Denote by $\hat{\alpha}_w$ the PWE obtained that way and let $K_{0,w}(\mathrm{v}) = -\dfrac{\phi(\mathrm{v}'\hat{\alpha}_w)}{1 - \Phi(\mathrm{v}'\hat{\alpha}_w)}$; $K_{1,w}(\mathrm{v}) = \dfrac{\phi(\mathrm{v}'\hat{\alpha}_w)}{\Phi(\mathrm{v}'\hat{\alpha}_w)}$. Next compute the PWE $\hat{\beta}_{t,\mathrm{LV}w}$ and $\hat{C}_{t,LVw}$ of β_t, C_t in the regression models $E_{S_t}(Y_t \mid \mathrm{x}, \mathrm{v}) = \mathrm{x}'\beta_{t,\mathrm{LV}} + C_{t,\mathrm{LV}} K_{t,w}(\mathrm{v})$, $t = 0,1$. Estimate $\hat{\tau}_{\mathrm{LV},w} = \hat{\bar{\mathrm{x}}}_w'(\hat{\beta}_{1,\mathrm{LV}w} - \hat{\beta}_{0,\mathrm{LV}w})$.

A PWE under the IV method is computed as follows. First fit the logistic model with x and h as the covariates, similarly to the computation of the weighted propensity scores used for the B-H and DR estimators. Denote by $\hat{g}_{w,i} = \hat{g}_w(\mathrm{x}_i, h_i)$ the corresponding PWE. Next compute the PWE $\hat{\theta}_{\mathrm{IV}w} = (\sum_{i=1}^{n} w_i \hat{z}_i \tilde{\mathrm{x}}_i')^{-1} \sum_{i=1}^{n} w_i \hat{z}_i y_i$, where $\hat{z}_i' = [\hat{g}_{w,i}\mathrm{x}_i', (1 - \hat{g}_{w,i})\mathrm{x}_i']$. The ATE estimator is $\hat{\tau}_{\mathrm{IV},w} = \hat{\bar{\mathrm{x}}}_w'(\hat{\beta}_{1,\mathrm{IV}w} - \hat{\beta}_{0,\mathrm{IV}w})$, where $\hat{\beta}_{t,\mathrm{IV}w}$ is the PWE of β_t as defined from $\hat{\theta}_{\mathrm{IV}w}$, $t = 0,1$.

<u>*New method*</u>. Instead of maximizing the likelihood (4.3), we maximize the 'pseudo likelihood' $L_{S,w}[\alpha_t, \theta_t; \{y_{ti}, i \in S_t; z_j, j \in S\}] =$

$$= \prod_{i=1}^{n_t} [\Pr(i \in S_t \mid y_{ti}, \mathrm{v}_i, i \in S; \alpha_t)]^{w_i} [f_p(y_{ti} \mid \mathrm{x}_i; \theta_t)]^{w_i} \prod_{\substack{j \in S \\ j \notin S_t}} [1 - \Pr(j \in S_t \mid z_j, j \in S; \alpha_t, \theta_t)]^{w_j},$$

yielding the pseudo maximum likelihood estimates $(\hat{\theta}_{t,w}, \hat{\alpha}_{t,w})$. Denote by $\hat{p}_{ti,w} = \Pr(i \in S_t \mid y_{ti}, \mathrm{v}_i, i \in S; \hat{\alpha}_{t,w})$ the PWE of the treatment assignment probability p_{ti}. When the covariates are known for all the population units, estimate, $\hat{\tau}_{p,w} = \frac{1}{N}\sum_{i=1}^{N} E_p(Y_{1i} \mid \mathrm{x}_i; \hat{\theta}_{1,w}) - \frac{1}{N}\sum_{i=1}^{N} E_p(Y_{0i} \mid \mathrm{x}_i; \hat{\theta}_{0,w})$. Otherwise, one can estimate,

$\tilde{\tau}_{p,w} = \tilde{\mu}_{1p,w} - \tilde{\mu}_{0p,w}$, where $\tilde{\mu}_{tp,w} = \sum_{i \in S_t} [w_i E_p(Y_{ti} \mid \mathrm{x}_i ; \hat{\theta}_{t,w}) / \hat{p}_{ti,w}] / \sum_{i \in S_t} (w_i / \hat{p}_{ti,w})$, or use the weighted combined estimator $\hat{\tau}_{C,w} = \hat{\mu}_{1C,w} - \hat{\mu}_{0C,w}$, where

$$\hat{\mu}_{tC,w} = \frac{\sum_{i \in S_t} w_i [y_{ti} - \hat{E}_p(Y_{ti} \mid \mathrm{x}_i ; \hat{\theta}_{t,w})] / \hat{p}_{ti,w}}{\sum_{i \in S_t} (w_i / \hat{p}_{ti,w})} + \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j \hat{E}_p(Y_{tj} \mid \mathrm{x}_j ; \hat{\theta}_{t,w}) \text{ for } t = 0,1 .$$

APPENDIX C: MAXIMIZATION OF THE LIKELIHOOD (4.3)

The mle in Table 3 were computed by using the maximization routine *nlm* in R (Development Core Team, 2004). The choice of the initial values is crucial for the convergence of the maximization algorithm. However, empirical investigation shows that for given value of the coefficient δ_t of y_{ti} in the logistic model the maximization is not sensitive to the choice of the initial values for the other parameters. We applied therefore the following algorithm:

1. Define a grid of plausible values for δ_t around zero. Maximize the likelihood for each value δ_t with respect to the other parameters using as initial values for β_t and σ_t^2 the values obtained by fitting a linear model to the sample data, and zeroes for γ_{0t} and γ_t. The parameters maximizing the likelihood over all the grid values of δ_t are taken as the initial values.
2. Maximize the likelihood with respect to all the parameters (including δ_t) using as initial values the values obtained in Step 1.

APPENDIX D: POWERS OF GOODNESS OF FIT TEST STATISTICS

In Section 4.6 we considered three goodness of fit test statistics and applied them for testing the model (4.8) fitted to the PISA data. In order to study the powers of the three test statistics in the case of misspecified population models we simulated new data sets for each of the two groups (public and private schools) from the same models as before, except that the residual terms in the two population models were generated from the skew *t*-distribution defined by Azzalini and Capitanio (2003). The true population means and hence the ATE remain unchanged. The skew *t*-distribution depends on four parameters: location (ξ), scale (w), shape (α) and degrees of freedom (υ). The normal and *t*-distributions are members of this family of

distributions. For example, the case $\xi = 0, w = 1, \alpha = 0, \upsilon = \infty$ defines the standard normal distribution.

We generated 100 sets of residuals for each group from the following 3 distributions: (A) $\xi = 0, w = 1, \alpha = 0, \upsilon = 6$, (B) $\xi = -1.16, w = 1.55, \alpha = 2.5, \upsilon = \infty$, (C) $\xi = -1.24, w = 1.45, \alpha = 2.5, \upsilon = 6$. Distribution A defines the standard *t*-distribution with 6 degrees of freedom. Distribution B defines a skewed distribution with relatively short tails, while Distribution C defines a skewed distribution with a heavy right tail. The three densities are plotted in Figure 1. The location parameters were chosen in such a way that all the three distributions have mean zero, implying that the true ATEs are the same. The standard deviations equal 1.22, 1.04 and 1.27 respectively. Next we fitted the models that assume that the residuals are normal (Equation 4.8), such that the true models are misspecified.

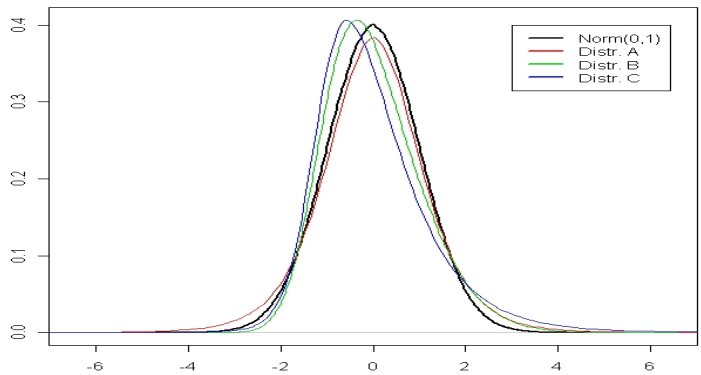

Figure 1. Densities of residuals under the four alternative distributions

Table D1 shows the percentage of samples that the misspecified models have been rejected by the three test statistics when using the conventional 10%, 5%, 2.5% and 1% significance levels. The percentages estimate the power of the tests. As can be seen, all the tests reject the three misspecified models for the private schools in literally all the samples, and the powers of the Cramer-von Misses test, and in particular, the powers of the Anderson-Darling test are acceptable for the three misspecified models for the public schools as well. We mention in this regard that the three misspecified distributions were chosen to be sufficiently close to the normal distribution (see Figure 1). Further empirical results not shown indicate that for only mild larger distortions from the normal distribution, the powers of the three tests for the public schools are likewise close to 1.

Table D1 also shows the average of the estimates of the mean score in each group and the ATE, and the corresponding standard errors of the estimates (in parenthesis), over the 100 data sets. The empirical averages deviate now significantly from the corresponding true values under all the three misspecified models and with larger Standard errors than under the correct model. However, the estimates of the ATE are still highly negative, as under the correct model.

The conclusion from this simulation study is that even mild distortions from normality can affect the magnitude of the estimates of the ATE (but not their sign), but these mild distortions can be detected by the goodness of fit test statistics.

Table D1. Average estimates (S.E.) and percent of samples with rejected model.

Distribution A	**Private schools** $\hat{\mu}_{1,S} = 6.01(0.14)$, $\hat{\mu}_{1,C} = 5.99(0.18)$				**Public schools** $\hat{\mu}_{0,S} = 7.34(0.13)$, $\hat{\mu}_{0,C} = 7.09(0.39)$			
ATE	$\hat{\tau}_S = -1.32(0.20)$; $\hat{\tau}_C = -1.10(0.43)$							
Sig. level	0.10	0.05	0.025	0.01	0.10	0.05	0.025	0.01
K-S	100	100	99	96	75	64	55	52
C-M	100	100	100	100	91	87	79	72
A-D	100	100	100	100	94	90	88	83

Distribution B	**Private schools** $\hat{\mu}_{1,S} = 5.96(0.19)$, $\hat{\mu}_{1,C} = 6.01(0.15)$				**Public schools** $\hat{\mu}_{0,S} = 6.84(0.10)$, $\hat{\mu}_{0,C} = 6.89(0.26)$			
ATE	$\hat{\tau}_S = -0.88(0.22)$; $\hat{\tau}_C = -0.88(0.31)$							
Sig. level	0.10	0.05	0.025	0.01	0.10	0.05	0.025	0.01
K-S	100	100	100	99	72	64	51	47
C-M	100	100	100	100	83	77	66	63
A-D	100	100	100	100	92	79	76	70

Distribution C	**Private schools** $\hat{\mu}_{1,S} = 5.42(0.20)$, $\hat{\mu}_{1,C} = 5.49(0.67)$				**Public schools** $\hat{\mu}_{0,S} = 6.76(0.08)$, $\hat{\mu}_{0,C} = 6.90(0.46)$			
ATE	$\hat{\tau}_S = -1.34\ (0.22)$; $\hat{\tau}_C = -1.41\ (0.83)$							
Sig. level	0.10	0.05	0.025	0.01	0.10	0.05	0.025	0.01
K-S	100	100	100	100	67	56	40	39
C-M	100	100	100	100	85	75	71	67
A-D	100	100	100	100	93	88	79	78

APPENDIX E: PROOF OF PART (a) of LEMMA 1

Let $f_p^{(1)}(y;\theta^{(1)})$, $f_p^{(2)}(y;\theta^{(2)})$ be two different *pdf*s with corresponding PARs $p^{(1)}(y;\alpha^{(1)}), p^{(2)}(y;\alpha^{(2)})$ satisfying the requirements of Lemma 1. Suppose to the contrary that the pairs $[f_p^{(j)}, p^{(j)}]$ $(j=1,2)$ yield the same sample model for all $y \in J^*$, or equivalently,

(E.1) $$K\frac{f_p^{(2)}(y;\theta^{(2)})}{f_p^{(1)}(y;\theta^{(1)})} = \frac{p^{(1)}(y;\alpha^{(1)})}{p^{(2)}(y;\alpha^{(2)})}, \quad \forall y \in J^*,$$

where $K = \int f_p^{(1)}(y;\theta^{(1)})p^{(1)}(y;\alpha^{(1)})dy / \int f_p^{(2)}(y;\theta^{(2)})p^{(2)}(y;\alpha^{(2)})dy$ is a positive constant. By taking the limit $y \to \infty$ in both sides of (E.1), we get a contradiction; the right hand side tends to a positive constant whereas the left hand side cannot have a positive limit because $\lim_{y\to\infty} R_p(y;\theta^{(1)},\theta^{(2)}) = \lim_{y\to\infty}\frac{f_p^{(2)}(y;\theta^{(2)})}{f_p^{(1)}(y;\theta^{(1)})} = 0,\infty$ or does not exist.

APPENDIX F: PROOF OF RESULT 1

Let $f_p^{(j)}(y \mid \mathrm{x};\beta_0^{(j)},\beta^{(j)},\sigma^{2(j)}) = N(\beta_0^{(j)} + \mathrm{x}'\beta^{(j)},\sigma^{2(j)})$ be two different normal *pdf*s with logistic *PAR*s $p^{(j)}(y,\mathrm{v};\gamma_0^{(j)},\delta^{(j)},\gamma^{(j)}) = [1+\exp(-\gamma_0^{(j)} - \delta^{(j)}y - \mathrm{v}'\gamma^{(j)})]^{-1}$, $j=1,2$. Suppose to the contrary of the result that the pairs $[f_p^{(j)}, p^{(j)}]$ yield the same sample *pdf* for all $y \in (-\infty,\infty)$ or equivalently, after some algebra, that

(F1) $$K\exp(Q_0 + Q_1 y + Q_2 y^2) = \frac{\exp(\gamma_0^{(1)} + \delta^{(1)}y + \mathrm{v}'\gamma^{(1)})}{1+\exp(\gamma_0^{(1)} + \delta^{(1)}y + \mathrm{v}'\gamma^{(1)})} \times \frac{1+\exp(\gamma_0^{(2)} + \delta^{(2)}y + \mathrm{v}'\gamma^{(2)})}{\exp(\gamma_0^{(2)} + \delta^{(2)}y + \mathrm{v}'\gamma^{(2)})},$$

where

$$Q_0 = \ln(\sigma^{(1)}/\sigma^{(2)}) + \frac{\beta_0^{2(1)}}{2\sigma^{2(1)}} - \frac{\beta_0^{2(2)}}{2\sigma^{2(2)}} + \mathrm{x}'(\frac{\beta_0^{(1)}}{\sigma^{2(1)}}\beta^{(1)} - \frac{\beta_0^{(2)}}{\sigma^{2(2)}}\beta^{(2)}) + \frac{(\mathrm{x}'\beta^{(1)})^2}{2\sigma^{2(1)}} - \frac{(\mathrm{x}'\beta^{(2)})^2}{2\sigma^{2(2)}},$$

$Q_1 = \frac{\beta_0^{(2)}}{\sigma^{2(2)}} - \frac{\beta_0^{(1)}}{\sigma^{2(1)}} + \mathrm{x}'(\frac{\beta^{(2)}}{\sigma^{2(2)}} - \frac{\beta^{(1)}}{\sigma^{2(1)}})$, $Q_2 = \frac{1}{2\sigma^{2(1)}} - \frac{1}{2\sigma^{2(2)}}$ and K is a constant.

Consider the following cases:

Case 1. $\delta^{(1)} \geq 0, \delta^{(2)} > 0$. Then, by Lemma 1, $\beta_0^{(1)} + \mathrm{x}'\beta^{(1)} = \beta_0^{(2)} + \mathrm{x}'\beta^{(2)}$, $\gamma_0^{(1)} + \mathrm{v}'\gamma^{(1)} = \gamma_0^{(2)} + \mathrm{v}'\gamma^{(2)}$ for every x and v, implying $\beta_0^{(1)} = \beta_0^{(2)}, \beta^{(1)} = \beta^{(2)}$, $\gamma_0^{(1)} = \gamma_0^{(2)}, \gamma^{(1)} = \gamma^{(2)}$, such that the sample pdf is identifiable under this condition.

Case 2. $\delta^{(1)} \le 0, \delta^{(2)} < 0$. Same proof as for Case 1.

Case 3. $\delta^{(1)} > 0, \delta^{(2)} < 0$. Re-write (F.1) for this case as,

(F2) $$K\exp(Q_0+Q_1y+Q_2y^2)\exp(\gamma_0^{(2)}+\delta^{(2)}y+\mathrm{v}'\gamma^{(2)})=\frac{\exp(\gamma_0^{(1)}+\delta^{(1)}y+\mathrm{v}'\gamma^{(1)})(1+\exp(\gamma_0^{(2)}+\delta^{(2)}y+\mathrm{v}'\gamma^{(2)}))}{1+\exp(\gamma_0^{(1)}+\delta^{(1)}y+\mathrm{v}'\gamma^{(1)})}$$

By taking the limit $y \to \infty$ on both sides of (F.2), the right hand side tends to 1 for every v and x. The left hand side has the same limit only if the following three equalities hold. (i) $Q_2 = 0$, (ii) $Q_1 + \delta^{(2)} = 0$, (iii) $\ln K + Q_0 + \gamma_0^{(2)} + \mathrm{v}'\gamma^{(2)} = 0$. Using the expressions for Q_1 and Q_2 defined below (F.1) yields the alternative equalities,

(i) $\sigma^{(1)} = \sigma^{(2)} = \sigma$. (ii) $\beta^{(1)} = \beta^{(2)} = \beta$, $\beta_0^{(2)} - \beta_0^{(1)} = -\sigma^2\delta^{(2)}$, and by (i)-(iii),

(F.3) $$\begin{cases} \beta_0^{(2)} - \beta_0^{(1)} = -\sigma^2\delta^{(2)} \\ 2\sigma^2 \ln K + (\beta_0^{2(1)} - \beta_0^{2(2)} + 2\sigma^2\gamma_0^{(2)}) + 2(\beta_0^{(1)} - \beta_0^{(2)})\mathrm{x}'\beta + 2\sigma^2\mathrm{v}'\gamma^{(2)} = 0 \end{cases}.$$

Distinguish between the following two cases:

(a) $\mathrm{x}'=(\tilde{\mathrm{x}}',\mathrm{x}'^*)$, none of the elements of x^* is included in v ($\tilde{\mathrm{x}}$ and v may contain common covariates). Denote by $\beta' = (\tilde{\beta}', \beta'^*)$ the coefficients of $\mathrm{x}=(\tilde{\mathrm{x}}',\mathrm{x}'^*)'$. The second equality in (F.3) is then,

$$2\sigma^2(\ln K + \gamma_0^{(2)}) + \beta_0^{2(1)} - \beta_0^{2(2)} + 2(\beta_0^{(1)} - \beta_0^{(2)})\tilde{\mathrm{x}}'\tilde{\beta} + 2(\beta_0^{(1)} - \beta_0^{(2)})\mathrm{x}'^*\beta^* + 2\sigma^2\mathrm{v}'\gamma^{(2)} = 0.$$

This equality cannot hold for every x and v unless $\beta_0^{(1)} = \beta_0^{(2)}$ (provided $\beta^* \ne 0$ as assumed), in which case we get $f_p^{(1)} = f_p^{(2)}$ in contradiction to the assumption.

(b) $\mathrm{v}'=(\tilde{\mathrm{v}}',\mathrm{v}'^*)$, none of the elements of $\mathrm{v}*$ is included in x ($\tilde{\mathrm{v}}$ and x may contain common covariates). Denote by $\gamma'^{(2)} = (\tilde{\gamma}'^{(2)}, \gamma'^{*(2)})$ the coefficients of $\mathrm{v}=(\tilde{\mathrm{v}}',\mathrm{v}'^*)'$ in the logit PAR. The second equality in (F.3) is in this case,

$$2\sigma^2(\ln K + \gamma_0^{(2)}) + \beta_0^{2(1)} - \beta_0^{2(2)} + 2(\beta_0^{(1)} - \beta_0^{(2)})\mathrm{x}'\beta + 2\sigma^2\tilde{\mathrm{v}}'\tilde{\gamma}^{(2)} + 2\sigma^2\mathrm{v}'^*\gamma^{*(2)} = 0.$$

This equality cannot hold for every x and v unless $\beta_0^{(1)} = \beta_0^{(2)}$ (provided $\gamma^{*(2)} \ne 0$ as assumed), in which case we get again $f_p^{(1)} = f_p^{(2)}$ in contradiction to the assumption.

Remark 12. if $\mathrm{x} = \mathrm{v}$, the second equation in (F.3) implies $2\sigma^2(\ln K + \gamma_0^{(2)}) + \beta_0^{2(1)} - \beta_0^{2(2)} + 2\mathrm{x}'[(\beta_0^{(1)} - \beta_0^{(2)})\beta + \sigma^2\gamma^{(2)}] = 0$, and coupled with the first equation, $\beta_0^{(1)} = -\frac{\gamma_0^{(2)} + \ln K}{\delta^{(2)}} + \sigma^2\delta^{(2)}/2;\ \beta_0^{(2)} = -\frac{\gamma_0^{(2)} + \ln K}{\delta^{(2)}} - \sigma^2\delta^{(2)}/2;\ \gamma^{(2)} = -\delta^{(2)}\beta$.

If these equalities hold, the left hand side in (F.2) equals 1 for all values of y and x. The right hand side equals 1 for all values of y and x if $\gamma_0^{(1)} = -\gamma_0^{(2)}$; $\delta^{(1)} = -\delta^{(2)}$ and $\gamma^{(1)} = -\gamma^{(2)}$. Thus, in this case we have two sets of parameters for which the sample *pdf* is the same, and hence is not identifiable. In practice, the above three parameter relationships are very unrealistic. They imply, in particular, that the parameters of the population model and the parameters of the PAR are not distinct in the sense of Rubin (1976).

REFERENCES

Abadie A, Drukker D., Herr L.G, and Imbens G. (2001). Implementing Matching Estimators for Average Treatment Effects in Stata. *The Stata Journal*, **1**, 1-18.

Abadie, A. and Imbens, G.W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica* **74**, 235-267.

Azzalini A. and Capitanio, A. (2003). Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew *t* Distribution. *Journal of the Royal Statistical Society,* Series B **65**, 367-389.

Babu, G. J. and Feigelson, E.D. (2006). Astrostatistics: Goodness-of-Fit and All That! In: *Astronomical Data Analysis Software and Systems XV*, ASP Conference Series, eds. C. Gabriel, C. Arviset, D. Ponz and E. Solano **35**. pp. 127-136.

Babu, G.J. and Rao, C.R. (2004). Goodness-of-Fit Tests when Parameters are Estimated. *Sankhya, Series A* **66**, 63-74.

Brewer, K.R.W. (1963). Ratio Estimation and Finite Populations: Some Results Deductable from the Assumption of an Underlying Stochastic Process. *Australian Journal of Statistics* **5**, 93-105.

Dronkers, J. and Avram, S. (2010). A Cross-National Analysis of the Relations of School Choice and Effectiveness Differences Between Private-Dependent and Public Schools. *Educational Research and Evaluation* **16**, 151–175.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*. CRC Press.

Greenlees, J.S., Reece W.S. and Zieschang, K.D. (1982). Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed. *Journal of the American Statistical Association* **77**, 251-261.

Hajek, J. (1971). Comment on “An Essay on the Logical Foundations of Survey Sampling, Part One”. *The Foundations of Survey Sampling*. eds. V.P.Godambe and D.A. Sprott, p. 236, Holt, Rinehart and Winston.

Hanushek, E. (2002). Publicly Provided Education. In: *Handbook of Public Economics,* eds. A. J. Auerbach and M. Feldstein. Amsterdam: North Holland.

Heckman, J. and Vytlacil, E. (2006). Econometric Evaluation of Social Programs. In: *Handbook of Econometrics* **6**, eds. J. Heckman and E. Leamer. Amsterdam: North Holland.

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.

Hoxby, C.M. (2000). Does Competition Among Public Schools Benefit Students and Taxpayers? *The American Economic Review* **90**, 1209-1238.

Imbens G. and Angrist J. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrics*, **62**, 467-475.

Landsman, V. (2008), Estimation of Treatment Effects in Observational Studies by Recovering the Assignment Probabilities and the Population Model. PhD Dissertation, Hebrew University of Jerusalem, Israel.

Lee, J. and Berger, J.O. (2001). Semiparametric Bayesian Analysis of Selection Models. *Journal of the American Statistical Association* **96**, 1397-1409.

Little, R. J. (2004). To Model or not to Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association* **99**, 546-556.

Lunceford, J.K. and Davidian, M. (2004). Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study. *Statistics in Medicine* **23**, 2937-2960.

Maddala, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press. Cambridge, UK.

McCaffrey D., Ridgeway, G., and Morral, A. (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, **9**, 403-425.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting Generalized Linear Models Under informative Sampling. In: *Analysis of survey Data*, eds. C. Skinner and R. Chambers, New York: Wiley.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under Informative Sampling. *Handbook of Statistics* 29B; *Sample Surveys: Inference and Analysis*, eds. D. Pfeffermann and C.R. Rao. Amsterdam: North Holland.

PISA 2000 *Technical Report* (2002), Edited by Ray Adams and M.Wu. OECD. Paris.

Qin, J. and Zhang, B. (2007). Empirical-Likelihood-Based Inference in Missing Response Problems and its Application in Observational Studies. *Journal of the Royal Statistical Society*, *Series B* **69**, 101-122.

R Development Core Team (2004). R: *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Rosenbaum P. (2002). *Observational Studies, Second Edition.* Springer-Verlag.

Rosenbaum, P.R. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Treatment Effects. *Biometrika* **70**, 41-55.

Rotnitzky, A. and Robins, J. (1997). Analysis of Semi-Parametric Regression Models With Non-Ignorable Non-Response. *Statistics in Medicine* **16**, 81-102.

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* **66**, 688-701.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika* **63**, 581-592.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag: New York.

Smith, T.M.F. and Sugden, R.A. (1988). Sampling and Assignment Mechanisms in Experiments, Surveys and Observational Studies. *International Statistical Review* **56**, 165-180.

StataCorp. 2004. *Stata Statistical Software: Release 7.* College Station, TX: StataCorp LP.

Stephens, M. A. (1986). Tests Based on EDF Statistics. In: *Goodness- of- fit techniques*, eds. R.B. D'Agostino and M.A. Stephens. Marcel Dekkers: New York. pp. 97-193.

Sverchkov, M. and Pfeffermann, D. (2004). Prediction of Finite Population Totals Based on the Sample Distribution. *Survey Methodology* **30**, 79-92.

Vandenberghe, V. and Robin, S. (2004). Evaluating the Effectiveness of Private Education Across Countries: A Comparison of Methods. *Labour Economics* **11**, 487-506.

Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge: MA.