Empir Econ (2015) 48:1287–1325
DOI 10.1007/s00181-014-0814-x

Latent variables and propensity score matching: a simulation study with application to data from the Programme for International Student Assessment in Poland

Maciej Jakubowski

Received: 10 August 2011 / Accepted: 27 February 2014 / Published online: 6 June 2014

Abstract This paper examines how including latent variables can benefit propensity score matching. Latent variables can be estimated from the observed manifest variables and used in matching. This paper demonstrates the benefits of such an approach by comparing it with a method where the manifest variables are directly used in matching. Estimating the propensity score on the manifest variables introduces a measurement error that can be limited with estimating the propensity score on the estimated latent variable. We use Monte Carlo simulations to test how the proposed approach behaves under distinct circumstances found in practice, and then apply it to real data. Using the estimated latent variable in the propensity score matching limits the measurement error bias of the treatment effects' estimates and increases their precision. The benefits are larger for small samples and with better information about the latent variable available.

Keywords Program evaluation · Matching · Treatment effects · Measurement error · Human capital · PISA · Tracking

JEL Classification C14 · C15 · C21

1 Introduction

This paper demonstrates how incorporating latent variable modeling into propensity score matching can limit measurement error in the propensity score and, in effect, increase precision of the estimates of treatment effects. This is not the first paper to introduce latent variable modeling into propensity score matching. The seminal work of James Heckman and others was helpful when developing the framework

M. Jakubowski (✉)
Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland
e-mail: mjakubowski@uw.edu.pl

we present (see Abbring and Heckman 2007, Sect. 2.7 for discussion and further references). However, our approach differs in one important aspect. Heckman and others modeled residuals from outcome equations across quasi-experimental groups assuming that there are latent traits behind them. We assume, instead, that values of the latent variable are associated with values of the manifest variables that are observable and can be used to estimate the latent variable. We present the benefits of using the estimated latent variable in the propensity score matching, showing that it can notably limit the measurement error bias and lower the variance of treatment estimators. We demonstrate this with Monte Carlo simulations and real data examples.

The paper is organized as follows. Section 2 shows how latent variables modeling can be introduced into propensity score matching using the measurement error model of the relation between the latent trait and the manifest variables. Section 3 provides evidence from a Monte Carlo study on how the proposed approach increases efficiency of the matching estimators of the treatment effects. Section 4 empirically applies this approach to data. Section 5 concludes.

2 Modeling latent variables in propensity score matching

2.1 Modeling latent variables

Latent variables can reflect either hypothetical constructs or existing phenomena which cannot be directly measured but are often reflected in observed variables that are proxies of measured phenomena. These observable manifestations are correlated with latent variables but also contain an independent component, or, in other words, manifest variables contain a signal about the latent variable and the random component (often called the "measurement error") that is uncorrelated with this variable.

A relation between the latent variable and the manifest variables can be presented using the one-factor model or the congeneric measurement error model (Joreskog 1971; Skrondal and Rabe-Hesketh 2004). We assume that the observed j-th variable is measured with error and on a scale specific to that variable. Values for a set of such manifest variables are observed for each i-th individual. This is modeled through the following equation:

$$M_{ij} = \delta_j + \lambda_j \eta_i + E_{ij}, \tag{1}$$

where η is the latent variable or common factor and M_{ij} are observed realizations of manifest variables. We assume independent error terms E_{ij} and $N(0, \sigma_j)$. This model can also be interpreted as a measurement error model where true scores η are reflected in each j-th variable with random error on the scale defined by δ_j and λ_j. In factor analysis λ_j are called factor loadings and δ_j are called intercepts. To identify this model some restrictions are needed, for example, that $\lambda_1 = 1$ or that the $\mathrm{Var}(\eta) = 1$. This model assumes that there is only one latent variable behind the manifest variables; however, dealing with several latent variables is quite straightforward within this framework.

If the model we present above is true, the latent variable can be estimated from the manifest variables using the factor analysis approach. Specification of a latent variable factor model has to be driven by theoretical considerations and carefully

tested empirically. Usually, models assuming different numbers of common factors are estimated and compared on how they fit the data. A problem arises when such models all seem to be plausible, in which case theoretical considerations can play an important role. In this paper, we abstract from these issues, assuming that a latent variable model properly reflects the latent structure behind the data. Moreover, we assume there is only one latent variable behind each set of manifest variables. Our approach can be easily extended to more complex situations, for example, involving more latent variables and allowing for correlations with other variables in a model. However, simple models considered in this paper illustrate the main benefits of incorporating latent variables in matching. Our general findings should also hold under more complex circumstances.

We are not aware of any other study, other than the efforts of James Heckman and his colleagues described in the introduction, which attempts to model latent traits in the propensity score matching. The usefulness of latent variable modeling in economic research can be reconsidered when taking into account modern developments in statistics. Recent work demonstrates that current approaches are much more reliable, theory driven and more adverse to ad hoc interpretations (see Skrondal and Rabe-Hesketh 2004, for an extensive discussion and a unifying framework; for discussion and examples of latent variable modeling in econometrics see Kmenta 1991; Wansbeek and Meijer 2000; Aigner et al. 1984). In addition to this, there are circumstances, such as evaluating labor market training or school programs, where latent traits like attitudes play an important role in the choices of participants and non-participants. In labor market studies, it was shown that job satisfaction, even if measured through a single simple question, significantly changes quasi-experimental estimates (White and Killeen 2002). Large economic literature discusses how personality traits affect labor market outcomes. For example, Girtz (2012) shows how self-esteem and locus of control, two personality traits modeled as latent variables, affect wages of adults.

In evaluation studies, attitudes toward work or other personal traits are used to control for selection and factors driving the outcomes. For example, in evaluation studies in Germany researchers very often use data from the German Socio-Economic Panel (GSOEP) which contain information on attitudes or personality traits which can be used by researchers in propensity score matching (for examples see Lechner 2000; Barg and Beblo 2009; Heineck and Anger 2010). In educational studies, a number of works demonstrate the importance of student attitudes or latent family characteristics (OECD 2009; Jakubowski and Pokropek 2009). Behavioral economics often looks into the ways personality traits or attitudes affect people's behavior, for example, in experiments analyzing risk aversion or economic preferences (see Fairlie and Holleran 2012; Grabner et al. 2009; Ovchinnikova et al. 2009; Ben-Ner et al. 2008).

Survey responses are often used to reflect latent traits instead of modeling these traits directly. In studies of anti-poverty programs direct responses about household possessions are typically used (see Jalan and Ravallion 2003), although they could be modeled as latent traits reflecting household wealth and socio-economic position (we use a similar example in this paper).[1] In a well-known paper by Agodini and

[1] It is sometimes argued against defining household's wealth as a latent trait. It should be noted, however, that in many surveys income questions are not asked due to confidentiality reasons or fear that because of these questions respondents will refuse to answer the survey. Instead of direct measures of income, some

Dynarski (2004) on propensity score matching, student responses to questions about time use and attitudes toward learning were added to the list of matching covariates instead of being used to model the latent characteristics behind them. In evaluation studies mentioned above, researchers prefer to use single questions rather than scales summarizing responses to questionnaire items which are reflecting some latent traits. For example, Lechner (2000) uses single responses to chosen questions about people's optimism, although they are part of a set of questions which could be used to build a scale reflecting overall optimism of a person.

We propose an approach where the latent variable is estimated from the manifest variables and directly used in the propensity score matching. Our simulation results and empirical examples demonstrate that modeling the latent variable might decrease bias and increase precision of propensity score matching estimates of treatment effects.

2.2 Propensity score matching with latent variables

Consider a situation where we want to compare outcomes between two groups in which the latent variable is unbalanced. One of these quasi-experimental groups is affected by a treatment, while the other remains unaffected and serves as a baseline reference group. We call subjects in the first group the "treated" and subjects in the latter group the "controls." We assume that the latent variable affects outcomes in both groups, and that the imbalance in the latent variable creates bias when comparing group outcomes.

For observational studies, a matching approach was proposed to balance covariates among groups of treated and controls (Rubin 1973). Propensity score matching is currently the most popular version of this approach and is based on balancing covariates through matching conducted on a propensity score (Rosenbaum and Rubin 1983). The propensity score is usually estimated by logit or probit and reflects the probability of being selected to the group of treated. Matching based on the propensity score instead of matching on all covariates solves the so-called curse of dimensionality that makes normal matching inadvisable or even impossible in smaller samples. After balancing covariates by using matching, simple outcome comparisons provide unbiased estimates of treatment effects, assuming that all differences between the two groups are observed and taken into account when estimating the propensity score (see Heckman et al. 1998, for detailed assumptions).

Consider that not only the imbalance of the observed covariates, but also the imbalance of the latent variable, poses a potential barrier to estimating the treatment effects. In this case, a researcher would like to include the latent variable in matching; however, it is not observed. Instead, matching has to be conducted on the observed variables, including the manifest variables that are only proxies of the latent trait and, by assump-

Footnote 1 continued
authors recommend using household possessions as a basis for wealth comparisons (McKenzie 2005). Often survey questionnaires thus contain a list of household possessions that are later used to estimate household wealth. Similarly, a list of questions can be used to assess socio-economic status, although SES is usually defined as a latent trait reflecting a theoretical concept rather than a real phenomenon, similarly to personal traits in psychology or behavioral economics.

tion, reflect it with a random error. Estimating the propensity score on the manifest variables thus introduces additional noise into matching. Intuitively, the greater the error, more often are subjects mismatched, which affects the quality of matching estimators. The smaller the error is or the stronger a signal from the latent variable reflected in the manifest variables is, more negligible is the fact that matching is not conducted directly on the latent variable.

This paper discusses how estimating the latent variable, and conducting matching on this estimate rather than on a set of manifest variables, can increase the quality of matching in some situations, particularly in smaller samples or when a relatively weak signal about the latent variable is available in the manifest variables. If the latent variable model is correct, the estimated latent variable should reflect the latent variable with more precision than observable proxies. This will benefit matching, as less error is introduced.

More formally, consider first a hypothetical situation where the latent variable is directly observed and can be used for matching. In this case, a propensity score is given by:

$$p(\mathbf{X}, \eta) \equiv \Pr\{D = 1|\mathbf{X}, \eta\} = E\{D|\mathbf{X}, \eta\}, \tag{2}$$

where $D = \{0, 1\}$ is the indicator of treatment exposure, η is the latent variable that has to be balanced together with other covariates contained in the vector $\mathbf{X}$. Rosenbaum and Rubin (1983) show that if the treatment assignment is random conditionally on multidimensional vector $(\mathbf{X}, \eta)$, it is also random when conditioning on the $p(\mathbf{X}, \eta)$. In this case, after successful matching which balances the latent variable and other covariates among quasi-experimental groups, the average treatment effect on the treated (ATT) can be estimated through comparisons of expected outcomes in a group of treated and matched controls (Becker and Ichino 2002):

$$\text{ATT} = E\left\{E\left\{Y_1|D = 1, p(\mathbf{X}, \eta)\right\} - E\left\{Y_0|D = 0, p(\mathbf{X}, \eta)\right\}|D = 1\right\}, \tag{3}$$

where the outer expectation is over the distribution of $p(\mathbf{X}, \eta)|D = 1$, while Y_1 and Y_0 are potential outcomes in case of treatment and no treatment, respectively.

Assumptions needed to identify treatment effects using propensity score matching require that the propensity score is defined for all treated (common support) and that a researcher observes a set of covariates such that after controlling for these covariates the potential outcomes are independent of the treatment. Formally, for the average treatment effect on the treated, these two assumptions can be written as:

$$Y_0 \perp D|\,(\mathbf{X}, \eta) \tag{4}$$

$$p(\mathbf{X}, \eta) < 1. \tag{5}$$

If these two assumptions are satisfied, we call the treatment assignment strongly ignorable and the formula (3) provides an unbiased ATT estimate.

Battistin and Chesher (2009), extending the seminal work by Cochran and Rubin (1973), show that even if strong ignorability holds when covariates are measured without error, it does not mean that this assumption will hold if some covariates are measured with error. Cochran and Rubin analyze a relatively simple situation where

outcome equations are linear in $\mathbf{X}$ and matching is conducted on only one covariate. Battistin and Chesher analyze a more general setting, but in both cases it is shown that measurement error in covariates can have substantial bias on the ATT estimate, with the sign of bias hard to predict under more complex circumstances. In general, both papers show that the bias increases with larger measurement error variance and with stronger effect of the error-prone covariate on the outcome. Battistin and Chesher propose a method to approximate this bias, which can be applied for cases with relatively small measurement error variance and typical distribution of $\mathbf{X}$.

According to both papers, the measurement error bias in matching is difficult to assess analytically. Even the sign of the bias might be impossible to evaluate under more complex circumstances or with large measurement error variance. In our application, we focus on a specific case where observed manifest variables are error-prone reflections of the latent variable. In this case, measurement error can be limited by attempts to estimate the latent trait and use it directly in matching. We provide insights based on Monte Carlo simulations and present empirical application, as general analytical solutions are too difficult to derive in this case. If we estimate the propensity score using information on the latent variable reflected in a set of manifest variables, the propensity score is given by:

$$p(\mathbf{X}, \mathbf{M}) \equiv \Pr\{D = 1 \mid \mathbf{X}, \mathbf{M}\} = E\{D|\mathbf{X}, \mathbf{M}\}. \tag{6}$$

As $p(\mathbf{X},\mathbf{M})$ is an error-prone reflection of $p(\mathbf{X}, \eta)$, matching does not necessarily have the balancing property that is needed to obtain unbiased ATT. In other words, matching on $p(\mathbf{X}, \mathbf{M})$ will not necessarily eliminate all the differences between the treated and controls in terms of the latent variable η.

Instead of matching on the manifest variables, we propose estimating the latent variable from the manifest variables in the first step. We assume that the latent structure and the model to estimate it follow the one described by the set of equations (1), in which one latent factor is reflected in the observed manifest variables. Here, the latent variable can be estimated by the factor analysis model, and the latent variable estimate $\hat{\eta}$ can be used to obtain the propensity score:

$$p(\mathbf{X}, \hat{\eta}) \equiv \Pr\{D = 1 \mid \mathbf{X}, \hat{\eta}\} = E\{D|\mathbf{X}, \hat{\eta}\}. \tag{7}$$

As $\hat{\eta}$ is a less noisy measure of the latent variable η than the set of manifest variables $\mathbf{M}$, it follows that matching on $p(\mathbf{X}, \hat{\eta})$ should give results closer to matching on $P(\mathbf{X}, \eta)$ than the matching on $p(\mathbf{X}, \mathbf{M})$.

Obtaining exact formulas for the bias in the ATT or for the variance of the matching estimators is difficult or even impossible for matching methods (see Abadie and Imbens 2009, for attempts to derive analytical formulas). As noted, Battistin and Chesher (2009) and Cochran and Rubin (1973) show that measurement error in matching covariates will cause bias in the ATT matching estimate. We are not able to provide analytical formulas for the biased matching estimator variance. Instead, we use simulation to provide evidence that the measurement error increases it. We then propose a method which might decrease matching estimator variance by limiting measurement error in matching covariates.

We described above three propensity scores that can be used for balancing the latent variable and covariates through propensity score matching: a hypothetical propensity score $P(\mathbf{X}, \eta)$ estimated on an unobservable latent variable, a propensity score $p(\mathbf{X}, \mathbf{M})$ estimated from the observed manifest variables, and finally a propensity score $p(\mathbf{X}, \hat{\eta})$ obtained through an estimate of the latent variable from the manifest variables. Through a simulation study in which matching is conducted under various circumstances commonly found in the practice of empirical research, we try to establish how bias in the matching estimate of treatment effects differs when using these three different propensity scores. In the simulation, we compare the results obtained with the hypothetical propensity score estimated using the unobserved latent variable with results obtained with error-prone propensity scores used in practice. In Sect. 3, we also apply the strategy suggested by the simulation results to data from an educational study.

3 Simulation study

In the simulation study, we analyze two outcome equations and two different selection to treatment mechanisms. The data were simulated for four variables with X_1, η and X_2 having a normal distribution with mean zero and covariance matrix given below by vector $\mathbf{M}$ and matrix $\mathbf{V}$:

$$\begin{pmatrix} X_1 \\ \eta \\ X_2 \end{pmatrix} = N(\mathbf{M}, \mathbf{V}), \quad \text{where } \mathbf{M} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & 0.5 \\ -1 & -0.5 & 1 \end{pmatrix} \tag{8}$$

and X_3 distributed $N(0,1)$ and uncorrelated with other variables.

Four selection equations were studied. In each case, different error term was simulated. In the first model selection to treatment is random:

$$\text{Model A:} \quad D_{\mathrm{A}} = 1\,(U > 0.5) \quad U \sim N\,(0, 1)\,.$$

In the following three models treatment selection depends on a set of covariates which include the latent variable. The distribution of covariates among the treated and control groups depends on the distribution of the error term. Three different models specified below assume different distributions of the error term U providing different matching scenarios:

$$\begin{aligned}
&\text{Model B:} \quad D_{\mathrm{B}} = 1(\eta + X_1 - X_2 - X_3 + U > 1), \quad U = \sqrt{30}Z, \quad Z \sim N\,(0, 1) \\
&\text{Model C:} \quad D_{\mathrm{C}} = 1(\eta + X_1 - X_2 - X_3 + U > 1), \quad U = \sqrt{5}Z, \quad Z \sim N\,(0, 1) \\
&\text{Model D:} \quad D_{\mathrm{D}} = 1(\eta + X_1 - X_2 - X_3 + U > 3), \quad U = 3Z, \quad Z \sim \chi^2\,(1)\,.
\end{aligned}$$

These selection equations give different distributions of covariates among the treated and controls and different proportion of the treated in a sample. This is reflected in the varying distribution of the propensity score among models depicted on Fig. 1 for an exemplary sample.

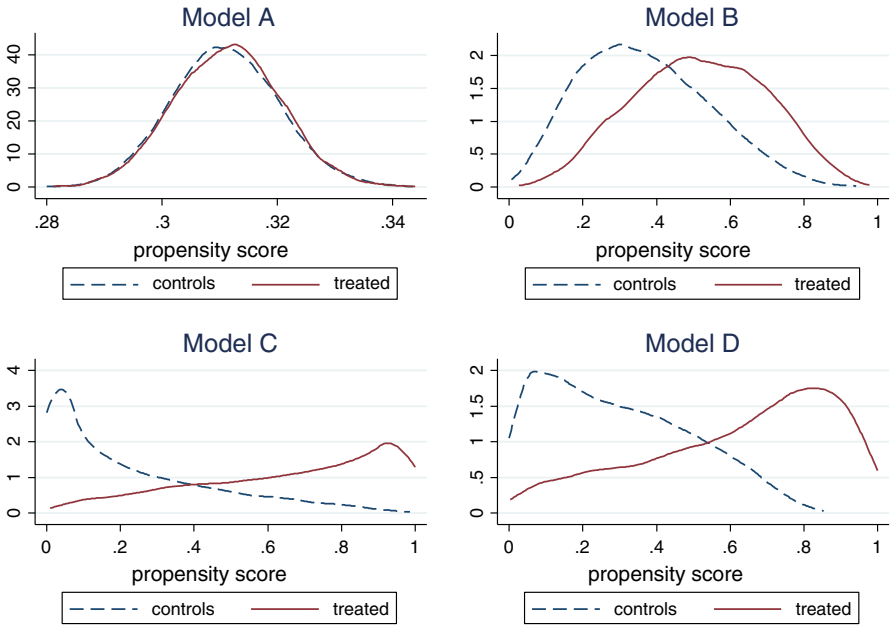

Fig. 1 Propensity score distributions for different models

In Model A, all the treated and controls are on the same support. In Model B around 1 % of the treated might not be on the same support, as well as around 5 % of the treated in Model C and around 1/3 in Model D. Thus, Models C and D are much more demanding matching scenarios with many treated without good matches. The proportion of treated also differs between models, with around 30 % assigned to treatment in Model A and around 40 % assigned to treatment in Models B, C and D.

We study three outcome equations. In all equations the outcome differs among the treated ($D = 1$) and controls ($D = 0$) by a known constant equal 3. Equation 1 represents a simple situation where all covariates are linearly related to the outcome, while in Eq. 2 the relationship between the outcome and covariates, including the latent variable, is more complex and non-linear.

$$\text{Outcome 1:} \quad y = -10 + \eta + 3D + X_1 - X_2 - X_3 + \varepsilon$$
$$\text{Outcome 2:} \quad y = -10 + \eta^3 + 3D + X_1^2 - 2X_2 - X_3^3 + \varepsilon.$$

The third equation adds an interaction between the latent variable and treatment to the Eq. 1, so the effect of the latent variable on the outcome is doubled for the treated:

$$\text{Outcome 3:} \quad y = -10 + \eta + 3D + \eta D + X_1 - X_2 - X_3 + \varepsilon.$$

In all cases ε is independently and normally distributed N(0,1). In all simulations, 10,000 random draws were studied.

Although values of the latent variable are observed in our simulation, we assume that a researcher observes only manifest variables that are generated by a set of equations:

$$M_j = \eta + kZ_j, \tag{9}$$

where M_j denotes the j-th manifest variable constructed from the latent variable η by adding a random noise Z_j specific to each manifest variable. Correlation between the latent variable and the manifest variables depends on a signal-to-noise ratio captured in the parameter k that is studied in the simulation. For example, with $k = 2$ the signal-to-noise ratio equals 1:2, which means that correlation between the latent variable and a manifest variable is close to 0.45. We studied also the results for value of k equal to 1 where correlation between the latent variable and a manifest variable is close to 0.7. This gives a typical range found in empirical research. In practice, when correlation of manifest variables (commonly called "items") is weaker than 0.4, that is usually taken as a sign that this variable has no relation to the latent construct. In such a case, the variable is usually dropped and other manifest variables are used.

We also varied the number of manifest variables from which a researcher can estimate the latent variable. Usually, the higher the number of manifest variables is, the better an estimate of the latent variable. We simulated data with 5 and 10 manifest variables, a range that covers typical situations. The quality of the estimated latent variable depends also on the sample size. We studied sample sizes with 500 and 5,000 observations, which provides evidence on the impact of measurement error in small samples and samples typically found in surveys.

For each simulated sample propensity score matching was conducted three times. First, matching was conducted on the latent variable that is normally unobserved, which provides a proper baseline for further comparisons. Second, matching was conducted on a set of manifest variables. Finally, matching was conducted on the estimated latent variable using information reflected in the manifest variables. We estimated the latent variable through a basic one-dimensional factor model using a standard procedure in Stata software (see Stata documentation on the *-factor-* command).[2] This model reflects the process in which manifest variables were generated, assuming that the model used for the latent variable estimation was correctly specified. Obviously, that is not always the case in empirical research. However, we do not study how mistakes in estimation of the latent variable can affect the quality of matching, but simply assume this step was conducted properly.

Main results are presented in Tables 1 and 2. They demonstrate how using the manifest variables instead of the latent one lowers the quality of matching estimates and how matching based on the estimated latent variable can help. The first part of the table presents results from 1-to-1 nearest neighbor propensity score matching without any additional restrictions, while the second part present results with the common

[2] Statistical codes in Stata used to estimate latent variables and conduct matching and simulations are available upon request from the authors. Simulations include all steps needed to derive final estimates. Thus, for the matching on the estimated latent variable it includes estimates the latent variable, estimating the propensity score and final matching on the propensity score.

Table 1 Part 1—Estimation without common support restriction

	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
						Outcome Eq. 1; 500 obs.						
Mean	3.000	2.996	3.004	3.042	3.084	3.083	3.245	3.349	3.347	3.809	3.938	3.905
SD	0.279	0.299	0.284	0.161	0.184	0.169	0.321	0.344	0.330	0.447	0.417	0.422
RMSE	0.279	0.299	0.284	0.166	0.202	0.188	0.404	0.490	0.479	0.924	1.026	0.999
	100 %	107.2 %	101.8 %	100 %	121.7 %	113.3 %	100 %	121.3 %	118.6 %	100 %	111.0 %	108.1 %
						Outcome Eq. 1; 5,000 obs.						
Mean	3.000	2.999	3.000	3.007	3.050	3.049	3.087	3.199	3.198	3.714	3.839	3.830
SD	0.087	0.094	0.089	0.046	0.049	0.049	0.130	0.130	0.129	0.397	0.316	0.316
RMSE	0.087	0.094	0.089	0.046	0.070	0.069	0.156	0.238	0.237	0.817	0.897	0.888
	100 %	108.0 %	102.3 %	100 %	152.2 %	150.0 %	100 %	152.6 %	151.9 %	100 %	109.8 %	108.7 %
						Outcome Eq. 2; 500 obs.						
Mean	3.002	2.990	2.993	3.256	3.370	3.368	4.135	4.372	4.364	5.554	5.917	5.799
SD	0.708	0.759	0.718	0.644	0.713	0.663	1.199	1.240	1.197	1.668	1.372	1.461
RMSE	0.708	0.759	0.718	0.693	0.803	0.758	1.651	1.849	1.814	3.051	3.223	3.157
	100 %	107.2 %	101.4 %	100 %	115.9 %	109.4 %	100 %	112.0 %	109.9 %	100 %	105.6 %	103.5 %
						Outcome Eq. 2; 5,000 obs.						
Mean	3.003	2.999	3.000	3.051	3.183	3.180	3.519	3.844	3.847	5.100	5.588	5.556
SD	0.224	0.236	0.227	0.205	0.219	0.215	0.619	0.587	0.582	1.783	1.033	1.061
RMSE	0.224	0.236	0.227	0.211	0.285	0.281	0.807	1.028	1.027	2.755	2.786	2.767
	100 %	105.4 %	101.3 %	100 %	135.1 %	133.2 %	100 %	127.4 %	127.3 %	100 %	101.1 %	100.4 %

Table 1 continued

	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
						Outcome Eq. 3; 500 obs.						
Mean	3.000	2.996	3.004	3.398	3.440	3.438	3.871	3.975	3.973	4.328	4.457	4.424
SD	0.290	0.312	0.299	0.176	0.200	0.186	0.331	0.355	0.341	0.459	0.431	0.436
RMSE	0.290	0.312	0.299	0.435	0.483	0.476	0.932	1.038	1.031	1.405	1.519	1.490
	100 %	107.6 %	103.1 %	100 %	111.0 %	109.4 %	100 %	111.4 %	110.6 %	100 %	108.1 %	106.0 %
						Outcome Eq. 3; 5,000 obs.						
Mean	3.000	2.999	3.000	3.363	3.406	3.406	3.713	3.825	3.824	4.233	4.359	4.350
SD	0.091	0.098	0.093	0.050	0.055	0.055	0.131	0.132	0.132	0.398	0.317	0.317
RMSE	0.091	0.098	0.093	0.367	0.410	0.409	0.725	0.836	0.834	1.296	1.395	1.386
	100 %	107.7 %	102.2 %	100 %	111.7 %	111.4 %	100 %	115.3 %	115.0 %	100 %	107.6 %	106.9 %

Mean estimate, standard deviation, root mean squared error (also as the percentage comparing to the model with latent variable) of the average treatment effect. Results for noise-to-signal ratio 1:1 and five manifest variables

Simulation results for the propensity score 1-to-1 nearest neighbor matching without replacement, propensity score estimated by probit, noise-to-signal ratio 1:1 and five manifest variables

support restriction imposed and caliper matching (with caliper 0.05).[3] The latter are more suited for comparisons in models with strong selection and unequal distribution of covariates among treated and controls. In fact, for models C and D only results with these additional restrictions should be considered. For these models, due to strong selection to treatment, distribution of covariates do not fully overlap and unrestricted matching estimates can be seriously biased.

Generally, results presented in Tables 1 and 2 suggest that measurement error can introduce serious bias in the estimate of treatment effect. Mean estimates of the simulated ATT effects are close to the true value of 3.0 mainly for matching on the latent variable and for large samples. Mean ATT estimates are seriously biased when matching is not conducted on the latent variable, especially in smaller samples and for models with strong selection. Under more demanding circumstances even the estimates from matching on the latent variable are biased, although this bias is always smaller than the bias for matching on the manifest variables or matching on the estimated latent variable. Thus, measurement error bias is present under almost all circumstances and adds to other sources of bias in matching.

Importantly, in smaller samples estimates obtained from matching on the estimated latent variable are marginally less biased, have slightly smaller variance and have smaller root mean squared error. Matching on the estimated latent variable outperforms matching on the set of manifest variables in smaller samples, while in large samples these differences partly disappear. It is worth noting that even unbiased estimates of the ATT, which can be obtained for simpler models from both matching on the manifest variables and on the estimated latent variable, have slightly larger variance for matching on the manifest variables. The benefits are thus twofold: first, matching on the estimated latent variables increases precision of the estimate. Second, it gives marginally less biased estimate of the ATT comparing to matching on observed manifest variables. Therefore, it is possible to say that matching on the estimated latent variable decreases bias and increases precision of the ATT estimates under all models. On the other hand, however, these gains are modest in the smaller sample and very small in the large sample. Both methods still introduce bias in the estimation, while in larger samples the difference in performance of both methods is rather small.

Tables 1 and 2 results confirm findings by Battistin and Chesher (2009) and Cochran and Rubin (1973) that ATT estimates from matching on error-prone covariates can be seriously biased. Additional evidence here is that measurement error in observed manifestations of the latent variable not only introduces bias but also decreases precision of the matching estimator. These results suggest that estimating latent variable and using it directly in matching helps to mitigate the impact of measurement error to some extent, providing less biased and more precise results, especially in smaller samples. Under the most demanding circumstances in the model with strong selection and no overlapping support, matching on the manifest variables without common support restriction and caliper might even result in smaller standard errors although

[3] In all cases, we use the propensity score matching with replacement. Propensity score was estimated using probit regression. Common support is imposed by dropping treatment observations whose propensity score is higher than the maximum or less than the minimum propensity score of the controls. The Stata procedure –psmatch2- was used to conduct propensity score matching (see Leuven and Sianesi 2003).

Table 2 Part 2—Estimation with common support restriction and imposed caliper (0.05 of the propensity score)

	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
						Outcome Eq. 1; 500 obs.						
Mean	3.000	2.996	3.004	3.000	3.042	3.041	2.985	3.094	3.095	3.041	3.211	3.168
SD	0.280	0.299	0.285	0.150	0.170	0.158	0.203	0.221	0.212	0.179	0.207	0.194
RMSE	0.280	0.299	0.285	0.150	0.175	0.164	0.203	0.240	0.232	0.183	0.296	0.257
	100 %	106.8 %	101.8 %	100 %	116.7 %	109.3 %	100 %	118.2%	114. %3 %	100 %	161.7 %	140.4 %
						Outcome Eq. 1; 5,000 obs.						
Mean	3.000	2.999	3.000	3.000	3.043	3.043	2.995	3.109	3.109	3.008	3.191	3.182
SD	0.087	0.094	0.089	0.045	0.048	0.048	0.080	0.083	0.083	0.058	0.082	0.080
RMSE	0.087	0.094	0.089	0.045	0.065	0.064	0.080	0.137	0.137	0.058	0.208	0.199
	100 %	108.0 %	102.3 %	100 %	144.4 %	142.2 %	100 %	171.3 %	171.3 %	100 %	358.6 %	343.1 %
						Outcome Eq. 2; 500 obs.						
Mean	2.995	2.987	2.987	2.996	3.124	3.116	2.945	3.203	3.211	3.109	3.520	3.384
SD	0.700	0.756	0.714	0.535	0.593	0.557	0.681	0.721	0.697	0.557	0.609	0.582
RMSE	0.700	0.756	0.714	0.535	0.606	0.569	0.683	0.749	0.728	0.567	0.801	0.697
	100 %	108.0 %	102.0 %	100 %	113.3 %	106.4 %	100 %	109.7 %	106.6 %	100 %	141.3 %	122.9 %

Table 2 continued

	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
						Outcome Eq. 2; 5,000 obs.						
Mean	3.002	2.999	2.999	2.999	3.130	3.127	2.981	3.308	3.307	3.004	3.427	3.399
SD	0.223	0.236	0.227	0.184	0.196	0.195	0.329	0.325	0.325	0.200	0.248	0.237
RMSE	0.223	0.236	0.227	0.184	0.236	0.233	0.329	0.448	0.447	0.200	0.494	0.464
	100 %	105.8 %	101.8 %	100 %	128.3 %	126.6 %	100 %	136.2 %	135.9 %	100 %	247.0 %	232.0 %
						Outcome Eq. 3; 500 obs.						
Mean	3.000	3.002	3.006	3.302	3.343	3.348	3.422	3.551	3.554	3.144	3.409	3.342
SD	0.292	0.311	0.301	0.163	0.188	0.174	0.216	0.237	0.231	0.204	0.243	0.226
RMSE	0.292	0.311	0.301	0.343	0.391	0.389	0.474	0.600	0.600	0.250	0.476	0.410
	100 %	106.5 %	103.1 %	100 %	114.0 %	113.4 %	100 %	126.6 %	126.6 %	100 %	190.4 %	164.0 %
						Outcome Eq. 3; 5,000 obs.						
Mean	2.999	3.000	3.001	3.344	3.388	3.388	3.529	3.654	3.655	3.107	3.416	3.401
SD	0.091	0.099	0.094	0.050	0.054	0.054	0.086	0.090	0.089	0.064	0.098	0.096
RMSE	0.091	0.099	0.094	0.347	0.392	0.392	0.536	0.660	0.661	0.125	0.427	0.412
	100 %	108.8 %	103.3 %	100 %	113.0 %	113.0 %	100 %	123.1 %	123.3 %	100 %	341.6 %	329.6 %

Mean estimate, standard deviation, root mean squared error (also as the percentage comparing to the model with latent variable) of the average treatment effect. Results for noise-to-signal ratio 1:1 and five manifest variables

Simulation results for the propensity score 1-to-1 nearest neighbor matching without replacement, propensity score estimated by probit, noise-to-signal ratio 1:1 and five manifest variables

with larger bias and overall larger RMSE. In fact, in both methods the measurement error bias remains large, particularly for models with strong selection and non-linear relationship between outcome and latent variable.

These results are due to smaller measurement error in the propensity score obtained using the estimated latent variable. Measurement errors in binary choice models like logit or probit, which are usually applied to estimate propensity score, can bias estimates in a complex way (see Carroll et al. 1995). The exact bias is hard to assess without additional information on measurement errors. The propensity score obtained from the probit or logit model with error-prone covariates will therefore usually be biased. As already mentioned, bias in the propensity score affects its balancing property which in turn biases the final treatment effect estimates from matching.

Table 3 below demonstrates results from the simulation study in which the bias in the propensity score and balancing tests were analyzed for Models B, C and D. In these models, selection to treatment is not random and results in imbalance in the distribution of covariates among treated and controls. Our balancing test is based on comparisons of the standardized percentage bias, that is the percentage difference of the sample means in the treated and controls group before and after matching as a percentage of the square root of the average of the sample variances in the treated and control groups (see Rosenbaum and Rubin 1985). Table 3 reports the absolute standardized bias for the unobserved latent variable before and after matching, but also the mean absolute standardized bias for all matching covariates.[4]

Average correlation across 10,000 random draws between the propensity score obtained using the latent variable and the propensity score obtained using the estimated latent variable was around 0.99, while correlation between the propensity score obtained using the latent variable and the propensity score obtained using the manifest variables was slightly lower around 0.97–0.98. As Table 3 shows, this results in slightly better balancing property of matching on the estimated latent variable rather than on the manifest variables (after imposing common support for the most demanding model D). It is also clear from the results in Table 3 that while matching on the estimated latent variable is superior to matching on the manifest variables, it also introduces bias due to measurement error, although somewhat smaller in the magnitude.

Below we provide additional simulation results for different numbers of manifest variables and for different signal-to-noise ratios. For brevity, below we provide results for Model C and outcome Eq. 1 only. Results for other models and outcome equations provide similar conclusions and are presented in the Appendix (Tables 9, 10, 11). Table 4 provides results for different signal-to-noise ratios defined by the value of parameter k in Eq. (9), and compares results obtained with 5 manifest variables to those obtained with 10 manifest variables. As previously, estimates are given for sample size of 500 and 5,000.[5]

Results in Table 4 show that precision always increases with the number of manifest variables available, but the benefits from having more manifest variables are

[4] Results for other models as well as for other balancing tests gave similar conclusions and are available upon request from the authors.

[5] Results for larger number of manifest variables, different sample sizes, and other models are available from the authors.

Table 3 Standardized percentage bias before and after matching

	Percentage bias for the latent variable			Mean percentage bias for all covariates		
	Latent	Manifest	Estimated latent	Latent	Manifest	Estimated latent
Model B						
Before matching		66.90			59.08	
After matching (no common support)	5.58	8.38	8.08	5.80	6.66	6.37
After matching (with common support)	5.00	7.52	7.43	5.32	6.07	5.87
Model C						
Before matching		121.32			107.46	
After matching (no common support)	15.12	22.85	22.75	15.55	17.36	16.96
After matching (with common support)	8.50	14.90	14.77	8.82	10.40	10.26
Model D						
Before matching		101.77			89.98	
After matching (no common support)	30.90	42.90	44.39	31.80	31.69	32.07
After matching (with common support)	7.94	16.98	15.99	8.22	10.82	10.35

Simulation results for the propensity score 1-to-1 nearest neighbor matching without replacement, propensity score estimated by probit, noise-to-signal ratio 1:1, five manifest variables, sample size 500

augmented by using the estimated latent variable instead of matching directly on the manifest variables. For example, with $k = 1$ and sample size of 500, root mean squared error is 0.490 for matching on five manifest variables and 0.479 when matching on the latent variable estimated from those five variables. With 10 manifest variables available, root mean squared error goes down to 0.475 for matching on the manifest variables and to 0.445 for matching on the estimated latent variables. Thus, at least for smaller samples, the benefits of matching on the estimated latent variable increase with the number of manifest variables available. For larger samples these differences are relatively small.

Modest benefits of observing less noisy manifestations for all the models, but here again matching on the estimated latent variable proves superior to matching on the manifest variables mainly for smaller samples. For large samples, the benefits of matching on the estimated latent variable are again marginal for both more and less noisy manifestations.

Results of propensity score matching differ according to the specific matching method applied. The results presented above are based on the 1-to-1 nearest neighbor method, while results for other matching methods could not only provide different point estimates but also vary in terms of estimators' variance (see

Table 4 Simulation results for larger number of manifest variables and different signal-to-noise ratios (Model C and outcome Eq. 1)

Signal-to-noise k=	500 obs.			5,000 obs.		
	Latent	Manifest	Estimated latent	Latent	Manifest	Estimated latent
5 Manifest variables						
Signal-to-noise $k = 1$						
Mean	3.245	3.349	3.347	3.087	3.199	3.198
SD	0.321	0.344	0.330	0.130	0.130	0.129
RMSE	0.404	0.490	0.479	0.156	0.238	0.237
	100.0 %	121.3 %	118.6 %	100.0 %	152.6 %	151.9 %
Signal-to-noise $k = 2$						
Mean	3.245	3.459	3.459	3.087	3.322	3.321
SD	0.321	0.350	0.337	0.130	0.131	0.131
RMSE	0.404	0.578	0.569	0.156	0.348	0.347
	100.0 %	143.1 %	140.8 %	100.0 %	223.1 %	222.4 %
10 Manifest variables						
Signal-to-noise $k = 1$						
Mean	3.245	3.315	3.307	3.088	3.152	3.152
SD	0.320	0.355	0.323	0.128	0.131	0.129
RMSE	0.403	0.475	0.445	0.155	0.201	0.199
	100.0 %	117.9 %	110.4 %	100.0 %	129.7 %	128.4 %
Signal-to-noise $k = 2$						
Mean	3.245	3.412	3.404	3.088	3.259	3.259
SD	0.320	0.358	0.330	0.128	0.131	0.129
RMSE	0.403	0.546	0.522	0.155	0.290	0.289
	100.0 %	135.5 %	129.5 %	100.0 %	187.1 %	186.5 %

Simulation results for Model C and outcome Eq. 1. The propensity score 1-to-1 nearest neighbor matching without replacement and with propensity score estimated by probit

Abadie and Imbens 2009). In the appendix, we present additional results calculated for other popular propensity score matching methods: the nearest neighbor 1-to-5 matching (where 5 best matches are used instead of 1 to construct the counterfactual outcome), local linear matching (where counterfactual outcome is constructed using local linear regression with tricube kernel function and 0.8 bandwidth) and kernel matching (with epanechnikov kernel and 0.6 bandwidth). Appendix Tables 12 and 13 present comparisons of simulation results for the four different methods and sample size 500, five manifest variables and signal-to-noise ratio 1:1.

For the least demanding model A the nearest neighbor 1-to-1 method performs well, while for more difficult models with strong selection and complex relationship between covariates and outcome kernel matching outperforms other methods. This reflects the well-known choice between bias and efficiency in matching with 1-to-1 method: providing results with the smallest bias if the pool of good matches is easily available and kernel or local linear regression matching increasing efficiency in case

when interpolation from several possible matches is necessary (Caliendo and Kopeinig 2008).

Matching based on the estimated latent variable outperforms matching based on the manifest variables in all cases except the simplest model A and gives similar results for Model B. In these models, matching on manifest variables with the kernel method provides the most precise results, as in this case all matches provide good basis for interpolation. In cases where there is strong selection between treated and controls, matching on the estimated latent variable usually performs better. These findings suggest that, while matching methods in fact provide different results, the conclusion that in smaller samples under most circumstances matching on the estimated latent variable is better than matching on the manifest variables holds despite the choice of the matching method.

4 Empirical application to human capital research

The simulation results suggest that estimating the latent variable from the manifest variables instead of using these variables directly in matching can limit the measurement error bias and increase the precision of the treatment estimates, especially with moderate sample sizes and several manifest variables available in a dataset. The suggested procedure has three steps. In the first step, the researcher has to estimate the latent variable from the observed manifest variables. This step is crucial and has to be based on a theoretically and empirically sound theory that relates observed manifest variables to the latent variable. In the remaining steps, the usual propensity score matching approach is applied. The propensity score is estimated on a set of matching covariates that includes the estimated latent variable. Matching is conducted on this propensity score, and the average treatment effect on the treated is calculated.

In this section, we apply this approach to see how selecting students into different school programs or tracks affects their achievement. This is the topic extensively analyzed in economics of education literature (see Hanushek and Woessmann 2006; Brunello and Checchi 2007; Duflo et al. 2011; Pekkarinen 2008; and for reviews Meier and Schütz 2007; Woessmann 2009). In this research, outcomes of students or adults who attended different school tracks are compared. For example, students of general-comprehensive schools are compared in outcomes with students of vocational schools. Obviously, students who go to comprehensive and vocational schools differ in many aspects, mainly in their academic achievement, as students usually are selected to different schools based on their school grades or exams. Thus, observing students' pre-selection school outcomes is crucial when assessing the impact of tracking. We use a unique dataset collected in Poland which used internationally developed tests to measure achievement of upper-secondary school students but was also linked to these students' outcomes in comprehensive lower secondary schools. In addition, the survey collected extensive information on socio-economic background of families which can be used to additionally control for selection of students into different school types.

In our application, we use a subset of data collected in the Programme for International Student Assessment (PISA). The PISA study is conducted by the OECD every three years and measures the achievement of 15-year-olds across all OECD countries

and other countries that join the project (see OECD 2007, 2009, for a detailed description of the PISA 2006 study). We use data for Poland from the PISA 2006 national study that extended the sample to cover 16- and 17-year-olds (10th and 11th grade in the Polish school system). Another, already mentioned, unique feature of the Polish dataset is supplementary information on student scores in national exams, which extends significantly the possibilities of evaluating school policies, as prior scores can be used to control for student ability or for intake levels of skills and knowledge.

We apply the approach proposed in this paper to evaluate differences in the magnitude of student progress across two types of upper secondary education: general–vocational and vocational. We use data for 16- and 17-year-olds only, as 15-year-olds are in comprehensive lower secondary schools. In 2006, there were four types of upper secondary educational programs: general, technical, general–vocational and vocational. The general–vocational schools were introduced by the reform of 2000, to replace some vocational schools with more comprehensive education but existing evidence suggests that they might not be effective in helping students develop more comprehensive skills (Jakubowski et al. 2010). The following empirical example seeks to establish whether, in fact, these schools equip students with a set of comprehensive skills not taught in purely vocational schools. In the PISA sample, we have slightly more than 1,000 observations of 16- and 17-year-olds attending these two types of upper secondary schools.

The PISA tests are suitable instruments to capture the extent to which different schools teach comprehensive skills, as they aim at testing general student literacy in mathematics, reading and science, using a general framework that defines internationally comparable measures of literacy. PISA tries to capture skills and knowledge needed in adult life, rather than those simply reflecting schools' curricula. This makes comparisons between schools more objective, and internationally developed instruments assure that they are not biased toward curriculum used in one type of Polish school.

To compare the impact of distinct types of schools on student outcomes it is necessary to control for student selection into these schools. This selection is probably based on previous student skills and knowledge, but also on other important student and family characteristics. The unique feature of the Polish PISA dataset is that students' prior scores on national exams are linked to data obtained through internationally comparable instruments. These are scores from the obligatory national exam conducted at the end of lower secondary school (at age 15) that contains two parts, one for mathematics and science, and one for humanities. We combine both scores in one measure reflecting the level of student intake knowledge and skills across disciplines.

We also use detailed data on student and family characteristics. In PISA, student background information is available in two types of indicators. In the first type, variables directly reflect student responses about their observable and easy-to-define characteristics. Among those, we use dummies denoting student gender and school grade level, parents' highest level of education measured on the ISCED scale and parents' highest occupation status measured on the ISEI scale (the International Socio-Economic Index of occupational status, see Ganzeboom et al. 1992). The second type of variables summarizes responses to several questions (so-called items) that reflect a common latent characteristic (OECD 2009, pp. 303–349). Here, we use student

responses about more than 20 types of home possessions, including consumption, educational, and cultural goods. The original PISA dataset contains four indices that summarize information on household goods: *homepos* for all home possessions, *wealth* for consumption goods (e.g., TV, DVD player, number of cars), *cultpos* for cultural possessions (e.g., poetry books), and *hedres* for educational resources (e.g., study desk). We re-estimated these indices to include additional available information. Therefore, we took one item from the *wealth* index ("having a microscope") and added it to educational resources under *hedres*. We also estimated the *cultpos* index including a question about the number of books at home that was originally not considered.

The latent constructs were estimated using the principal component analysis models based on polychoric and polyserial correlations matrix as student responses were measured on ordinal scale (see Kolenikov and Angeles 2009). The estimated indices have relatively high reliabilities in a range from 0.6 to 0.8 that are higher for Poland than most OECD countries (OECD 2009, p. 317). Correlations between items used in the same index were from 0.3 to 0.5, which is the range modeled in our simulation study. Correlations between *cultpos*, *wealth*, and *hedres* indices were relatively low, between 0.37 and 0.56. Responses within each index were also more strongly correlated with each other than with responses from other indices. This suggests that they might represent different constructs (see Jakubowski and Pokropek 2009, for additional information on these indices).

We conducted the propensity score matching three times, to see how results are affected by including the estimated latent variable instead of a set of manifest variables. First, we included all the manifest variables in the list of matching covariates, or, more precisely, we included all dummy variables denoting household items. Second, we included three indices estimated from the manifest variables and reflecting three latent variables: household wealth, household cultural possessions and household educational resources. Finally, we estimated only one latent variable using all manifest variables and reflecting all possible home possessions. We expected that the first approach would differ from the second in terms of bias and the precision of the estimates of the average treatment effect. More precisely, we expected that the second approach would provide estimates that are different and have smaller standard errors, in line with the theory and simulation results presented above. Furthermore, we expected that the third approach, based on only one latent variable instead of three, could be less efficient, as restricting a latent dimension to one limits the amount of relevant information provided in matching if there are, in fact, three distinct latent variables behind the values of the manifest variables. This example demonstrates typical problems that arise in empirical research: first, whether to estimate the latent trait instead of using manifest variables directly in matching; and second, how many latent variables should be estimated.

Tables 5 and 6 presents results for this empirical exercise with additional results presented in the appendix. Table 5 shows imbalance between students who go to vocational (control group) and general–vocational (treated group) schools in terms of the estimated indices and other matching covariates: gender, grade, HISEI (the highest occupational status of parents as measured by the International Socio-Economic Index of occupational status, see Ganzeboom et al. 1992; OECD 2009, p. 305), and lower secondary national exam scores. The table also shows how propensity score matching helps to balance these variables. Summary statistics for all matching covariates and

propensity score equations (probit regression) are presented in the appendix. Table 5 compares variable means between these two groups of students and percentage bias before and after matching. It also shows reduction in absolute percentage bias and t-test results for the nearest neighbor 1-to-1 propensity score matching with replacement, for the same 1-to-1 matching but with the common support restriction and caliper (0.05) and for kernel matching (with epanechnikov kernel and 0.6 bandwidth).

The results show important differences between students who went to general–vocational and vocational schools in terms of matching covariates. Propensity score matching is, however, able to reduce these differences to construct comparable groups of control and treated students. The mean absolute percentage bias across all covariates goes down from 49.8 to 5.7 % with 1-to-1 matching. Imposing common support for 1-to-1 matching reduces mean absolute percentage bias to 4.8 %, while with kernel matching mean absolute percentage is further reduced to 4.5 %

Table 6 shows outcome differences before and after matching. The results are presented for all three methods: the nearest neighbor 1-to-1 matching with and without common support and caliper and for kernel matching. For matching on the manifest variables only one set of results is presented per matching method, while for matching on the estimated latent variables (one overall index of home possessions or three separate indices) two sets of results are presented. The first set of results presented in columns (2), (4) and (6) was obtained by estimating the latent variable before matching, so the same index was used for all bootstrapped samples. These results reflect common practice of not re-estimating indices available in the original dataset and using the same index across all resampled samples. Clearly, this might result in biased variance estimates as calculations omit the first step of estimating latent variable. In columns (3), (5), and (7) results are presented with the latent variables re-estimated in each bootstrap sample. These results include all sources of variance in our three-step estimation method.

First, note that the differences in outcomes go down after matching. This was expected, as differences in outcomes between different types of schools are driven mainly by differences in student characteristics. However, the outcome differences remain quite large and in favor of general–vocational schools, even after adjusting for intake scores, gender, parents' education and occupation, and family resources. The average treatment effects vary across the sets of results, with the biggest gap being found in reading. Results with common support imposed and caliper matching (last column) are similar to those obtained with unrestricted 1-to-1 matching. Differences are visible only for matching on manifest variables, which is due to larger amount of observations excluded from the analysis in this case. Common support restriction was not satisfied for only 6 and 11 observations when matching on the one estimated index and matching on the three estimated indices, respectively, but for matching on the manifest variables 25 treated observations had to be excluded. Kernel matching generally provides results similar to 1-to-1 matching without restrictions, although with marginally smaller standard errors.

The results mainly demonstrate the benefits of using the estimated latent variables in matching that are related to the precision of estimates. While the average treatment effects are reasonably similar across matching methods, the standard errors are 20–40 % higher for matching conducted on the manifest variables. This is also true for

Table 5 Balance of matching covariates

Variable	Matching method	Mean		%Bias	% Reduction in absolute bias	t test	
		Treated	Control			T stat	p value
Gender (female = 1)	Unmatched	0.57	0.31	53.7		8.73	0.000
	NN 1-to-1	0.57	0.58	−1.4	97.5	−0.20	0.842
	NN 1-to-1 with common support and caliper 0.05	0.57	0.57	−0.5	99.1	−0.07	0.946
	kernel matching	0.57	0.59	−4.6	91.5	−0.68	0.499
HISEI (the highest occupational status of parents)	Unmatched	41.37	35.73	46.9		7.66	0.000
	NN 1-to-1	41.37	40.30	8.9	81.0	1.35	0.179
	NN 1-to-1 with common support and caliper 0.05	41.12	40.23	7.4	84.2	1.10	0.270
	kernel matching	41.37	42.02	−5.4	88.5	−0.78	0.436
Lower secondary school exam score in humanities	Unmatched	31.20	24.62	95.5		15.45	0.000
	NN 1-to-1	31.20	31.54	−5.0	94.8	−0.74	0.461
	NN 1-to-1 with common support and caliper 0.05	30.90	31.16	−3.7	96.1	−0.56	0.574
	kernel matching	31.20	31.07	1.9	98.0	0.29	0.775
Lower secondary school exam score in math–science	Unmatched	21.23	15.50	79.0		12.98	0.000
	NN 1-to-1	21.23	21.50	−3.7	95.4	−0.48	0.629
	NN 1-to-1 with common support and caliper 0.05	20.81	21.25	−6.1	92.3	−0.82	0.413
	kernel matching	21.23	20.92	4.3	94.6	0.58	0.563

Table 5 continued

Variable	Matching method	Mean		%Bias	% Reduction in absolute bias	t test	
		Treated	Control			T stat	p value
Grade	Unmatched	0.57	0.47	19.4		3.14	0.002
	NN 1-to-1	0.57	0.59	−5.2	73.1	−0.80	0.424
	NN 1-to-1 with common support and caliper 0.05	0.56	0.58	−3.6	81.6	−0.54	0.590
	kernel matching	0.57	0.53	5.5	71.8	0.83	0.407
The estimated index of cultural possessions	Unmatched	−0.38	−0.81	39.2		6.35	0.000
	NN 1-to-1	−0.38	−0.43	4.6	88.3	0.69	0.490
	NN 1-to-1 with common support and caliper 0.05	−0.40	−0.44	3.5	91.0	0.52	0.603
	Kernel matching	−0.38	−0.39	1.3	96.6	0.20	0.838
The estimated index of family wealth	Unmatched	−0.40	−0.69	21.0		3.38	0.001
	NN 1-to-1	−0.40	−0.37	−1.7	91.8	−0.26	0.797
	NN 1-to-1 with common support and caliper 0.05	−0.40	−0.38	−1.8	91.3	−0.27	0.789
	Kernel matching	−0.40	−0.42	1.4	93.3	0.21	0.832
The estimated index of household educational resources	Unmatched	−0.14	−0.57	43.6		7.02	0.000
	NN 1-to-1	−0.14	−0.28	15.0	65.7	2.39	0.017
	NN 1-to-1 with common support and caliper 0.05	−0.15	−0.27	12.1	72.1	1.91	0.057
	Kernel matching	−0.14	−0.25	11.6	73.4	1.85	0.065

The propensity score 1-to-1 nearest neighbor matching without replacement. The propensity score kernel matching with epanechnikov kernel and 0.6 bandwidth. Propensity score estimated by probit. HISEI stands for the highest occupational status of parents as defined by Ganzeboom et al. (1992) (see also OECD 2009, p. 305)

Table 6 Empirical example: achievement difference between general–vocational and vocational schools

Latent variables included as	Outcome	Outcome difference (ATT)						
		Before matching	1-to-1 matching		1-to-1 matching with common support and caliper = 0.05		Kernel matching	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
All manifest variables (household items)	Mathematics	71.0	53.0		54.7		54.0	
			(12.0)		(10.5)		(10.4)	
	Reading	107.1	62.9		69.2		65.3	
			(19.9)		(14.9)		(17.2)	
	Science	76.8	56.9		61.5		51.2	
			(12.7)		(10.1)		(10.7)	
Three estimated indices: household wealth, cultural possessions, and educational resources	Mathematics	71.0	54.6	57.1	52.5	55.2	55.6	55.5
			(10.1)	(10.8)	(10.0)	(10.6)	(9.6)	(9.7)
	Reading	107.1	66.4	65.1	66.1	64.7	71.4	71.4
			(14.4)	(14.7)	(14.2)	(13.6)	(14.3)	(14.0)
	Science	76.8	51.7	53.4	50.5	52.3	55.0	55.0
			(9.9)	(10.3)	(9.6)	(9.8)	(9.1)	(9.1)
One estimated index: home possessions	Mathematics	71.0	56.1	65.3	54.9	64.2	55.8	55.8
			(10.7)	(10.0)	(10.3)	(10.0)	(9.7)	(9.3)
	Reading	107.1	70.9	80.5	70.2	79.9	73.3	73.5
			(14.8)	(14.8)	(14.0)	(14.0)	(14.5)	(13.4)
	Science	76.8	55.3	61.9	54.4	61.2	55.3	55.4
			(10.1)	(9.5)	(9.9)	(9.6)	(9.1)	(8.6)

Number of treated: 461; Number of controls: 607; Standard errors are given in parentheses and were obtained by bootstrapping clusters at the school level; In columns (2), (4) and (6) standard errors do not include estimating the latent variable. In columns (3), (5), and (7) standard errors do include the additional step to estimate the latent variable

standard errors that include error for the estimation of the latent variable as they differ only slightly from the errors calculated without this step. This suggests that in this case adding errors related to the estimation of the latent variable was not crucial for the results.

The results are quite similar for matching on the three estimated latent variables and for matching on the one estimated variable. Standard errors are slightly higher when matching on only one instead of three latent variables, which suggests that our data have three-dimensional latent structure. However, standard errors are also much lower in this case, compared to matching directly on all manifest variables.

This example confirms our main findings from the simulation study: benefits are gained from matching on estimated latent variables rather than on a set of manifest variables, at least with moderate sample sizes and a well-developed latent variables framework. Matching on three estimated latent variables lowered standard errors of the treatment effects. Matching on one latent variable gave similar results, which were still better than those obtained with matching directly on manifest variables.

5 Summary

This paper demonstrates how modeling and including the latent variable in the propensity score matching might limit the bias and increase precision of the treatment estimates in comparison to a more standard approach when matching is conducted directly on the set of manifest variables, even if they reflect the same latent trait. Including the estimated latent variable can limit the measurement error bias in the propensity score matching estimates introduced by a noisy signal contained in the manifest variables. We present simulation studies demonstrating how our approach helps limit the bias in average treatment effect estimates, and the range of efficiency gains provided by incorporating the latent variable in matching. Finally, we apply these to real educational data, showing the importance of our findings on an empirical example.

We find that in various models studied estimating the latent variable and using this estimate for propensity score matching decreases the measurement error bias in the average treatment effects and lowers the variance of the average treatment estimators in smaller samples. While in some cases the bias and the variance are smaller even for large samples, in general, matching on the estimated latent variable provides similar results or only modest gains compared to matching on manifest variables with large sample sizes. Thus, while in larger samples the choice might be less important, in smaller samples we suggest following our three-step strategy: estimate the latent variable using observed manifest variables; use this estimate to estimate the propensity score; and match on this propensity score to estimate the average treatment effects.

Appendix

See Tables 7, 8, 9, 10, 11, 12, 13 and 14.

Table 7 Summary statistics of matching covariates for the empirical application

Variable	Mean	SD	Min.	Max.
Gender (female = 1)	0.43	0.49	0	1
HISEI (the highest occupational status of parents)	38.16	12.23	16	85
Lower secondary school exam score in humanities	27.46	7.62	5	47.17
Lower secondary school exam score in math–science	17.98	7.69	2	44.77
Grade	0.51	0.50	0	1
The estimated index of cultural possessions	−0.62	1.12	−2.14	1.46
The estimated index of family wealth	−0.57	1.44	−4.00	3.44
The estimated index of household educational resources	−0.38	1.01	−4.71	0.90

Statistics calculated for a sample of 1068 students with data available for all covariates; HISEI stands for the highest occupational status of parents as defined by Ganzeboom et al. (1992) (see also OECD 2009, p. 305)

Table 8 Probit regression for the propensity score

Outcome: general–vocational school = 1 vs. vocational school = 0								
Variable	Coef.	SE	Variable	Coef.	SE	Variable	Coef.	SE
Gender (female = 1)	0.906	0.103	Gender	0.830	0.098	Gender	0.839	0.094
HISEI (the highest occupational status of parents)	0.019	0.004	HISEI	0.020	0.004	HISEI	0.019	0.004
Lower secondary school exam score in humanities	0.044	0.008	Score in humanities	0.047	0.008	Score in humanities	0.045	0.008
Lower secondary school exam score in math–science	0.053	0.008	Score in math–science	0.051	0.008	Score in math–science	0.051	0.008
Grade	0.393	0.093	Grade	0.364	0.089	Grade	0.367	0.089
Which of the following are in your home?			The estimated index of family wealth	0.025	0.039	The estimated index of all household possessions	0.131	0.029
Room of your own	0.083	0.111						
Link to the internet	0.260	0.109						
Dishwasher	-0.212	0.126						
DVD	0.125	0.134						
Cable TV with at least 30 channels	0.157	0.104						
Digital camera	-0.097	0.112						
How many of these are there at your home?								
Cellular phones	0.044	0.064						
Televisions	0.018	0.073						

Table 8 continued

Outcome: general–vocational school = 1 vs. vocational school = 0								
Variable	Coef.	SE	Variable	Coef.	SE	Variable	Coef.	SE
Computers	−0.108	0.124						
Cars	−0.096	0.061						
Rooms with a bath or shower	−0.151	0.090						
Which of the following are in your home?			The estimated index of cultural possessions	0.000	0.045			
Classic literature	0.015	0.114						
Books of poetry	0.142	0.112						
Works of art (e.g. paintings)	−0.142	0.107						
How many books are there in your home?	0.054	0.074						
Which of the following are in your home?			The estimated index of household educational resources	0.206	0.056			
Desk study to study at	0.175	0.169						
Quiet place to study	0.052	0.162						
Computer you can use for school work	0.490	0.171						
Educational software	0.243	0.125						
Your own calculator	−0.632	0.190						
Books to help with your school work	0.040	0.183						
Dictionary	0.144	0.235						
Telescope or microscope	−0.426	0.174						
Constant	−4.005	0.406	Constant	−3.632	0.264	Constant	−3.535	0.257

Sample size of 1,068 students. The variable names listed in the left column are original names of questionnaire items available in the PISA dataset from which the indices listed on the right were estimated (for details see OECD 2009, p. 316). HISEI stands for the highest occupational status of parents as defined by Ganzeboom et al. (1992) (see also OECD 2009, p. 305)

Table 9 Additional simulation results for models A, B and D and outcome Eq. 1

	Model A			Model B			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
Noise: 1, Cov: 5, Obs: 500									
Mean	3.000	2.996	3.004	3.042	3.084	3.083	3.809	3.938	3.905
SD	0.279	0.299	0.284	0.161	0.184	0.169	0.447	0.417	0.422
RMSE	0.279	0.299	0.284	0.166	0.202	0.188	0.924	1.026	0.999
	100.0 %	107.2 %	101.8 %	100.0 %	121.7 %	113.3 %	100.0 %	111.0 %	108.1 %
Noise: 1, Cov: 5, Obs: 5,000									
Mean	3.000	2.999	3.000	3.007	3.050	3.049	3.714	3.839	3.830
SD	0.087	0.094	0.089	0.046	0.049	0.049	0.397	0.316	0.316
RMSE	0.087	0.094	0.089	0.046	0.070	0.069	0.817	0.897	0.888
	100.0 %	108.0 %	102.3 %	100.0 %	152.2 %	150.0 %	100.0 %	109.8 %	108.7 %
Noise: 1, Cov: 10, Obs: 500									
Mean	2.997	2.999	2.998	3.040	3.068	3.062	3.815	3.946	3.870
SD	0.278	0.306	0.278	0.161	0.196	0.166	0.445	0.415	0.429
RMSE	0.278	0.306	0.278	0.166	0.208	0.177	0.929	1.033	0.970
	100.0 %	110.1 %	100.0 %	100.0 %	125.3 %	106.6 %	100.0 %	111.2 %	104.4 %
Noise: 1, Cov: 10, Obs: 5,000									
Mean	2.999	2.999	3.000	3.007	3.032	3.032	3.708	3.808	3.790
SD	0.089	0.096	0.088	0.046	0.049	0.048	0.393	0.333	0.345
RMSE	0.089	0.096	0.088	0.046	0.058	0.058	0.809	0.874	0.862
	100.0 %	107.9 %	98.9 %	100.0 %	126.1 %	126.1 %	100.0 %	108.0 %	106.6 %
Noise: 2, Cov: 5, Obs: 500									
Mean	3.000	2.998	3.002	3.042	3.132	3.130	3.809	4.014	3.988
SD	0.279	0.299	0.281	0.161	0.193	0.178	0.447	0.399	0.404

Table 9 continued

	Model A			Model B			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
RMSE	0.279	0.299	0.281	0.166	0.234	0.221	0.924	1.090	1.068
	100.0 %	107.2 %	100.7 %	100.0 %	141.0 %	133.1 %	100.0 %	118.0 %	115.6 %
Noise: 2, Cov: 5, Obs: 5,000									
Mean	3.000	3.001	3.000	3.007	3.099	3.098	3.714	3.932	3.927
SD	0.087	0.095	0.091	0.046	0.052	0.052	0.397	0.274	0.275
RMSE	0.087	0.095	0.091	0.046	0.112	0.111	0.817	0.971	0.967
	100.0 %	109.2 %	104.6 %	100.0 %	243.5 %	241.3 %	100.0 %	118.8 %	118.4 %
Noise: 2, Cov: 10, Obs: 500									
Mean	2.997	3.001	3.001	3.040	3.107	3.105	3.815	4.011	3.944
SD	0.278	0.317	0.286	0.161	0.203	0.173	0.445	0.403	0.410
RMSE	0.278	0.317	0.286	0.166	0.229	0.202	0.929	1.088	1.029
	100.0 %	114.0 %	102.9 %	100.0 %	138.0 %	121.7 %	100.0 %	117.1 %	110.8 %
Noise: 2, Cov: 10, Obs: 5,000									
Mean	2.999	2.998	3.000	3.007	3.074	3.073	3.708	3.890	3.880
SD	0.089	0.098	0.088	0.046	0.052	0.051	0.393	0.289	0.297
RMSE	0.089	0.098	0.088	0.046	0.090	0.089	0.809	0.935	0.929
	100.0 %	110.1 %	98.9 %	100.0 %	195.7 %	193.5 %	100.0 %	115.6 %	114.8 %

Mean estimate, standard deviation and root mean squared error (also as the percentage comparing to the model with latent variable) of the average treatment effect. Results for different noise-to-signal ratios, sample size and the number of manifest variables

Table 10 Simulation results for four models and outcome Eq. 2

	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
Noise: 1, Cov: 5, Obs: 500												
Mean	3.002	2.990	2.993	3.256	3.370	3.368	4.135	4.372	4.364	5.554	5.917	5.799
SD	0.708	0.759	0.718	0.644	0.713	0.663	1.199	1.240	1.197	1.668	1.372	1.461
RMSE	0.708	0.759	0.718	0.693	0.803	0.758	1.651	1.849	1.814	3.051	3.223	3.157
	100 %	107.2 %	101.4 %	100	115.9 %	109.4 %	100 %	112.0 %	109.9 %	100 %	105.6 %	103.5 %
Noise: 1, Cov: 5, Obs: 5,000												
Mean	3.003	2.999	3.000	3.051	3.183	3.180	3.519	3.844	3.847	5.100	5.588	5.556
SD	0.224	0.236	0.227	0.205	0.219	0.215	0.619	0.587	0.582	1.783	1.033	1.061
RMSE	0.224	0.236	0.227	0.211	0.285	0.281	0.807	1.028	1.027	2.755	2.786	2.767
	100 %	105.4 %	101.3 %	100 %	135.1 %	133.2 %	100 %	127.4 %	127.3 %	100 %	101.1 %	100.4 %
Noise: 1, Cov: 10, Obs: 500												
Mean	3.007	3.008	3.001	3.257	3.319	3.319	4.158	4.332	4.292	5.556	5.967	5.713
SD	0.711	0.777	0.713	0.642	0.760	0.660	1.179	1.268	1.186	1.582	1.345	1.455
RMSE	0.711	0.777	0.713	0.692	0.824	0.733	1.652	1.839	1.754	3.006	3.257	3.078
	100 %	109.3 %	100.3 %	100 %	119.1 %	105.9 %	100 %	111.3 %	106.2 %	100 %	108.3 %	102.4 %
Noise: 1, Cov: 10, Obs: 5,000												
Mean	3.002	3.000	3.002	3.053	3.129	3.130	3.500	3.708	3.704	5.112	5.587	5.525
SD	0.223	0.241	0.225	0.205	0.218	0.213	0.617	0.603	0.592	1.754	1.125	1.212
RMSE	0.223	0.241	0.225	0.211	0.253	0.249	0.794	0.930	0.920	2.746	2.821	2.801
	100 %	108.1 %	100.9 %	100 %	119.9 %	118.0 %	100 %	117.1 %	115.9 %	100 %	102.7 %	102.0 %
Noise: 2, Cov: 5, Obs: 500												
Mean	3.002	2.994	2.998	3.256	3.494	3.488	4.135	4.597	4.599	5.554	5.997	5.895
SD	0.708	0.778	0.729	0.644	0.723	0.685	1.199	1.219	1.163	1.668	1.252	1.295

Table 10 continued

	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
RMSE	0.708	0.778	0.729	0.693	0.876	0.841	1.651	2.009	1.977	3.051	3.247	3.171
	100 %	109.9 %	103.0 %	100 %	126.4 %	121.4 %	100 %	121.7 %	119.7 %	100 %	106.4 %	103.9 %
Noise: 2, Cov: 5, Obs: 5,000												
Mean	3.003	3.002	3.000	3.051	3.327	3.327	3.519	4.159	4.157	5.100	5.605	5.579
SD	0.224	0.236	0.234	0.205	0.225	0.222	0.619	0.533	0.536	1.783	0.845	0.901
RMSE	0.224	0.236	0.234	0.211	0.397	0.395	0.807	1.276	1.275	2.755	2.738	2.732
	100 %	105.4 %	104.5 %	100 %	188.2 %	187.2 %	100 %	158.1 %	158.0 %	100 %	99.4 %	99.2 %
Noise: 2, Cov: 10, Obs: 500												
Mean	3.007	3.007	3.008	3.257	3.430	3.432	4.158	4.542	4.520	5.556	6.065	5.846
SD	0.711	0.785	0.725	0.642	0.776	0.676	1.179	1.249	1.160	1.582	1.219	1.325
RMSE	0.711	0.785	0.725	0.692	0.887	0.803	1.652	1.985	1.912	3.006	3.299	3.139
	100 %	110.4 %	102.0 %	100 %	128.2 %	116.0 %	100 %	120.2 %	115.7 %	100 %	109.7 %	104.4 %
Noise: 2, Cov: 10, Obs: 5,000												
Mean	3.002	2.998	3.000	3.053	3.252	3.253	3.500	3.998	3.997	5.112	5.634	5.586
SD	0.223	0.244	0.229	0.205	0.225	0.218	0.617	0.555	0.547	1.754	0.904	0.930
RMSE	0.223	0.244	0.229	0.211	0.338	0.334	0.794	1.142	1.137	2.746	2.785	2.748
	100 %	109.4 %	102.7 %	100 %	160.2 %	158.3 %	100 %	143.8 %	143.2 %	100 %	101.4 %	100.1 %

Mean estimate, standard deviation and root mean squared error (also as the percentage comparing to the model with latent variable) of the average treatment effect. Results for different noise-to-signal ratios, sample size and the number of manifest variables

Table 11 Simulation results for four models and outcome Eq. 3

	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
Noise: 1, Cov: 5, Obs: 500												
Mean	3.000	2.996	3.004	3.398	3.440	3.438	3.871	3.975	3.973	4.328	4.457	4.424
SD	0.290	0.312	0.299	0.176	0.200	0.186	0.331	0.355	0.341	0.459	0.431	0.436
RMSE	0.290	0.312	0.299	0.435	0.483	0.476	0.932	1.038	1.031	1.405	1.519	1.490
	100 %	107.6 %	103.1 %	100 %	111.0 %	109.4 %	100 %	111.4 %	110.6 %	100 %	108.1 %	106.0 %
Noise: 1, Cov: 5, Obs: 5,000												
Mean	3.000	2.999	3.000	3.363	3.406	3.406	3.713	3.825	3.824	4.233	4.359	4.350
SD	0.091	0.098	0.093	0.050	0.055	0.055	0.131	0.132	0.132	0.398	0.317	0.317
RMSE	0.091	0.098	0.093	0.367	0.410	0.409	0.725	0.836	0.834	1.296	1.395	1.386
	100 %	107.7 %	102.2 %	100 %	111.7 %	111.4 %	100 %	115.3 %	115.0 %	100 %	107.6 %	106.9 %
Noise: 1, Cov: 10, Obs: 500												
Mean	2.996	2.999	2.997	3.395	3.424	3.418	3.870	3.940	3.931	4.333	4.464	4.388
SD	0.291	0.320	0.291	0.176	0.210	0.182	0.330	0.365	0.334	0.458	0.429	0.443
RMSE	0.291	0.319	0.291	0.432	0.473	0.456	0.930	1.008	0.989	1.409	1.526	1.457
	100 %	109.6 %	100 %	100 %	109.5 %	105.6 %	100 %	108.4 %	106.3 %	100 %	108.3 %	103.4 %
Noise: 1, Cov: 10, Obs: 5,000												
Mean	2.999	2.999	3.000	3.363	3.388	3.388	3.713	3.777	3.777	4.227	4.327	4.309
SD	0.093	0.100	0.093	0.050	0.054	0.053	0.130	0.133	0.131	0.394	0.335	0.346
RMSE	0.093	0.100	0.093	0.366	0.392	0.392	0.725	0.789	0.788	1.289	1.369	1.354
	100 %	107.5 %	100 %	100 %	107.1 %	107.1 %	100 %	108.8 %	108.7 %	100 %	106.2 %	105.0 %

Table 11 continued

	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
Noise: 2, Cov: 5, Obs: 500												
Mean	3.000	2.997	3.002	3.398	3.488	3.486	3.871	4.085	4.085	4.328	4.534	4.507
SD	0.290	0.316	0.300	0.176	0.212	0.199	0.331	0.363	0.349	0.459	0.416	0.421
RMSE	0.290	0.316	0.300	0.435	0.532	0.525	0.932	1.144	1.140	1.405	1.589	1.565
	100 %	109.0 %	103.4 %	100 %	122.3 %	120.7 %	100 %	122.7 %	122.3 %	100 %	113.1 %	111.4 %
Noise: 2, Cov: 5, Obs: 5,000												
Mean	3.000	3.001	3.000	3.363	3.455	3.455	3.713	3.948	3.947	4.233	4.451	4.446
SD	0.091	0.101	0.097	0.050	0.059	0.058	0.131	0.134	0.134	0.398	0.276	0.277
RMSE	0.091	0.101	0.097	0.367	0.459	0.459	0.725	0.958	0.957	1.296	1.477	1.473
	100 %	111.0 %	106.6 %	100 %	125.1 %	125.1 %	100 %	132.1 %	132.0 %	100 %	114.0 %	113.7 %
Noise: 2, Cov: 10, Obs: 500												
Mean	2.996	3.000	3.000	3.395	3.463	3.460	3.870	4.037	4.029	4.333	4.529	4.462
SD	0.291	0.333	0.302	0.176	0.218	0.191	0.330	0.369	0.342	0.458	0.419	0.425
RMSE	0.291	0.333	0.302	0.432	0.511	0.499	0.930	1.101	1.085	1.409	1.585	1.523
	100 %	114.4 %	103.8 %	100 %	118.3 %	115.5 %	100 %	118.4 %	116.7 %	100 %	112.5 %	108.1 %
Noise: 2, Cov: 10, Obs: 5,000												
Mean	2.999	2.998	2.999	3.363	3.430	3.429	3.713	3.884	3.884	4.227	4.409	4.399
SD	0.093	0.103	0.093	0.050	0.057	0.057	0.130	0.134	0.131	0.394	0.291	0.299
RMSE	0.093	0.103	0.093	0.366	0.434	0.433	0.725	0.895	0.894	1.289	1.438	1.431
	100 %	110.8 %	100 %	100 %	118.6 %	118.3 %	100 %	123.4 %	123.3 %	100 %	111.6 %	111.0 %

Mean estimate, standard deviation and root mean squared error (also as the percentage comparing to the model with latent variable) of the average treatment effect. Results for different noise-to-signal ratios, sample size and the number of manifest variables

Table 12 Simulation results for four models, outcome Eq. 1 and different propensity score matching methods: nearest neighbor 1-to-1 matching (1-to-1), nearest neighbor 1-to-5 matching (1-to-5), local linear matching (llr) and kernel matching (kernel)

Outcome 1	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
1-to-1 NN												
Mean	3.000	2.996	3.004	3.042	3.084	3.083	3.245	3.349	3.347	3.809	3.938	3.905
SD	0.279	0.299	0.284	0.161	0.184	0.169	0.321	0.344	0.330	0.447	0.417	0.422
RMSE	0.279	0.299	0.284	0.166	0.202	0.188	0.404	0.490	0.479	0.924	1.026	0.999
	100 %	107.2 %	101.8 %	100 %	121.7 %	113.3 %	100 %	121.3 %	118.6 %	100 %	111.0 %	108.1 %
1-to-5 NN												
Mean	2.998	3.000	3.000	3.102	3.144	3.141	3.434	3.532	3.523	3.875	4.000	3.965
SD	0.154	0.165	0.155	0.132	0.145	0.138	0.246	0.257	0.249	0.260	0.259	0.258
RMSE	0.154	0.165	0.155	0.167	0.204	0.198	0.499	0.591	0.579	0.913	1.033	0.999
	100 %	107.1 %	100.6 %	100 %	122.2 %	118.6 %	100 %	118.4 %	116.0 %	100 %	113.1 %	109.4 %
LLR												
Mean	2.999	2.999	3.000	3.040	3.081	3.080	3.244	3.348	3.343	3.806	3.940	3.906
SD	0.126	0.138	0.130	0.132	0.148	0.138	0.294	0.314	0.302	0.374	0.359	0.362
RMSE	0.126	0.138	0.130	0.138	0.169	0.160	0.382	0.469	0.457	0.889	1.006	0.976
	100 %	109.5 %	103.2 %	100 %	122.5 %	115.9 %	100 %	122.8 %	119.6 %	100 %	113.2 %	109.8 %
Kernel												
Mean	3.000	3.000	3.000	3.059	3.100	3.101	3.189	3.294	3.292	3.166	3.379	3.317
SD	0.139	0.124	0.144	0.116	0.130	0.123	0.246	0.264	0.253	0.195	0.226	0.211
RMSE	0.139	0.124	0.144	0.130	0.164	0.159	0.310	0.395	0.386	0.256	0.441	0.380
	100 %	89.2 %	103.6 %	100 %	126.2 %	122.3 %	100 %	127.4 %	124.5 %	100 %	172.3 %	148.4 %

Mean estimate, standard deviation, root mean squared error (also as the percentage comparing to the model with latent variable) of the average treatment effect. Results for sample size of 500, noise-to-signal ratio 1:1 and five manifest variables

Table 13 Simulation results for four models, outcome Eq. 2 and different propensity score matching methods: nearest neighbor 1-to-1 matching (1-to-1), nearest neighbor 1-to-5 matching (1-to-5), local linear matching (llr) and kernel matching (kernel)

Outcome 2	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
1-to-1 NN												
Mean	3.002	2.990	2.993	3.256	3.370	3.368	4.135	4.372	4.364	5.554	5.917	5.799
SD	0.708	0.759	0.718	0.644	0.713	0.663	1.199	1.240	1.197	1.668	1.372	1.461
RMSE	0.708	0.759	0.718	0.693	0.803	0.758	1.651	1.849	1.814	3.051	3.223	3.157
	100 %	107.2 %	101.4 %	100 %	115.9 %	109.4 %	100 %	112.0 %	109.9 %	100 %	105.6 %	103.5 %
1-to-5 NN												
Mean	3.007	3.004	3.008	3.518	3.610	3.613	4.715	4.913	4.896	5.847	6.091	6.005
SD	0.500	0.538	0.514	0.517	0.562	0.535	0.851	0.863	0.849	0.865	0.815	0.811
RMSE	0.500	0.538	0.514	0.732	0.830	0.813	1.915	2.098	2.077	2.975	3.196	3.112
	100 %	107.6 %	102.8 %	100 %	113.4 %	111.1 %	100 %	109.6 %	108.5 %	100 %	107.4 %	104.6 %
LLR												
Mean	2.995	2.995	2.998	3.211	3.329	3.326	4.116	4.356	4.344	5.571	5.923	5.803
SD	0.471	0.506	0.485	0.570	0.627	0.589	1.118	1.147	1.103	1.413	1.217	1.279
RMSE	0.471	0.506	0.485	0.607	0.708	0.674	1.580	1.777	1.739	2.934	3.167	3.081
	100 %	107.4 %	103.0 %	100 %	116.6 %	111.0 %	100 %	112.5 %	110.1 %	100 %	107.9 %	105.0 %
Kernel												
Mean	3.095	3.032	3.094	3.190	3.306	3.309	3.778	4.022	4.024	3.342	3.920	3.719
SD	0.474	0.476	0.487	0.458	0.518	0.482	0.978	1.013	0.981	0.678	0.730	0.695
RMSE	0.484	0.477	0.496	0.496	0.601	0.573	1.250	1.439	1.418	0.759	1.174	1.000
	100 %	98.6 %	102.5 %	100 %	121.2 %	115.5 %	100 %	115.1 %	113.4 %	100 %	154.7 %	131.8 %

Mean estimate, standard deviation, root mean squared error (also as the percentage comparing to the model with latent variable) of the average treatment effect. Results for sample size of 500, noise-to-signal ratio 1:1 and five manifest variables

Table 14 Simulation results for four models, outcome Eq. 3 and different propensity score matching methods: nearest neighbor 1-to-1 matching (1-to-1), nearest neighbor 1-to-5 matching (1-to-5), local linear matching (llr) and kernel matching (kernel)

Outcome 3	Model A			Model B			Model C			Model D		
	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent	Latent	Manifest	Est. latent
1-to-1 NN												
Mean	3.000	2.996	3.004	3.398	3.440	3.438	3.871	3.975	3.973	4.328	4.457	4.424
SD	0.290	0.312	0.299	0.176	0.200	0.186	0.331	0.355	0.341	0.459	0.431	0.436
RMSE	0.290	0.312	0.299	0.435	0.483	0.476	0.932	1.038	1.031	1.405	1.519	1.490
	100 %	107.6 %	103.1 %	100 %	111.0 %	109.4 %	100 %	111.4 %	110.6 %	100 %	108.1 %	106.0 %
1-to-5 NN												
Mean	3.000	3.003	3.004	3.455	3.497	3.495	4.060	4.158	4.151	4.401	4.527	4.490
SD	0.178	0.190	0.185	0.150	0.166	0.160	0.264	0.276	0.269	0.281	0.286	0.281
RMSE	0.178	0.190	0.185	0.479	0.524	0.520	1.093	1.191	1.182	1.429	1.554	1.516
	100 %	106.7 %	103.9 %	100 %	109.4 %	108.6 %	100 %	109.0 %	108.1 %	100 %	108.7 %	106.1 %
LLR												
Mean	3.001	3.002	3.003	3.392	3.434	3.434	3.871	3.977	3.970	4.337	4.474	4.432
SD	0.152	0.166	0.162	0.148	0.164	0.158	0.307	0.326	0.315	0.391	0.381	0.379
RMSE	0.152	0.166	0.162	0.419	0.463	0.462	0.923	1.030	1.019	1.393	1.522	1.482
	100 %	109.2 %	106.6 %	100 %	110.5 %	110.3 %	100 %	111.6 %	110.4 %	100 %	109.3 %	106.4 %
Kernel												
Mean	3.001	3.001	3.001	3.396	3.438	3.440	3.767	3.878	3.874	3.386	3.677	3.593
SD	0.187	0.170	0.194	0.132	0.149	0.143	0.279	0.295	0.284	0.223	0.262	0.246
RMSE	0.187	0.170	0.194	0.417	0.462	0.463	0.816	0.926	0.919	0.445	0.726	0.642
	100 %	90.9 %	103.7 %	100 %	110.8 %	111.0 %	100 %	113.5 %	112.6 %	100 %	163.1 %	144.3 %

Mean estimate, standard deviation, root mean squared error (also as the percentage comparing to the model with latent variable) of the average treatment effect. Results for sample size of 500, noise-to-signal ratio 1:1 and five manifest variables

References

Abadie A, Imbens G (2009) Matching on the estimated propensity score, NBER Working Papers 15301. National Bureau of Economic Research Inc., Cambridge

Abbring JH, Heckman JJ (2007) Econometric evaluation of social programs, Part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation, Chap. 72. In: Heckman JJ, Leamer EE (eds) Handbook of econometrics, vol 6. Elsevier, Amsterdam

Agodini R, Dynarski M (2004) Are experiments the only option? A look at dropout prevention programs. Rev Econ Stat 86(1):180–194

Aigner D, Hsiao C, Kapteyn A, Wansbeek T (1984) Latent variable models in econometrics, Chap. 23. In: Griliches Z, Intriligator MD (eds) Handbook of econometrics, vol 2, 1st edn. Elsevier, Amsterdam, pp 1321–1393

Barg K, Beblo M (2009) Does marriage pay more than cohabitation? J Econ Stud 36(6):552–570

Battistin E, Chesher A (2009) Treatment effect estimation with covariate measurement error, CeMMAP Working Papers CWP25/09. Centre for Microdata Methods and Practice, Institute for Fiscal Studies, London

Becker S, Ichino A (2002) Estimation of average treatment effects based on propensity scores. Stat J 2(4):358–377

Ben-Ner A, Kramer A, Levy O (2008) Economic and hypothetical dictator game experiments: incentive effects at the individual level. J Socio-Econ 37(5):1775–1784

Brunello G, Checchi D (2007) Does school tracking affect equality of opportunity? New international evidence. Econ Policy 22:781–861

Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. J Econ Surv 22(1):31–72, 02

Carroll RJ, Ruppert D, Stefanski LA (1995) Measurement error in nonlinear models. Chapman & Hall/CRC, London

Cochran W, Rubin D (1973) Controlling bias in observational studies: a review. Sankhyā (1961–2002) 35(4):417–446

Duflo E, Dupas P, Kremer M (2011) Peer effects. Teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. Am Econ Rev 101(5):1739–1774

Fairlie R, Holleran W (2012) Entrepreneurship training, risk aversion and other personality traits: evidence from a random experiment. J Econ Psychol 33(2):366–378

Ganzeboom H, De Graaf P, Treiman D (1992) A standard international socio-economic index of occupational status. Soc Sci Res 21(1):1–56

Girtz R (2012) The effects of personality traits on wages: a matching approach. LABOUR 26(4):455–471

Grabner C, Hahn H, Leopold-Wildburger U, Pickl S (2009) Analyzing the sustainability of harvesting behavior and the relationship to personality traits in a simulated Lotka–Volterra biotope. Eur J Oper Res 193(3):761–767

Hanushek E, Woessmann L (2006) Does educational tracking affect performance and inequality? Differences- in-differences evidence across countries. Econ J 116(510):C63–C76, 03

Heckman JJ, Ichimura H, Todd PE (1998) Matching as an econometric evaluation estimator. Rev Econ Stud 65:261–294

Heineck G, Anger S (2010) he returns to cognitive abilities and personality traits in Germany. Labour Econ 17(3):535–546

Jakubowski M, Pokropek A (2009) Family income or knowledge? Decomposing the impact of socio-economic status on student outcomes and selection into different types of schooling. Paper presented at the PISA research conference, Sept 2009, Kiel

Jakubowski M, Patrinos HA, Porta EE, Wiśniewski J (2010) The impact of the 1999 education reform in Poland, Policy Research Working Paper Series 5263. The World Bank, Washington, DC

Jalan J, Ravallion M (2003) Estimating the benefit incidence of an antipoverty program by propensity-score matching. J Bus Econ Stat Am Stat Assoc 21(1):19–30

Joreskog KG (1971) Statistical analysis of sets of congeneric tests. Psychometrika 36:109–132

Kmenta J (1991) Latent variables in econometrics. Stat Neerl 45:73–84

Kolenikov S, Angeles G (2009) Socioeconomic status measurement with discrete proxy variables: is principal component analysis a reliable answer? Rev Income Wealth 55:128–165

Lechner M (2000) An evaluation of public-sector-sponsored continuous vocational training programs in East Germany. J Human Resour 35(2):347–375

Leuven E, Sianesi B (2003) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Version 4.0.6. http://ideas.repec.org/c/boc/bocode/s432001.html
McKenzie D (2005) Measuring inequality with asset indicators. J Popul Econ 18(2):229–260, 06
Meier V, Schütz G (2007) The economics of tracking and non-tracking, Ifo Working Paper No. 50. Ifo Institute for Economic Research at the University of Munich
OECD (2007) PISA 2006 science competencies for tomorrow's world. OECD, Paris
OECD (2009) PISA 2006 Technical Report. OECD, Paris
Ovchinnikova N, Czap H, Lynne G, Larimer C (2009) 'I don't want to be selling my soul': two experiments in environmental economics. J Socio-Econ 38(2):221–229
Pekkarinen T (2008) Gender differences in educational attainment: evidence on the role of tracking from a Finnish quasi-experiment. Scand J Econ 110(4):807–825
Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
Rosenbaum P, Rubin D (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39(1):33–38
Rubin D (1973) Matching to remove bias in observational studies. Biometrics 29:159–183
Skrondal A, Rabe-Hesketh S (2004) Generalized latent variable modeling: multilevel, longitudinal and structural equation models. Chapman & Hall/CRC, Boca Raton
Wansbeek T, Meijer E (2000) Measurement error and latent variables in econometrics. Elsevier Science Limited, Oxford
White M, Killeen J (2002) The effect of careers guidance for employed adults on continuing education: assessing the importance of attitudinal information. J R Stat Soc Ser A 165(1):83–95
Woessmann L (2009) International evidence on school tracking: a review. CESifo DICE Rep 7(1):26–34, 04