Chilean and High School Dropout Calculations to "Testing the Correlated Random Coefficient Model"*

James J. Heckman
University of Chicago,
University College Dublin
Cowles Foundation, Yale University
and the American Bar Foundation

Daniel Schmierer
University of Chicago

Sergio Urzua
Northwestern University

March 4, 2009

*This document is available on the web at http://jenni.uchicago.edu/testing_random/

Contents

List of Figures

List of Tables

1 High School Diploma vs. High School Dropout

Estimating the returns to graduating high school versus dropping out has received less attention in the literature.* For this analysis, we follow the example in (Heckman and Vytlacil, 2007, p. 4953) and use data from the NLSY79. We let $D = 1$ if an individual's highest level of education is a high school diploma and $D = 0$ if the individual is a high school dropout (not a GED recipient). This gives a sample size of 928. The outcome of interest is the log of average hourly wages between ages 28 and 32.

In order to estimate the propensity scores, we run a probit for D on the following independent variables (Z): individual's cognitive and noncognitive abilities, father's highest grade completed, mother's highest grade completed, number of siblings, family income in 1979, wages and unemployment rates of local dropouts, wages and unemployment rates of local high school graduates, indicators for black and Hispanic, indicators for residence in the South and urban residence at age 14, and year of birth indicators.

Using the fitted values from this probit we form our estimates of the propensity score, $P(Z)$. We then regress the outcome variable on polynomials in the propensity score plus the following regressors (X): job tenure, job tenure squared, experience, experience squared, AFQT score, noncognitive score, marital status, and year of birth indicators.

First, we conduct the conditional moment test of the null hypothesis of no selection on the gain. The result of this test is shown in panel A of Table 3 and shows that we are unable to reject the null. We next implement our series test by estimating (14) for different degrees of the polynomial in $P(Z)$. Table 3, panel B, contains the probability values from these tests on this data and gives the results from our overall test for the presence of nonlinearity in these models. Linearity is not rejected in any of the models. This means that we cannot rule out the case of a constant-MTE and so it may not be necessary to deal with the additional complications of allowing for sorting into schooling based gains.

We then test for linearity by calculating the IV estimate using observations with propen-

*See, however, Heckman, Lochner, and Todd (2003, 2006, 2008) and Heckman and LaFontaine (2006).

Table 1: College participation vs. stopping at high school: tests for selection on the gain to treatment, excluding ability measures.

A. Conditional moment test[a]

Probability value of test:	0.5832
Outcome of test:	Do not reject

B. Series test[b]

Degree of polynomial		2	3	4	5
Joint test (no bias correction)		0.186	0.359	0.399	0.557
Joint test (with bias correction)		0.114	0.372	0.417	0.754
p-value of test (no bias correction):	0.186				
p-value of test (bias correction):	0.114				
Critical value:	0.024				
Outcome of test:	Do not reject				

C. IV estimates above and below the median[c]

	Whole sample	Below	Above	Prob. value of test
With interactions	0.1944	0.8999	4.1001	0.6239
(evaluated at mean X)	(2.3773)	(1.8888)	(2.8674)	
Without interactions	0.5103	0.6986	0.1432	0.4538
	(0.1938)	(0.3873)	(0.6219)	
Outcome of test:	Do not reject			

D. IV estimates by quartiles of the propensity score[d]

	1st quartile	2nd quartile	3rd quartile	4th quartile
Estimate:	1.2189	0.2688	0.6100	-0.4594
Standard error:	(0.8462)	(6.7533)	(11.2250)	(5.4924)
Smallest probability value from pairwise tests:		0.4761		
Outcome of test:		Do not reject		

E. IV estimates using different instruments

Instrument:	Local college	Local college * mother's education	Local wages at age 17
Estimate:	-0.0632	-0.2242	0.8375
Standard error:	(5.6508)	(11.5826)	(0.4143)

F. Test of equality of IV estimates using different instruments[e]

Instrument:	Local college	Local college * mother's education	Local wages at age 17
Local college	.	.	.
Local college * mother's education	0.809	.	.
Local wages at age 17	0.274	0.615	.

G. Test of heterogeneity in normal selection model[f]

Probability value of test:	0.0884
Outcome of test:	Do not reject

H. Treatment effects[g]

Degree of polynomial	2	3	4	5	Normal
ATE	0.3176	0.5471	0.5434	0.5706	0.3434
	(0.2171)	(0.3750)	(0.3817)	(0.4613)	(0.1677)
TT	1.0193	1.1887	0.7779	0.8119	0.6406
	(0.4847)	(0.5838)	(0.7442)	(0.7412)	(0.1981)
TUT	-0.4992	-0.0778	0.4437	0.4848	0.0238
	(0.7101)	(0.8404)	(1.0635)	(1.2331)	(0.2682)
IV	0.1944	0.1944	0.1944	0.1944	0.1944
	(2.3773)	(2.3773)	(2.3773)	(2.3773)	(2.3773)
p-value of test of equality of treatment effects:	0.3656	0.3489	0.5825	0.7618	0.0087

[a] See text for a description of this test.

[b] The probability values in panel B are from Wald tests for the joint tests. The standard errors are calculated using 100 bootstrap samples.

[c] The IV estimates in panel C are calculated using the method described in the paper; the test of equality is a Wald test using a covariance matrix which is constructed using 1,000 bootstrap samples.

[d] These IV estimates do not include interactions between the treatment and X. The size of the test is controlled using a bootstrap method as described in the text.

[e] The IV estimates underlying these tests are without interactions (between the treatment and X), and the probability values are from Wald tests for the equality of two estimates, using a variance constructed using 1,000 bootstrap samples.

[f] The probability value in panel F is calculated using a Wald test for whether the coefficient on the selection term is zero. The standard error is calculated using 100 bootstrap samples.

[g] The treatment effects in panel G are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). The IV estimate uses P(Z), the propensity score, as the

Table 2: College participation vs. stopping at high school: tests for selection on the gain to treatment, excluding ability measures from the outcome equations but not the choice equations.

A. Conditional moment test[a]

Probability value of test:	0.3222
Outcome of test:	Do not reject

B. Series test[b]

Degree of polynomial		2	3	4	5
Joint test (no bias correction)		0.010	0.001	0.001	0.014
Joint test (with bias correction)		0.000	0.000	0.000	0.000
p-value of test (no bias correction):	0.001				
p-value of test (bias correction):	0.000				
Critical value:	0.020				
Outcome of test:	Reject				

C. IV estimates above and below the median[c]

	Whole sample	Below	Above	Prob. value of test
With interactions	0.7283	3.3203	6.9047	0.9995
(evaluated at mean X)	(2.3773)	(2.7781)	(2.4119)	
Without interactions	0.7817	1.0148	0.9231	0.7988
	(0.0740)	(0.2436)	(0.3099)	
Outcome of test:	Do not reject			

D. IV estimates by quartiles of the propensity score[d]

	1st quartile	2nd quartile	3rd quartile	4th quartile
Estimate:	2.5649	0.6812	0.8527	17.4657
Standard error:	(9.1600)	(2.3186)	(1.0431)	(17.6828)
Smallest probability value from pairwise tests:		0.2028		
Outcome of test:		Do not reject		

E. IV estimates using different instruments[e]

Instrument:	Local college	Local college * AFQT	Local college * mother's education	Local wages at age 17
Estimate:	0.9084	0.9132	0.9121	0.9121
Standard error:	(0.0789)	(0.0786)	(0.0788)	(0.0773)

F. Test of equality of IV estimates using different instruments[f]

Instrument:	Local college	Local college * AFQT	Local college * mother's education
Local college	.	.	.
Local college * AFQT	0.594	.	.
Local college * mother's education	0.328	0.872	.
Local wages at age 17	0.864	0.950	0.998

G. Test of heterogeneity in normal selection model[g]

Probability value of test:	0.7268
Outcome of test:	Do not reject

H. Treatment effects[h]

Degree of polynomial	2	3	4	5	Normal
ATE	0.5378	0.7768	0.7900	0.9892	0.5886
	(0.0554)	(0.1031)	(0.1081)	(0.1647)	0.0531
TT	0.7711	1.0870	1.1762	1.4928	0.5627
	(0.1093)	(0.1576)	(0.2562)	(0.3204)	0.0721
TUT	0.3125	0.5770	0.5128	0.7050	0.6277
	(0.1008)	(0.1470)	(0.2059)	(0.2516)	0.0647
IV	0.7283	0.7283	0.7283	0.7283	0.7283
	(2.3773)	(2.3773)	(2.3773)	(2.3773)	(2.3773)
p-value of test of equality of treatment effects:	0.1151	0.0066	0.0444	0.0487	0.0110

[a] See text for a description of this test.

[b] The probability values in panel B are from Wald tests for the joint tests. The standard errors are calculated using 100 bootstrap samples.

[c] The IV estimates in panel C are calculated using the method described in the paper; the test of equality is a Wald test using a covariance matrix which is constructed using 1,000 bootstrap samples.

[d] These IV estimates do not include interactions between the treatment and X. The size of the test is controlled using a bootstrap method as described in the text.

[e] The IV estimates in panel D contain mother's AFQT and mother's AFQT squared as instruments in addition to those given in the table.

[f] The IV estimates underlying these tests are without interactions (between the treatment and X), and the probability values are from Wald tests for the equality of two estimates, using a variance constructed using 1,000 bootstrap samples.

[g] The probability value in panel F is calculated using a Wald test for whether the coefficient on the selection term is zero. The standard error is calculated using 100 bootstrap samples.

[h] The treatment effects in panel G are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). The IV estimate uses P(Z), the propensity score, as the instrument. In both panels the degree of the polynomial refers to the degree used to approximate E(Y|P) (the degree of the approximation to the MTE is one less).

Table 3: High school diploma vs. high school dropout: tests for selection on the gain to treatment.

A. Conditional moment test[a]	
Probability value of test:	0.0777
Outcome of test:	Do not reject

B. Series test[b]					
Degree of polynomial		2	3	4	5
Joint test (no bias correction)		0.653	0.408	0.598	0.794
Joint test (with bias correction)		1.000	0.494	0.899	1.000
p-value of test (no bias correction):	0.408				
p-value of test (bias correction):	0.494				
Critical value:	0.018				
Outcome of test:	Do not reject				

C. IV estimates above and below the median[c]				
	Whole sample	Below	Above	Prob. value of test
With interactions	0.3473	0.7300	-0.4379	0.2372
(evaluated at mean X)	(1.4028)	(1.5499)	(2.3797)	
Without interactions	0.4326	0.5902	-0.2483	0.5991
	(0.1639)	(0.2403)	(1.5926)	
Outcome of test:	Do not reject			

D. IV estimates by quartiles of the propensity score[d]				
	1st quartile	2nd quartile	3rd quartile	4th quartile
Estimate:	0.2813	0.4687	0.1870	-9.6874
Standard error:	(0.2920)	(8.9879)	(8.3019)	(78.0959)
Smallest probability value from pairwise tests:		0.6508		
Outcome of test:		Do not reject		

E. IV estimates using different instruments					
Instrument:	Father's education	Mother's education	Number of siblings	Family income	Local wages of graduates
Estimate:	0.6043	0.1728	0.7546	0.8821	8.7682
Standard error:	(0.3356)	(0.2247)	(5.9229)	(0.3018)	(110.3856)

F. Hausman-type test of equality of IV estimates using different instruments[e]				
Instrument:	Father's education	Mother's education	Number of siblings	Family income
Father's education	.	.	.	.
Mother's education	0.030	.	.	.
Number of siblings	0.648	0.312	.	.
Family income	0.204	0.006	0.690	.
Local wages of graduates	0.762	0.757	0.770	0.768

G. Test of heterogeneity in normal selection model[f]	
Probability value of test:	0.1373
Outcome of test:	Do not reject

H. Treatment effects[g]					
Degree of polynomial	2	3	4	5	Normal
ATE	0.1085	-0.5366	-0.9323	-2.4109	0.2193
	(0.3365)	(0.5130)	(1.1247)	(2.3749)	(0.1336)
TT	0.0180	-0.7129	-1.2845	-3.1440	0.1069
	(0.4810)	(0.6354)	(1.5713)	(3.0549)	(0.1420)
TUT	0.4517	-0.2280	0.0442	-0.5069	0.7320
	(0.4415)	(0.6838)	(0.9431)	(1.2011)	(0.3107)
IV	0.3473	0.3473	0.3473	0.3473	0.3473
	(1.4028)	(1.4028)	(1.4028)	(1.4028)	(1.4028)
p-value of test of equality of treatment effects:	0.8633	0.2578	0.4468	0.1204	0.1397

[b] The probability values in panel B are from Wald tests for the joint tests. The standard errors are calculated using 100 bootstrap samples.

[c] The IV estimates in panel C are calculated using the method described in the paper; the test of equality is a Wald test using a covariance matrix which is constructed using 1,000 bootstrap samples.

[d] These IV estimates do not include interactions between the treatment and X. The size of the test is controlled using a bootstrap method as described in the text.

[e] The IV estimates underlying these tests are without interactions (between the treatment and X), and the probability values are from Wald tests for the equality of two estimates, using a variance constructed using 1,000 bootstrap samples.

[f] The probability value in panel F is calculated using a Wald test for whether the coefficient on the selection term is zero. The standard error is calculated using 100 bootstrap samples.

[g] The treatment effects in panel G are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). In both panels the degree of the polynomial refers to the degree used to approximate E(Y|P) (the degree of the approximation to the MTE is one less).

sity scores in given intervals. Panel C reports the IV estimates when calculated on samples restricted to only those observations with propensity scores above or below the median, respectively. In addition, we report the p-value of the test of equality of these IV estimates. Panel D of Table 3 reports the IV estimates when separating the sample by the quartiles of the propensity scores. We test for the pairwise equality of these estimates across all pairs and, again, we control the size of the test using the bootstrap method of Romano and Wolf (2005). None of the tests using IV estimates over separate intervals of the propensity score are able to reject the null.

In order to determine whether different instruments are identifying a common treatment effect (as they should under the null hypothesis), panel E of Table 3 presents the IV estimates obtained using different instruments. Panel F reports the probability values of the pairwise tests of equality of those IV estimates. We can see that although some of the tests have small p values we need to recognize that the specification used in obtaining these estimates does not contain interactions between the treatment variable and the X variables and hence is also testing the maintained assumption of the equality of the effects of the X variable by treatment status. We can circumvent this problem by allowing for interactions between the X variables and the treatment, and once we do so we are unable to reject the joint equality of the treatment effects calculated using any two instruments.

Panel G of Table 3 gives the results of the test of whether the coefficient on the selection term in the normal selection model is zero. This is equivalent to a test for the correlated random coefficient if we assume the normal model is the true model and we present it as a benchmark against which to compare the other tests. We are unable to reject the null hypothesis of a correlated random coefficient using this test as well.

Using the estimated MTE, we can calculate the various treatment parameters by weighting the MTE by the weights given in Heckman and Vytlacil (2005) to get the treatment effects listed in panel H of Table 3. The estimated marginal treatment effects (MTE(x, u_D)) at the mean X, for various degrees of polynomials in $P(Z)$ are plotted in Figure A1. In

addition, we give the weights that IV implicitly places on the MTE, and the histogram of estimated propensity scores.

Inspection of the instruments used in this application shows that weak instruments may be a problem in this example. The F-statistic in the first stage of the IV estimate is 7.56 with 27 instruments.[†] The tables in Stock and Yogo (2005) indicate that an F-statistic above 11.36 is required in order for the bias of the two-stage least squares estimate to be at most 10% of the bias of the OLS estimate. Also, an F-statistic of at least 41.17 is required in order for a test on the two-stage least squares estimate with nominal size of 5% to have an actual size of 15%. Therefore, even in the absence of testing for selection on the gain to treatment, we have the problem that the instruments are weak in the traditional sense.

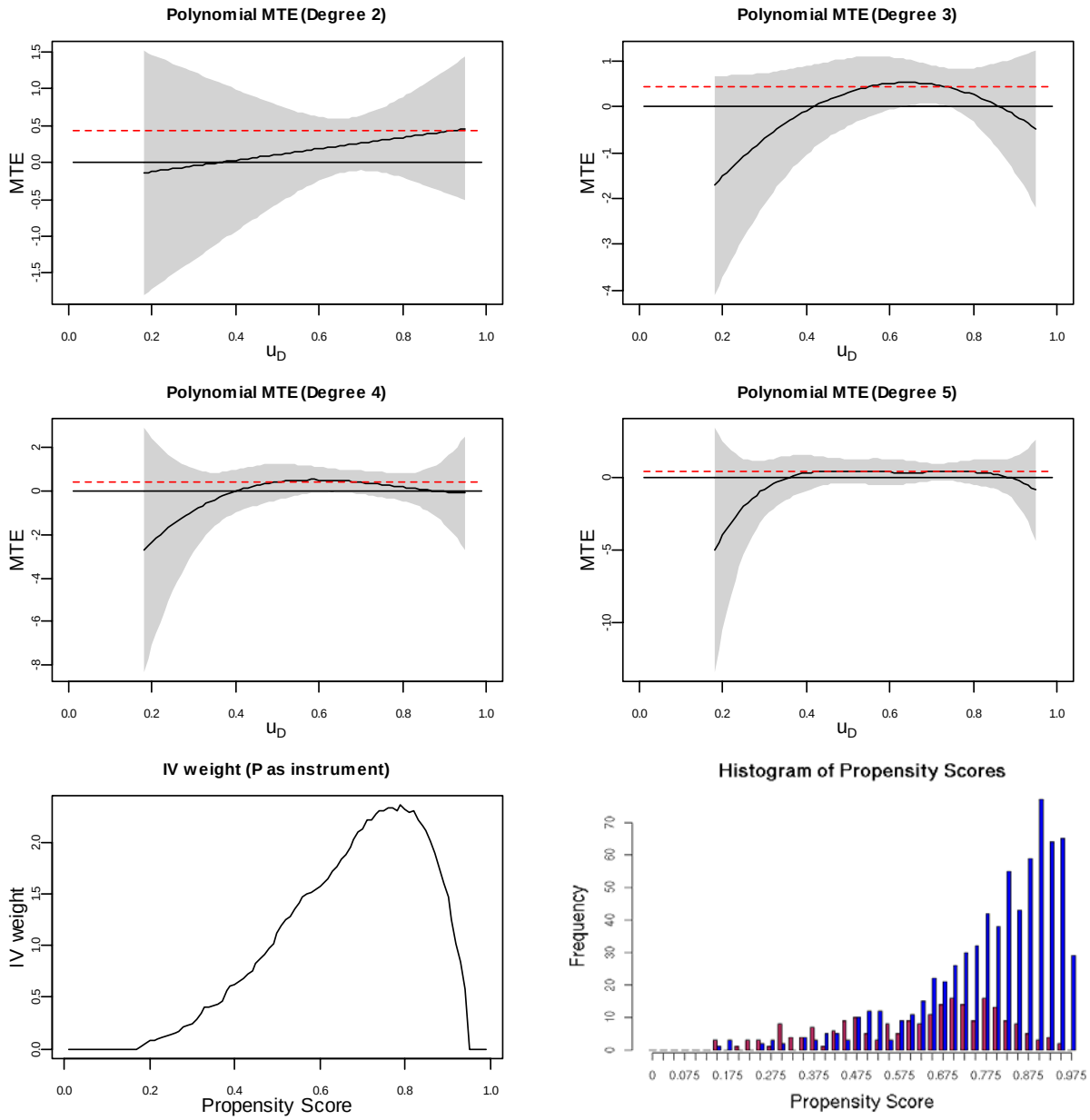
As in the case of measuring the returns to college, we want to check the robustness of our methods for measuring the return to high school graduation in datasets which do not contain ability measures. Many datasets used to estimate the return to high school graduation do not contain the ability measures that the NLSY does (e.g., CPS), so we repeat our analysis excluding those ability measures. The results of the tests for this specification are in Table 4.

The results in Table 4 are similar to those in Table 3 in that all of the tests fail to reject. The test of the equality of IV estimates using different instruments shows a few small p-values, but as discussed above, these pairwise tests are also testing the maintained hypotheses that the effects of the X variables are identical across treatment states. Once we allow for interactions between the X variables and treatment status, none of the pairwise tests are significant.

Finally, as in the case of the returns to college, we conduct the analysis assuming that ability measures do not enter the outcome equations. We still include the ability measures

[†]The instruments are: AFQT score, noncognitive score, mother's education, father's education, number of siblings, family income in 1979, wage of local high school dropouts at age 17, wage of local high school graduates at age 17, unemployment rate of local high school dropouts at age 17, unemployment rate of local high school graduates at age 17, indicator for black, indicator for hispanic, indicator for south residence at age 14, indicator for urban residence at age 14, seven year of birth dummies; as well as five additional variables from the outcome equations: job tenure, job tenure squared, experience, experience squared, and marital status.

Figure A1: High school diploma vs. high school dropout: estimates of marginal treatment effect for different models, IV weights and support of the estimated propensity score.



Note: The covariates in the outcome equations are: job tenure, job tenure squared, experience, experience squared, AFQT score, noncognitive score, marital status, and year of birth indicators. The instruments are: AFQT score, noncognitive score, father's highest grade completed, mother's highest grade completed, number of siblings, family income in 1979, wages and unemployment rates of local dropouts, wages and unemployment rates of local high school graduates, indicators for black and hispanic, indicators for south residence and urban residence at age 14, and year of birth indicators. The dependent variable in the probit is 1 if the individual's highest education is a high school diploma, and 0 if the individual is a high school dropout (GEDs are excluded). The E(Y|P,X) curve is found by regressing log hourly wages on the X's, P, P2, P3, and P4. The confidence intervals are found using 100 bootstraps. In the MTE graph, the horizontal red line indicates the IV estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 1,035.

Table 4: High school diploma vs. high school dropout: tests for selection on the gain to treatment, excluding ability measures.

A. Conditional moment test[a]

Probability value of test:	0.8587
Outcome of test:	Do not reject

B. Series test[b]

Degree of polynomial		2	3	4	5
Joint test (no bias correction)		0.383	0.580	0.769	0.889
Joint test (with bias correction)		0.440	0.929	1.000	1.000
p-value of test (no bias correction):	0.383				
p-value of test (bias correction):	0.440				
Critical value:	0.016				
Outcome of test:	Do not reject				

C. IV estimates above and below the median[c]

	Whole sample	Below	Above	Prob. value of test
With interactions	0.6841	0.9627	0.5759	0.9997
(evaluated at mean X)	(1.4010)	(1.8688)	(2.8445)	
Without interactions	0.5143	0.4554	0.8225	0.5222
	(0.1476)	(0.2112)	(0.5401)	
Outcome of test:	Do not reject			

D. IV estimates by quartiles of the propensity score[d]

	1st quartile	2nd quartile	3rd quartile	4th quartile
Estimate:	0.3754	-3.4379	-0.4340	2.0113
Standard error:	(0.6153)	(11.2694)	(17.4250)	(6.2937)
Smallest probability value from pairwise tests:		0.1256		
Outcome of test:		Do not reject		

E. IV estimates using different instruments

Instrument:	Father's education	Mother's education	Number of siblings	Family income	Local wages of graduates
Estimate:	0.7170	0.4088	0.7723	0.8526	4.1776
Standard error:	(0.2189)	(0.1614)	(0.5396)	(0.2016)	(484.4675)

F. Hausman-type test of equality of IV estimates using different instruments[e]

Instrument:	Father's education	Mother's education	Number of siblings	Family income
Father's education	.	.	.	.
Mother's education	0.034	.	.	.
Number of siblings	0.726	0.169	.	.
Family income	0.486	0.016	0.655	.
Local wages of graduates	0.356	0.344	0.351	0.361

G. Test of heterogeneity in normal selection model[f]

Probability value of test:	0.2411
Outcome of test:	Do not reject

H. Treatment effects[g]

Degree of polynomial	2	3	4	5	Normal
ATE	0.2739	0.7454	1.4305	6.3423	0.3949
	(0.2705)	(0.7193)	(1.5099)	(4.7602)	(0.1252)
TT	0.1406	0.6944	1.6131	7.7247	0.3235
	(0.3673)	(0.8580)	(1.9749)	(6.0451)	(0.1260)
TUT	0.7685	0.9587	0.7946	1.4500	0.6626
	(0.2675)	(0.4015)	(0.5503)	(0.6556)	(0.2066)
IV	0.6841	0.6841	0.6841	0.6841	0.6841
	(1.4010)	(1.4010)	(1.4010)	(1.4010)	(1.4010)
p-value of test of equality of treatment effects:	0.5075	0.8718	0.9198	0.5588	0.1763

[a] See text for a description of this test.

[b] The probability values in panel B are from Wald tests for the joint tests. The standard errors are calculated using 100 bootstrap samples.

[c] The IV estimates in panel C are calculated using the method described in the paper; the test of equality is a Wald test using a covariance matrix which is constructed using 1,000 bootstrap samples.

[d] These IV estimates do not include interactions between the treatment and X. The size of the test is controlled using a bootstrap method as described in the text.

[e] The IV estimates underlying these tests are without interactions (between the treatment and X), and the probability values are from Wald tests for the equality of two estimates, using a variance constructed using 1,000 bootstrap samples.

[f] The probability value in panel F is calculated using a Wald test for whether the coefficient on the selection term is zero. The standard error is calculated using 100 bootstrap samples.

[g] The treatment effects in panel G are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y | P) (and hence the polynomial used to approximate the MTE). In both panels the degree of the polynomial refers to the degree used to approximate E(Y | P) (the degree of the approximation to the MTE is one less).

in the choice equations, however. The results of the tests for this specification are in Table 5.

The results in Table 5 show that the majority of the tests do not reject the null of no selection on the gain to high school graduation. The null is only rejected in the model which assumes normality and in that case both the test of the significance of the coefficient on the control function and the test of the equality of the treatment effects reject the null. Because none of the other less parametric tests reject, we believe that this is not strong evidence against the null.

2 The Impact of the Chilean Voucher Program on Test Scores for Students at Different Margins of Choice

2.1 Introduction

A topic of much recent public policy interest is the effect of school vouchers on school quality. Proponents of vouchers argue that public schools, with effectively a local monopoly, are inefficient providers of education. Following this argument, the government may be able to improve the quality of the education provided to students by giving students vouchers to attend private schools which would compete to offer higher quality education. Unfortunately, measuring the potential effect of such a program has been difficult simply because there have not been many large-scale implementations of voucher programs. The case of Chile, however, is an exception in this regard.

Chile implemented reforms in 1981 which changed the structure of schooling in that country. With the 1981 reforms, the government began providing certain privately-run schools with a fixed per-pupil payment. In order to have existing public schools compete with these private schools, the government changed the funding of public schools to a per-pupil payment which was exactly the same payment given to subsidized private schools. In all, the 1981 reforms created three main types of schools in Chile:

Table 5: High school diploma vs. high school dropout: tests for selection on the gain to treatment, excluding ability measures only from the outcome equations.

A. Conditional moment test[a]

Probability value of test:	0.4653
Outcome of test:	Do not reject

B. Series test[b]

Degree of polynomial		2	3	4	5
Joint test (no bias correction)		0.539	0.252	0.401	0.638
Joint test (with bias correction)		1.000	0.194	0.426	0.999
p-value of test (no bias correction):	0.252				
p-value of test (bias correction):	0.194				
Critical value:	0.019				
Outcome of test:	Do not reject				

C. IV estimates above and below the median[c]

	Whole sample	Below	Above	Prob. value of test
With interactions	1.0219	0.5173	-2.7802	0.2254
(evaluated at mean X)	(1.9368)	(2.0560)	(1.9694)	
Without interactions	0.7086	1.2179	0.0461	0.1964
	(0.1180)	(0.3967)	(0.8430)	
Outcome of test:	Do not reject			

D. IV estimates by quartiles of the propensity score[d]

	1st quartile	2nd quartile	3rd quartile	4th quartile
Estimate:	0.8012	0.6311	0.1303	4.0430
Standard error:	(9.6753)	(3.0701)	(12.3667)	(6.3881)
Smallest probability value from pairwise tests:		0.1676		
Outcome of test:		Do not reject		

E. IV estimates using different instruments

Instrument:	Father's education	Mother's education	Number of siblings	Family income	Local wages of graduates
Estimate:	0.7734	0.6688	0.7928	0.8128	0.8567
Standard error:	(0.1232)	(0.1128)	(0.1262)	(0.1125)	(0.1481)

F. Hausman-type test of equality of IV estimates using different instruments[e]

Instrument:	Father's education	Mother's education	Number of siblings	Family income
Father's education	.	.	.	.
Mother's education	0.077	.	.	.
Number of siblings	0.720	0.143	.	.
Family income	0.509	0.053	0.694	.
Local wages of graduates	0.309	0.072	0.296	0.569

G. Test of heterogeneity in normal selection model[f]

Probability value of test:	0.0032
Outcome of test:	Reject

H. Treatment effects[g]

Degree of polynomial	2	3	4	5	Normal
ATE	0.7183	-0.1488	-0.3042	-1.7060	0.4504
	(0.1805)	(0.4764)	(1.0877)	(2.1889)	(0.0922)
TT	0.7374	-0.2477	-0.4642	-2.1672	0.3253
	(0.2431)	(0.5503)	(1.4596)	(2.7036)	(0.0940)
TUT	0.6508	0.1734	0.2415	-0.0997	0.9187
	(0.2333)	(0.3703)	(0.5481)	(0.8052)	(0.1771)
IV	1.0219	1.0219	1.0219	1.0219	1.0219
	(1.9368)	(1.9368)	(1.9368)	(1.9368)	(1.9368)
p-value of test of equality of treatment effects:	0.9738	0.5983	0.8995	0.6942	0.0014

[a] See text for a description of this test.

[b] The probability values in panel B are from Wald tests for the joint tests. The standard errors are calculated using 100 bootstrap samples.

[c] The IV estimates in panel C are calculated using the method described in the paper; the test of equality is a Wald test using a covariance matrix which is constructed using 1,000 bootstrap samples.

[d] These IV estimates do not include interactions between the treatment and X. The size of the test is controlled using a bootstrap method as described in the text.

[e] The IV estimates underlying these tests are without interactions (between the treatment and X), and the probability values are from Wald tests for the equality of two estimates, using a variance constructed using 1,000 bootstrap samples.

[f] The probability value in panel F is calculated using a Wald test for whether the coefficient on the selection term is zero. The standard error is calculated using 100 bootstrap samples.

[g] The treatment effects in panel G are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on the degree of the polynomial used to approximate E(Y|P) (and hence the polynomial used to approximate the MTE). In both panels the degree of the polynomial refers to the degree used to approximate E(Y|P) (the degree of the approximation to the MTE is one less).

1. Public schools: These schools are administered by local municipalities, of which there are about 300 in Chile. What had previously been centrally administered public schools became "municipal" schools after the 1981 reforms. They receive funding in the form of a fixed per-pupil payment from the government. These schools do not charge tuition and cannot deny admission to students.

2. Voucher (subsidized private) schools: These privately-run schools receive a per-pupil payment from the government which is identical to the payment given to municipal schools. Such schools can choose which students they would like to admit. Since 1996 these schools have been allowed to charge tuition to students.

3. Private (unsubsidized) schools: These independent, privately-run schools receive no funding from the government and charge tuition to students. They generally are comprised of students from privileged backgrounds and the tuitions in these schools is much higher than the tuition at (subsidized) voucher schools.

The introduction of privately-run voucher schools gradually increased the private enrollment rate in Chile throughout the 1980s and 1990s to over 40% by 1996 (Hsieh and Urquiola (2006)). Voucher school enrollment grew particularly fast in more populous, urban areas because these are the areas where new voucher schools tended to be founded. If voucher schools have some fixed costs of administration and maintenance then we would expect such schools to be founded in areas where these schools could maintain enrollments at an optimal level. See Hsieh and Urquiola (2006) and McEwan (2001) for further description and analysis of the Chilean educational system.

The issue which we address in this paper is: what is the effect of voucher school attendance on students at different margins of indifference to attending a voucher school? That is, which students potentially benefit and which students are potentially hurt by attending a voucher school? The variable which we will consider as the outcome is a student's score on an achievement test. Section 2.2 describes the model for this outcome and will make clear

what we mean by "students at different margins of indifference." Section 2.3 presents our empirical results.

2.2 Model and Background

The data which we will use to measure the quality of schools is a standardized achievement test score. Because we are interested in the effect of voucher schools on the students who attend the schools, we abstract from the indirect process through which this occurs. That is, we will not attempt to separate the effect of voucher schools on the quality of schooling and the effect of higher quality schooling on student test scores. Instead, we model the outcome (test score) as a direct function of whether or not a student attends a voucher school. For concreteness, let Y_i denote the test score of student i. Let D_i denote the choice of the type of school which student i attends, ie. $D_i \in \{\text{Public}, \text{Voucher}, \text{Private}\}$. Let X_i denote the observable characteristics of student i. We adopt a potential outcomes framework and write the potential test scores for individual i as

$$Y_{i,\text{Public}} = \mu_{\text{Public}}(X_i) + U_{i,\text{Public}}$$
$$Y_{i,\text{Voucher}} = \mu_{\text{Voucher}}(X_i) + U_{i,\text{Voucher}}$$
$$Y_{i,\text{Private}} = \mu_{\text{Private}}(X_i) + U_{i,\text{Private}}$$

We do not observe all of these potential outcomes for any student i, however, but rather we only observe one of the three. That is we observe

$$Y_i = \mathbf{1}\{D_i = \text{Public}\}Y_{i,\text{Public}} + \mathbf{1}\{D_i = \text{Voucher}\}Y_{i,\text{Voucher}} + \mathbf{1}\{D_i = \text{Private}\}Y_{i,\text{Private}}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise.

Because the unsubsidized private schools serve only a very small, elite segment of the

Table 6: Summary statistics, student characteristics

	Public	Voucher	Private
Gender (male = 1)	0.491	0.496	0.479
	(0.500)	(0.500)	(0.500)
Mother's highest grade completed	9.842	12.060	15.892
	(3.322)	(3.082)	(1.626)
Father's highest grade completed	9.912	12.116	16.395
	(3.466)	(3.234)	(1.678)
Number of family members	5.162	4.869	5.112
	(1.766)	(1.585)	(1.623)
Monthly family income:			
Less than 100,000 pesos	0.314	0.109	0.001
	(0.464)	(0.312)	(0.031)
Between 100,001 and 200,000 pesos	0.441	0.326	0.007
	(0.497)	(0.469)	(0.081)
Between 200,001 and 300,000 pesos	0.137	0.204	0.013
	(0.344)	(0.403)	(0.113)
Between 300,001 and 400,000 pesos	0.052	0.118	0.017
	(0.222)	(0.323)	(0.131)
Between 400,001 and 500,000 pesos	0.025	0.077	0.025
	(0.156)	(0.267)	(0.158)
More than 500,000 pesos	0.032	0.165	0.937
	(0.175)	(0.372)	(0.244)
Number of observations	51,198	61,152	10,443

Note: These summary statistics are constructed from a subsample of fourth grade students in the SIMCE 2005 sample. The sample is restricted to include only students in Regions 5, 8 and 13 (Santiago). The numbers in parentheses beneath each mean are the sample standard deviations.

population we will focus only on students for whom these schools are not an option. We do this because the observable characteristics (X) of those attending unsubsidized private schools are so different from those of students attending public and voucher schools that it is difficult to justify pooling those observations. This is shown in Table 6. Notice that the differences in family income between those attending unsubsidized private schools and other students are particularly glaring. The majority of both public and voucher school students have a monthly family income below 300,000 pesos while only 2% of students in unsubsidized private schools have family incomes this low.

In addition, because we are interested in measuring the effect of voucher schools relative

to the status quo of public schools, it seems reasonable to focus on just the students on the margins of choice between those two types of schools. Therefore, in practice we restrict the choice set faced by the students we will be considering to $D_i \in \{0, 1\}$ where $D_i = 0$ corresponds to attending a public school and $D_i = 1$ corresponds to attending a voucher school. Dropping individual subscripts, we can now simplify the potential outcomes to the standard

$$Y_0 = \mu_0(X) + U_0 \tag{1}$$

$$Y_1 = \mu_1(X) + U_1 \tag{2}$$

Finally, we need to specify a model for how students (or more likely their parents) choose which type of school the student attends. Let Z_i denote observable characteristics of student i which affect his or her likelihood of attending a voucher school (such as the distance to the nearest voucher school) and V_i denore unobservable characteristics of the student. Then

$$D_i = \mathbf{1}\{\mu_D(Z_i) - V_i \geq 0\}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function with the same definition as above.

We maintain assumptions about the structure of the unobservables in the outcome and choice equations that are standard to those in the treatment effect literature or equivalent to those used in the treatment effect literature. Our assumption are (A-1) through (A-5) given in Section 2 of the text.

The fundamental parameter which we seek to estimate is the marginal treatment effect (MTE), which is the average treatment effect for individuals at different margins of indifference of attending a voucher school. To better understand this parameter, note that if we let F_V be the CDF of the unobservable in the choice equation, V, then we can transform the

choice equation to

$$\begin{aligned} D_i &= \mathbf{1}\{\mu_D(Z_i) - V_i \geq 0\} \\ &= \mathbf{1}\{F_V(\mu_D(Z_i)) - F_V(V_i) \geq 0\} \end{aligned}$$

Define a new random variable $U_D = F_V(V)$ which is distributed Uniform$[0, 1]$ by construction. Also, let $P(z)$ be the propensity score, or probability of choosing $D = 1$, for an individual with observables $Z_i = z$. Then

$$P(z) = \Pr(D_i = 1 | Z_i = z) = \Pr(V_i \leq \mu_D(z)) = F_V(\mu_D(z))$$

Therefore, we can write

$$D_i = \mathbf{1}\{P(Z_i) \geq U_D\}$$

Now we define the marginal treatment effect as

$$\text{MTE}(x, u_D) = E(Y_1 - Y_0 | X = x, U_D = u_D)$$

This is the mean treatment effect for individuals with observables x and unobservable in the choice equation u_D. To see why this captures the treatment effect for individuals at different margins of indifference notice that for a fixed u_D we know that such an individual would need a Z_i such that $P(Z_i) = u_D$ to be indifferent between attending a voucher school or a public school. That is, hold all else fixed, the MTE evaluated at larger values of u_D will give the treatment effect for people who would need larger values of $P(Z_i)$ in order to be indifferent between attending a voucher or a public school – that is, people who are less likely to attend voucher school.

2.2.1 Estimation

We briefly describe the techniques we use for estimating $\text{MTE}(x, u_D)$. The estimation methods all rely on the relationship

$$\text{MTE}(x, u_D) = \left. \frac{\partial E(Y|X = x, P(Z) = p)}{\partial p} \right|_{p=u_D}$$

To see where this relationship comes from notice that, under our assumptions, we can write

$$\begin{aligned} E(Y|X = x, P(Z) = p) &= E(Y_0|X = x) + E(D(Y_1 - Y_0)|X = x, P(Z) = p) \\ &= E(Y_0|X = x) + E(Y_1 - Y_0|X = x, D = 1)p \\ &= E(Y_0|X = x) + \int_0^p E(Y_1 - Y_0|X = x, U = u)du. \end{aligned}$$

The integrand in the expression in the last line is $\text{MTE}(x, u)$. Therefore, differentiating with respect to p and evaluating the partial derivative at $p = u_D$ gives $\text{MTE}(x, u_D)$.

The different estimation methods we use differ in how they estimate $E(Y|X = x, P(Z) = p)$. We will consider a parametric method, which assumes the joint normality of the unobservables, and two semiparametric methods – one which uses ordinary polynomials in p and one which uses local polynomials in p. We will maintain the assumption that the outcome equations are linear in X, that is

$$Y_1 = X\beta_0 + U_0$$

$$Y_0 = X\beta_1 + U_1$$

With this linearity the expected value of the observed outcome Y is

$$\begin{aligned}
&E(Y|P(Z)=p, X=x) \\
&\quad = E(Y_0|P(Z)=p, X=x) + E(D(Y_1-Y_0)|P(Z)=p, X=x) \\
&\quad = \mu_0(x) + (\mu_1(x)-\mu_0(x))p + E(U_0|P(Z)=p) + E(U_1-U_0|D=1, P(Z)=p)p \\
&\quad = x\beta_0 + x(\beta_1-\beta_0)p + \kappa(p)
\end{aligned}$$

where $\kappa(p) = E(U_0 \mid P(Z)=p) + E(U_1 - U_0 \mid D=1, P(Z)=p)p$. Our first specification assumes that (U_0, U_1, V) are jointly normally distributed and therefore the marginal treatment effect is

$$\text{MTE}(x, u_D) = x(\beta_1-\beta_0) + \left(\frac{\sigma_{1V}-\sigma_{0V}}{\sigma_V}\right)\Phi^{-1}(u_D)$$

The other specification estimates $\text{MTE}(x, u_D)$ by taking the derivative

$$\frac{\partial E(Y|X=x, P(Z)=p)}{\partial p} = x(\beta_1-\beta_0) + \frac{\partial \kappa(p)}{\partial p}$$

and evaluating this at $p = u_D$. We use ordinary polynomials in p which set

$$\kappa(p) = \sum_{j=0}^{J} \phi_j p^j$$

With a high enough degree of the polynomial, J, this specification will be able to approximate any $\kappa(\cdot)$ function which satisfies standard regularity conditions.

2.2.2 Data

Our data for measuring the effect of voucher schools on students' test scores comes from the Sistema de Medición de la Calidad de la Educación (SIMCE), which is a national standardized test administered once a year to either 4th, 8th or 10th grade students. We use the data from the 2005 administration of the test, which was given to 96% of all of the 4th grade

students in Chile. In addition we use data from the 1998 and 2000 administrations of the Caracterización Socioeconómica (CASEN) survey.

Chile is divided geographically into thirteen regions. We focus in this study on three of those regions: Región Metropolitana, Región de Biobío, and Región de Valparaíso. These are the three largest regions by population and contain the three largest cities in Chile – Santiago, Concepción and Valparaíso. We consider these regions because some of the instruments we use are constructed from geographical variables in the CASEN survey which are only representative in certain regions of Chile.

We consider two different subsets of our data to use in the analysis. As discussed above, we seek to test the linearity of $E(Y|X = x, P(Z) = p)$ in p, holding x constant. Therefore, we will consider one subset of the data in which we condition on the categorical X variables and we impose linearity in the remaining X variables. This reduces the size of our sample significantly, however, and so we will also conduct the analysis using the entire sample, where we include all of the X conditioning variables linearly.

The regions we are considering are divided into 15 "provinces" and 144 "municipalities" in 2005. We have data on 105,124 students from these regions. In the subset in which we condition on a number of categorical X variables, we will look at only male students living in Santiago whose parents earn between 100,000 and 200,000 pesos per month. We have data on 12,789 such students from the 45 municipalities in Santiago.

Most of the instruments for voucher school attendance which we use are at the municipality level. Because voucher schools tend to be founded in more populous, faster growing and more urban municipalities we use these characteristics in constructing our instruments for voucher school attendance. Students who were in 4th grade in 2005 entered school in 2001. Therefore, we use data from the 2000 administration of the CASEN survey to construct instruments which are plausibly exogenous to the outcome we are measuring, but which predict the voucher school attendance of students in each municipality. In addition, we include as instruments proxies for the cost of voucher school attendance and for the relative desir-

ability of voucher schools. To proxy the cost of voucher school attendance we form a variable which is the difference between the average tuition of voucher schools in a municipality and the average tuition of public schools in that municipality.[3] As a measure of desirability we form variables which are the difference between the average test scores of the voucher school students in a municipality and the average test scores of the public school students in that municipality. Note that these instruments are similar to those used in Hsieh and Urquiola (2006).

The binary choice in this setting is whether a student attends a voucher school ($D = 1$) or a public school ($D = 0$). In our first stage we run a probit of voucher school attendance on the following independent variables (Z): population of municipality in 2000, urbanization of municipality in 2000, population growth rate of municipality between 1998 and 2000, difference in average tuition between voucher and public schools in the municipality in 2000, difference in average test scores between the voucher schools and the public schools in one's municipality in 2002[4], gender, mother's highest grade completed, father's highest grade completed, number of family members, indicators for household income categories and region indicators. We use the fitted values from this probit as our estimates of the propensity score $P(Z)$. Note that in the sample where we have conditioned on gender, region and household income category, we do not include those in the estimation of the propensity score.

2.3 Results

The examination used to measure scholastic performance has three components — math, verbal, and social and natural sciences. We consider each of these test scores as an outcome variable. Once we have our estimated propensity scores, we regress the outcome on the

[3]As noted above, public schools are not allowed to charge tuition. However, because respondents in the CASEN sometimes report positive tuition payments we surmise that these respondents are reporting other educational expenditures. Because presumably families with voucher school students report these additional expenditures as well, any non-tuition costs should be netted out by our differencing. The results are not sensitive to this construction and hold up if we impose zero tuition at public schools.

[4]We use test scores from 2002 because this was the year closest to 2001 in which the SIMCE was administered to 4th grade students.

Table 7: Tests for selection on the gain to treatment – males in Santiago

	Conditional moment test	Test of difference between linear and series estimator	Test of equality of IV estimates	
			With interactions	Without interactions
p-value	0.0001	0.4902	0.1730	0.0211
Critical value for p-value for rejecting H_0	0.05	0.0136	0.05	0.05
Result of test	**Reject**	Do not reject	Do not reject	**Reject**

Note: See text for a description of how test statistics were calculated and critical value of series test was obtained.

following controls (X) in addition to polynomial terms in the propensity score $P(Z)$: all of the Z variables excluding the municipality-level variables – population, population growth rate, urbanization, average tuition difference, and average test score difference.

We now carry out our tests for correlated random coefficients as described in Section 2.2 on our two subsets of the data. Table 7 presents the results of our tests in the subset of the data which includes only male students living in Santiago whose parents earn between 100,000 and 200,000 pesos per month. The first column reports that the conditional moment test of Bierens (1990) rejects the null hypothesis of the linearity of $E(Y|X = x, P(Z) = p)$ and hence it rejects the null hypothesis of no selection on the gain to treatment. The second column reports the result of the test in which we compare a parametric (linear) estimate of $E(Y|X = x, P(Z) = p)$ to a flexible series estimator of that function. This test is unable to reject the null hypothesis. Finally, the last two columns show the outcome of the test of the equality of the IV estimates using two separate samples – those with propensity scores above the median, and those with propensity scores below the median. We see that if we include interactions between all of the X variables and the treatment indicator we do not reject the null, but if we do not include those interactions then we do reject. This could be interpreted either as evidence that the interactions should be included or as evidence that including the interactions introduces so much error into the estimates that we lose the power to reject the null hypothesis.

Next we report the results of our tests in the whole sample. Table 8 shows that in this

Table 8: Tests for selection on the gain to treatment – all students in large regions

	Conditional moment test	Test of difference between linear and series estimator	Test of equality of IV estimates	
			With interactions	Without interactions
p-value	0.0002	0.0000	0.0000	0.0035
Critical value for p-value for rejecting H_0	0.05	0.0349	0.05	0.05
Result of test	**Reject**	**Reject**	**Reject**	**Reject**

Note: See text for a description of how test statistics were calculated and critical value of series test was obtained.

sample, all of the tests are able to reject the null hypothesis of no selection on the gain to treatment.

In order to see the effect that accounting for correlated random coefficients would have on our interpretation of the effect of voucher schools on students' test scores in Chile we can estimate the commonly used treatment effects, ATE and TT, and see how they compare to the results found using OLS or IV. Notice that our test in which we estimate $E(Y|X = x, P(Z) = p)$ using a series estimator immediately gives us estimates of the MTE (which is simply the derivative of $E(Y|X = x, P(Z) = p)$ with respect to p). Therefore we calculate the treatment effects as weighted averages of our estimates of the MTE, by using the weights given in Heckman and Vytlacil (2005). However, using these weights to form the treatment effect makes clear the problem of the support of the propensity score. That is, if we do not have full support (on $[0, 1]$) of the propensity score, then we cannot form the weights and we are unable to calculate the treatment effects. This problem presents itself in the subset of the data in which we condition on the X variables. In this subset we do not have full support and hence are unable to calculate the treatment effects. However, in the larger dataset, because we are using the data across many values of the X variables, we do have full support and hence can calculate the treatment effects. We give the estimates of the treatment effects for this data in table 9. The numbers presented in this table are the treatment effects averaged over the covariates X. Note that because our estimate of the MTE depends on the number of polynomial terms used to estimate $E(Y|X = x, P(Z) = p)$, our estimates of the treatment

Table 9: Treatment effect estimates – all students in large regions

	Degree of Series Estimate			
	2	3	4	5
Average treatment effect	-0.559	-15.958	-14.333	-14.665
Average effect of the treatment on the treated	-21.877	-34.205	-12.547	-12.951
Average effect of the treatment on the untreated	27.831	-1.584	-29.056	-29.585
Instrumental variables estimate	2.429	2.429	2.429	2.429
Instrumental variables estimate (using weights)	-3.907	-1.908	-1.735	-1.654

effects will also differ based on the degree of this polynomial. In the table we also show the estimate obtained using standard IV, namely the ratio of the covariances, described above with $P(Z)$ as our instrument, as well as the IV estimate found using the implicit weight that the IV estimator is placing on the MTE. These two estimates differ not only because our estimate of the MTE is not exact, but also because the weights themselves are estimated.

As table 9 shows, the treatment effect estimates tend to differ substantially from the estimate obtained using instrumental variables. This indicates that selection on the gain to treatment is likely present in this data and is important in the interpretation of the effects of the voucher school program. Using only the IV estimate the researcher would conclude that attending a voucher school has no effect on student achievement (as the standard deviation of the test scores is normalized to 50, against which the point estimate of 2.429 is negligible). However, looking at the more relevant ATE and TT we see that the effect is likely closer to -15 to -10 for these populations, which corresponds to around one third of a standard deviation decrease in test scores.

In figure A2 we graph the estimates of the $MTE(x, u_D)$ averaged over the covariates X for the subset of the data which includes just males in Santiago in a certain income range. This figure shows the estimates of the MTE obtained from the 4th and 5th degree series estimators. The panels in the second row plot the same MTEs as the panels in the first row, but simply fix the scale of the vertical axis so that they can be more easily compared. Also, we plot the $MTE(\bar{x}, u_D)$ only for $u_D \in [0.04, 0.82]$ because this is the range of the

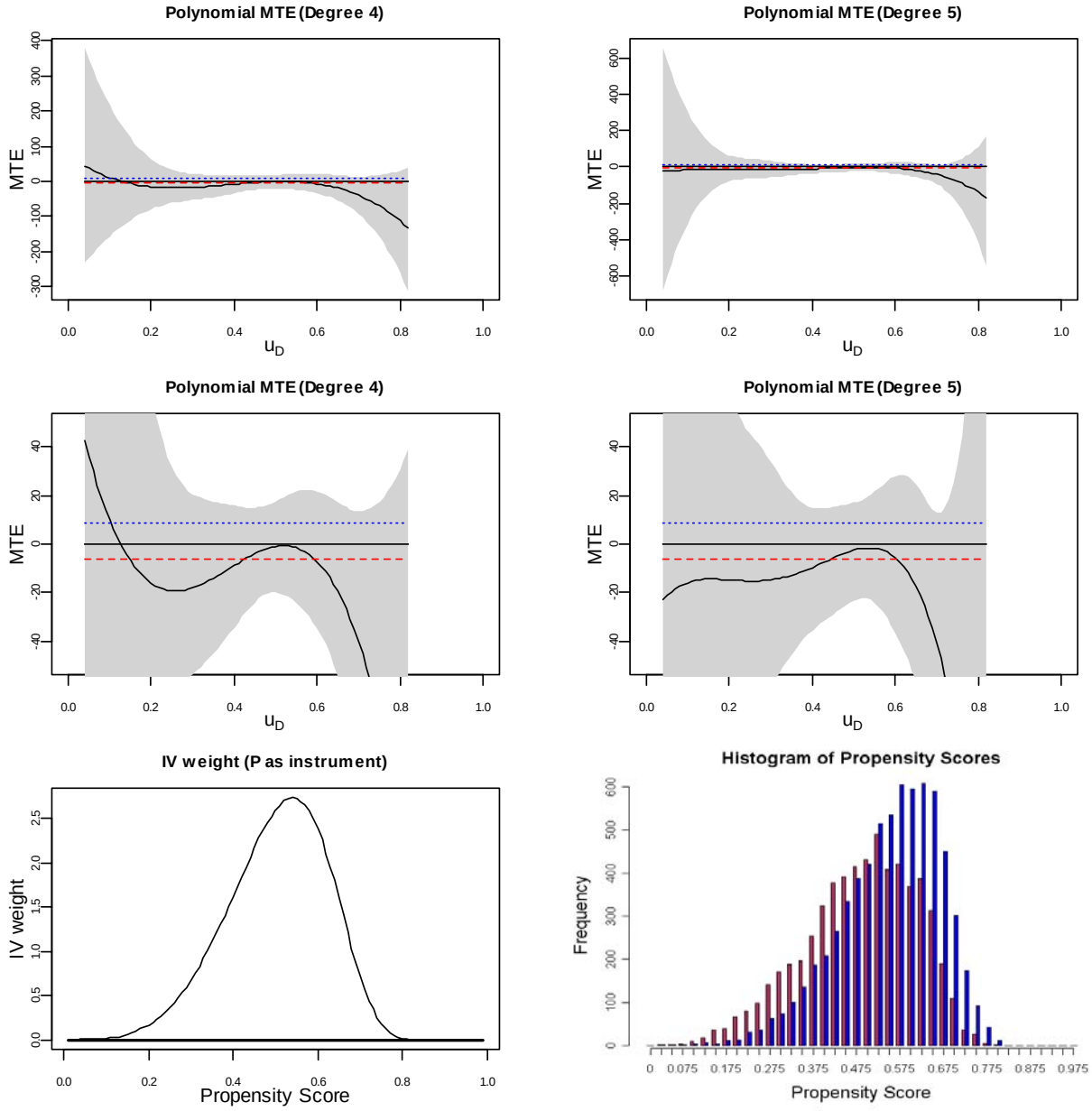
support of the propensity score. Hence, it is only over this range where we can actually estimate $MTE(x, u_D)$. We can see that the 95% pointwise confidence intervals are fairly large and this shows why we were unable to reject the null hypothesis of no selection on the gain to treatment using these estimates. Note, however, that both point estimates seem to indicate that those with high values of U_D (the unobservable in the choice equation), whose unobservables make them least likely to attend a voucher school, are indeed those who would get the lowest benefit from attending such a school, in accordance with our intuition. The wide standard error bands, however, preclude us from being able to make this statement one of statistical significance – at least in this subset of the data.

In the bottom two panels we plot the weights that the IV estimate is implicitly placing on the MTE (averaged over X) as well as a histogram of the estimated propensity scores, separated by treatment status. We can see that the support of the estimated propensity scores is limited and hence the IV estimate places weight only on the center of the MTE.

Figure A3 plots the estimates of $MTE(x, u_D)$ as well as the IV weights and a histogram of the propensity scores for the entire sample. We can see that the standard error bands on the MTE estimates are much narrower than those in figure A2 because we are able to use a much larger sample. Therefore we are able to say that those individuals whose unobservables make them unlikely to attend voucher schools (those with high values of U_D) get significantly negative returns from attending such schools. Also, there is a range of individuals with values of U_D between 0.5 and 0.7 for whom the voucher schools have a significantly positive effect.

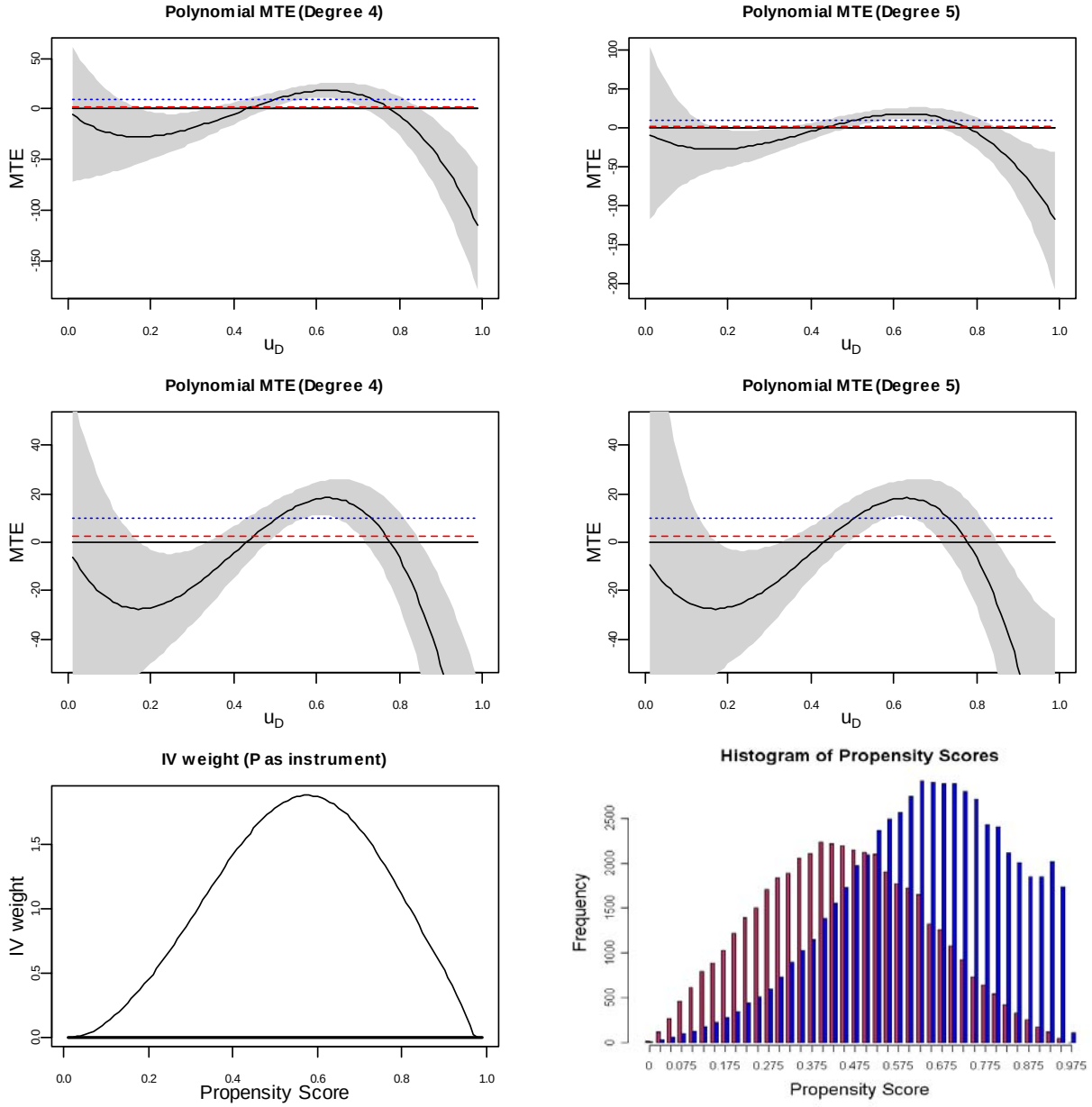
The bottom two panels of figure A3 plot the weights that the IV estimator is implicitly placing on the MTE (averaged over X) and a histogram of the estimated propensity scores (separated by treatment status). We can see from the histogram that we have near full support on the unit interval of estimated propensity scores in both treatment groups, which is what we need in order to be able to form the treatment effects commonly sought in the literature. The IV estimate places the most weight in the center of the interval, however, because that is where the mean of the propensity score lies.

Figure A2: MTE estimates, IV weights and Propensity Scores – males in Santiago



[a] The covariates in the outcome equations are: mother's highest grade completed, father's highest grade completed, and number of family members. The instruments are: population and urbanization of one's municipality in 2000, population growth rate between 1998 and 2000, difference between average tuition in voucher schools and average tuition in public schools in one's municipality, difference in average test scores in voucher schools and average test scores in public schools in one's municipality, in addition to all of the X variables. The dependent variable in the probit is 1 if the individual is enrolled in a voucher school, and 0 if the individual is enrolled in a public school. The confidence intervals are found using 100 bootstraps. In the MTE graph, the dashed red line indicates the IV estimate and the dotted blue line indicates the OLS estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 12,789.

Figure A3: MTE estimates, IV weights and Propensity Scores – all students in large regions



[a] The covariates in the outcome equations are: gender, mother's highest grade completed, father's highest grade completed, number of family members, and household income categories. The instruments are: population and urbanization of one's municipality in 2000, population growth rate between 1998 and 2000, difference between average tuition in voucher schools and average tuition in public schools in one's municipality, difference in average test scores in voucher schools and average test scores in public schools in one's municipality, in addition to all of the X variables. The dependent variable in the probit is 1 if the individual is enrolled in a voucher school, and 0 if the individual is enrolled in a public school. The confidence intervals are found using 100 bootstraps. In the MTE graph, the dashed red line indicates the IV estimate and the dotted blue line indicates the OLS estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 105,124.

2.4 The Impact of Vouchers on Test Scores

Our last empirical example examines the effect of school vouchers on test scores. We study this by using a large sample of Chilean fourth graders collected in 2005. The Chilean voucher system was implemented in the early 1980's and has received great attention in the literature (see McEwan and Carnoy, 2000, as well as the Web Appendix).

We estimate the propensity score using a probit model. We let $D = 1$ if the student attends a voucher school, and $D = 0$ if the student attends a local public school. The set of variables included in the probits includes population and urbanization in 2000, local population growth rate between 1993 and 2000, school tuition, and the difference in average test scores in voucher and public schools at the municipality level.

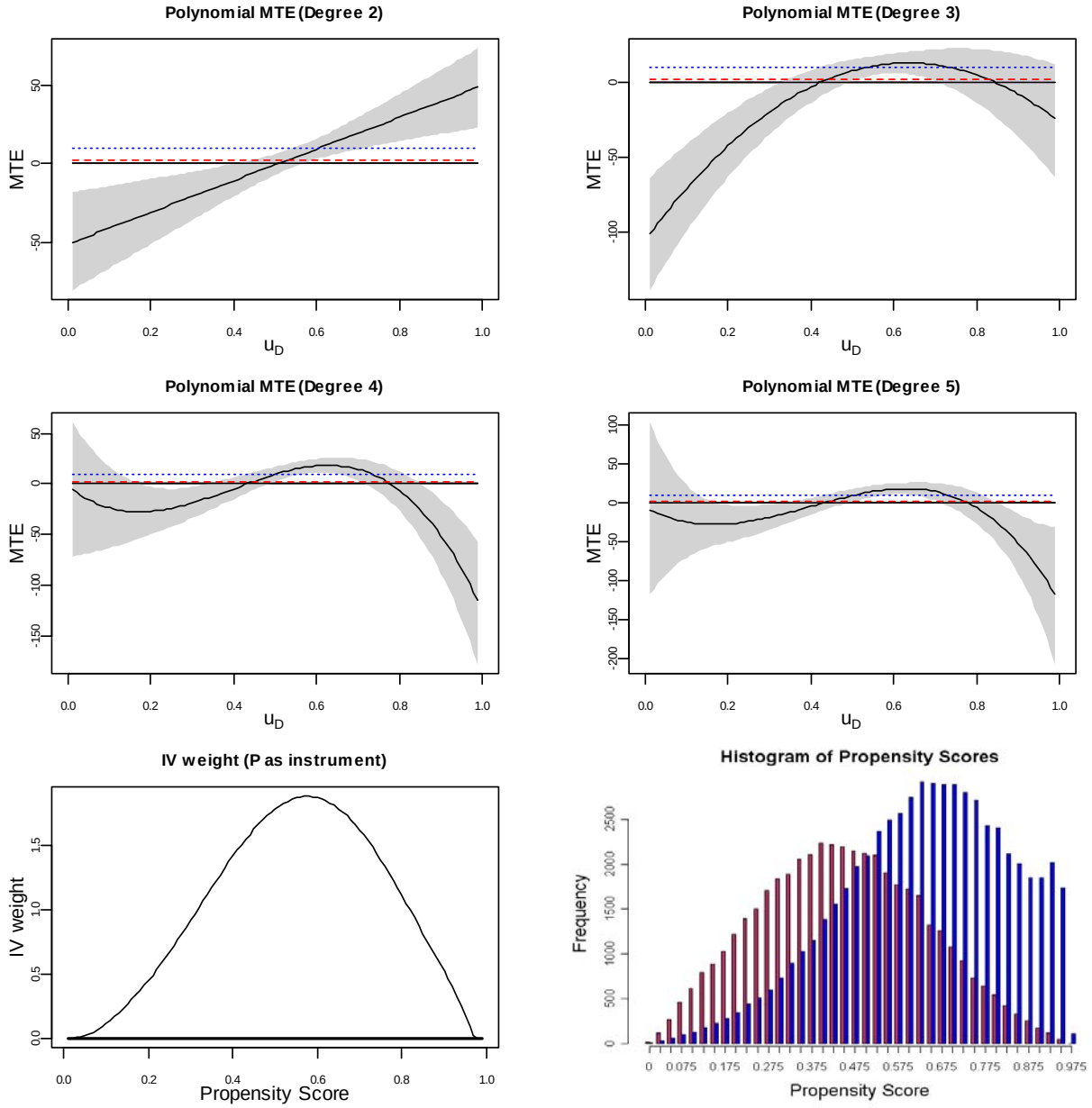
Using the fitted values from the probit we form our estimates of the propensity score. We then regress individual's math test score on a polynomial in the propensity score plus gender, mother's and father's highest grade completed, number of family members, and household income. See the data appendix for further details about our sample and variables.

Table 10 presents the results for our tests. We find strong evidence against the null hypothesis. All our tests reject linearity of MTE. Figure A4 plots the estimated MTE for different degrees of polynomials in $P(Z)$.

We investigate the effect of randomly reducing sample size on the performance of the tests and the variability of the estimated MTE. The full dataset contains over 100,000 observations, from which we randomly sample 1,000, 5,000, 10,000 or 20,000 observations. We estimate the MTE in each of these samples and conduct each of our tests for heterogeneity in each of these samples. The results of the tests on the reduced sample sizes are summarized in Table 11. It is only once we reach 20,000 observations that we are able to reject. At that sample size, we reject the null of no selection on the gain with all of our tests.[5] Note that in this exercise, we sample the full data randomly to construct the reduced samples and conduct

[5]As discussed in Lindley (1957) and Leamer (1978), fixing the size of a test when increasing the sample size has little justification. A better procedure is to trade off power and size as sample size increases instead of loading all gains due to sample size into power.

Figure A4: Chile voucher schools: estimates of marginal treatment effect for different models, IV weights and support of the estimated propensity score.



Note: The covariates in the outcome equations are: gender, mother's highest grade completed, father's highest grade completed, number of family members, and household income categories. The instruments are: population and urbanization of one's municipality in 2000, population growth rate between 1998 and 2000, difference between average tuition in voucher schools and average tuition in public schools in one's municipality, difference in average test scores in voucher schools and average test scores in public schools in one's municipality, in addition to all of the X variables. The dependent variable in the probit is 1 if the individual is enrolled in a voucher school, and 0 if the individual is enrolled in a public school. The confidence intervals are found using 100 bootstraps. In the MTE graph, the dashed red line indicates the IV estimate and the dotted blue line indicates the OLS estimate. In the histogram, the blue bars correspond to the D=1 group and the red bars to the D=0 group. The sample size is 105,124.
Source: Heckman, Schmierer and Urzua (2007).

Table 10: Chile voucher schools: tests for selection on the gain to treatment.

A. Conditional moment test[a]	
Probability value of test:	0.0002
Outcome of test:	Reject

B. Series test[b]					
Degree of polynomial		2	3	4	5
Joint test (no bias correction)		0.0001	0.0000	0.0000	0.0000
Joint test (with bias correction)		0.0004	0.0000	0.0000	0.0000
p-value of test (no bias correction):	0.0000				
p-value of test (bias correction):	0.0000				
Critical value:	0.0349				
Outcome of test:	Reject				

C. IV estimates above and below the median[c]				
	Whole sample	Below	Above	Prob. value of test
With interactions	2.4285	2.8725	16.0318	0.0000
(evaluated at mean X)	(2.5729)	(5.2842)	(5.5249)	
Without interactions	-1.0247	-14.0080	5.5534	0.0079
	(2.5088)	(4.6670)	(4.5419)	
Outcome of test:	Reject			

D. IV estimates by quartiles of the propensity score[d]				
	1st quartile	2nd quartile	3rd quartile	4th quartile
Estimate:	-19.0022	-9.6446	20.0183	2.7580
Standard error:	(10.2035)	(9.0290)	(8.4232)	(7.4222)
Smallest p-value from pairwise tests:		0.0027		
Critical value:		0.0082		
Outcome of test:		Reject		

E. IV estimates using different instruments			
Instrument:	Population	Urbanization	Pop. growth rate
Estimate:	0.2412	-0.1799	-18.9217
Standard error:	(5.5975)	(4.4280)	(6.7253)

F. Hausman-type test of equality of IV estimates using different instruments[e]			
Instrument:	Population	Urbanization	Pop. growth rate
Population	.	.	.
Urbanization	0.935	.	.
Population growth rate	0.000	0.015	.

G. Test of heterogeneity in normal selection model[f]	
Probability value of test:	0.0000
Outcome of test:	Reject

H. Treatment effects[g]					
Degree of polynomial	2	3	4	5	Normal
ATE	-0.559	-15.958	-14.333	-14.665	-1.295
	(3.3576)	(4.6168)	(4.6213)	(6.4529)	(2.8102)
TT	-21.877	-34.205	-12.547	-12.951	-18.245
	(15.6302)	(7.5952)	(9.3843)	(10.5544)	(2.7534)
TUT	27.831	-1.584	-29.056	-29.585	21.634
	(20.8510)	(7.3581)	(9.1279)	(9.9496)	(2.9973)
IV	2.429	2.429	2.429	2.429	2.429
	(2.5729)	(2.5729)	(2.5729)	(2.5729)	(2.5729)
p-value of test of equality of treatment effects:	0.000	0.000	0.000	0.000	0.000

[a] See text for a description of this test.

[b] The probability values in panel B are from Wald tests for the joint tests. The standard errors are calculated using 100 bootstrap samples.

[c] The IV estimates in panel C are calculated using the method described in the paper; the test of equality is a Wald test using a covariance matrix which is constructed using 1,000 bootstrap samples.

[d] These IV estimates do not include interactions between the treatment and X. The size of the test is controlled using a bootstrap method as described in the text.

[e] The IV estimates underlying these tests are without interactions (between the treatment and X), and the probability values are from Wald tests for the equality of two estimates, using a variance constructed using 100 bootstrap samples.

[f] The probability value in panel G is calculated using a Wald test for whether the coefficient on the selection term is zero. The standard error is calculated using 100 bootstrap samples.

[g] The treatment effects in panel H are calculated by weighting the estimated MTE by the weights from Heckman and Vytlacil (2005). Therefore, they vary depending on th degree of the polynomial used to approximate $E(Y|P)$ (and hence the polynomial used to approximate the MTE). The IV estimate uses $P(Z)$, the propensity score, as the instrument. The estimates differ not only because the estimate of the MTE is inexact, but also because the weights are estimated. In both panels the degree of the polynomial refers to the degree used to approximate $E(Y|P)$. Standard errors are in parentheses below the estimates.

Source: Heckman, Schmierer and Urzua (2007).

Table 11: Chile voucher schools: tests for selection on the gain to treatment in smaller samples.

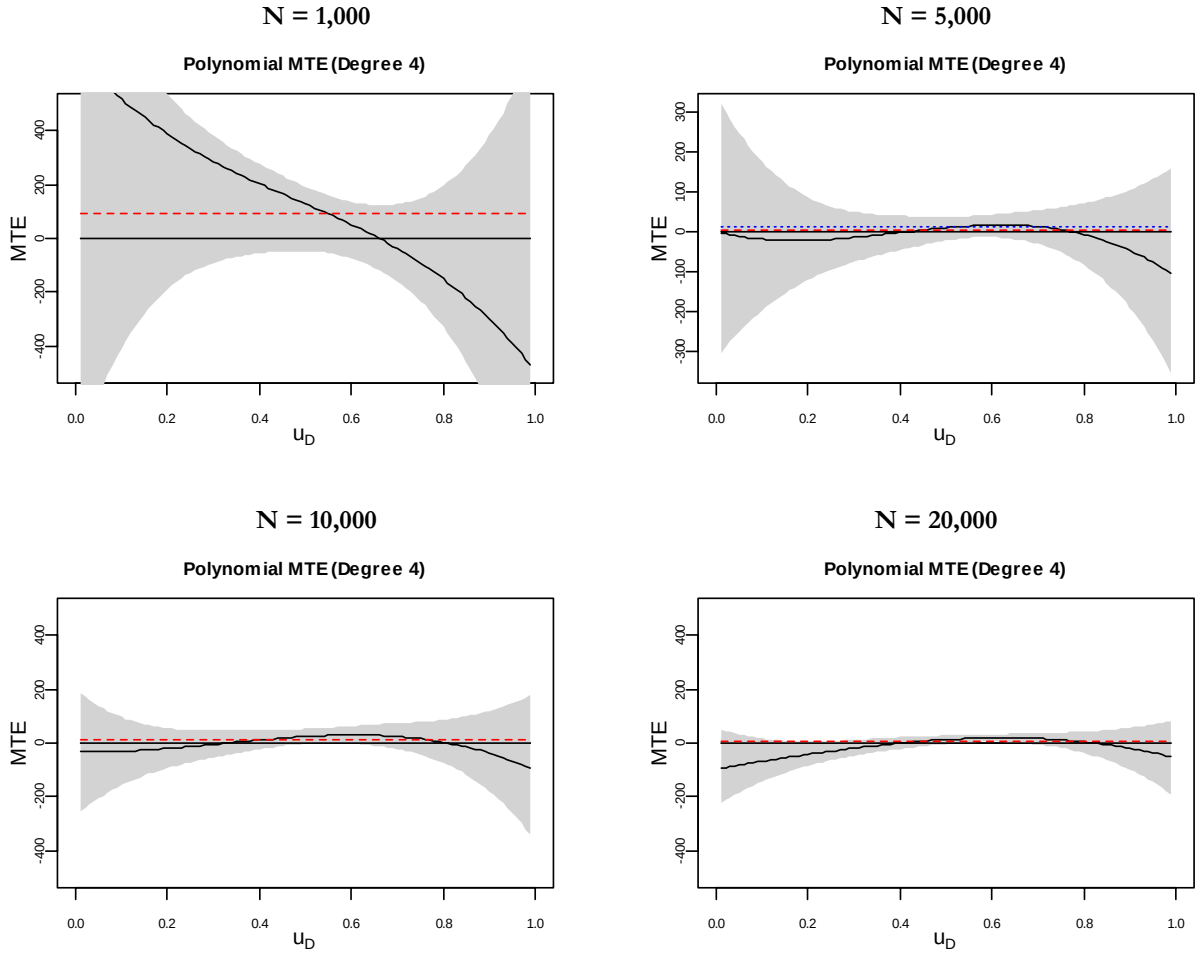
	Size of Sample			
	1,000	5,000	10,000	20,000
A. Conditional moment test				
Probability value of test:	0.7421	0.5440	0.5926	0.0469
Outcome of test:	DNR	DNR	DNR	**Reject**
B. Series test				
Probability value of test:	0.0730	0.1042	0.1617	0.0006
Outcome of test:	DNR	DNR	DNR	**Reject**
C. IV estimates above and below the median				
Probability value of test:	0.7385	0.1113	0.8539	0.0230
Outcome of test:	DNR	DNR	DNR	**Reject**
D. Test of heterogeneity in normal selection model				
Probability value of test:	0.7962	0.1371	0.0508	0.0024
Outcome of test:	DNR	DNR	DNR	**Reject**

Note: The tests are performed on subsamples of the data on school vouchers in Chile used in Heckman, Schmierer and Urzua (2007). "DNR" stands for "Do not reject."

the entire analysis on the fixed reduced sample. Therefore, any resampling procedures in our analysis are done on the same fixed reduced sample, acting as if each were the entire available sample.

Finally, in order to investigate the effect of reducing the sample size on the variability of the MTE, we plot estimates of the MTE from each of the samples in Figure A5. Notice that the point estimates have a similar shape but the confidence bands are much larger for the smaller sample sizes – so large that it is easy to see why we are unable to reject the null of a constant MTE.

Figure A5: Chile voucher schools: marginal treatment effect estimates in smaller samples.



[a] The covariates in the outcome equations are: gender, mother's highest grade completed, father's highest grade completed, number of family members, and household income categories. The instruments are: population and urbanization of one's municipality in 2000, population growth rate between 1998 and 2000, difference between average tuition in voucher schools and average tuition in public schools in one's municipality, difference in average test scores in voucher schools and average test scores in public schools in one's municipality, in addition to all of the X variables. The dependent variable in the probit is 1 if the individual is enrolled in a voucher school, and 0 if the individual is enrolled in a public school. The confidence intervals are found using 100 bootstraps. The MTE estimate shown is taken from a fourth degree polynomial estimate of E(Y|X,P). The sample sizes are 1,000, 5,000, 10,000 and 20,000 in each of the four panels, respectively.

References

Bierens, H. J. (1990, November). A consistent conditional moment test of functional form. *Econometrica 58*(6), 1443–1458.

Heckman, J. J. and P. A. LaFontaine (2006, July). Bias corrected estimates of GED returns. *Journal of Labor Economics 24*(3), 661–700.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2003). Fifty years of Mincer earnings regressions. Technical Report 9732, National Bureau of Economic Research.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006). Earnings equations and rates of return: The Mincer equation and beyond. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Chapter 7, pp. 307–458. Amsterdam: Elsevier.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2008, Spring). Earnings functions and rates of return. *Journal of Human Capital 2*(1), 1–31.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4875–5144. Amsterdam: Elsevier.

Hsieh, C.-T. and M. Urquiola (2006, September). The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program. *Journal of Public Economics 90*(8-9), 1477–1503.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Lindley, D. V. (1957). A statistical paradox. *Biometrika 44*, 187–192.

McEwan, P. J. (2001, August). The effectiveness of public, catholic, and non-religious private schools in Chile's voucher system. *Education Economics 9*(2), 103–128.

McEwan, P. J. and M. Carnoy (2000). The effectiveness and efficiency of private schools in chile's voucher system. *Educational Evaluation and Policy Analysis 22*(3), 213–239.

Romano, J. P. and M. Wolf (2005, March). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

Stock, J. H. and M. Yogo (2005). Testing for weak instruments in linear iv regression. In J. H. Stock and D. W. Andrews (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, Chapter 5. Cambridge University Press.