

Three Essays on
Semiparametric Econometric Evaluation:
Methods and Applications

Dissertation

zur Erlangung des Grades
Doktor der Wirtschaftswissenschaften (Dr. rer. pol.)
am Fachbereich Wirtschaftswissenschaften
der Universität Konstanz

vorgelegt von:

Ruben R. Seiberlich

Tag der mündlichen Prüfung: 09. Juli 2013

1. Referent: Prof. Dr. Winfried Pohlmeier
2. Referent: Prof. Dr. Thomas Hinz

*Meiner Familie:
Michael, Waltraut, Mascha und Steffi*

Danksagung

An dieser Stelle möchte ich den Personen danken, die wesentlich zur Entstehung dieser Arbeit beigetragen haben.

Mein Dank gilt meinem Doktorvater, Herrn Prof. Dr. Winfried Pohlmeier, der mich nicht nur in fachlicher Hinsicht im Verlauf meiner Promotion jederzeit unterstützt hat. Er hat bereits während meines Studiums mein Interesse an der Ökonometrie geweckt und mich bei meinen beiden Diplomarbeiten in vielerlei Hinsicht unterstützt. Ich habe mich an seinem Lehrstuhl immer sehr wohlfühlt und die Arbeit mit ihm hat meine Freude an der Forschung gefördert.

Ebenfalls möchte ich mich bei Prof. Dr. Thomas Hinz bedanken, der sich bereit erklärt hat, als Zweitgutachter meiner Dissertation zu fungieren.

Des Weiteren bedanke ich mich herzlich bei meinen Kollegen, die mich während meiner Promotion begleitet haben. Dabei gilt mein besonderer Dank Fabian Krüger, Derya Uysal, Peter Schanbacher, Laura Wichert, Lidan Großmaß, Hao Liu, Remi Piatek, Christoph Frey, Roxana Halbleib und Zahide Eylem Gevrek-Demiray, die stets zu einer angenehmen und freundschaftlichen Arbeitsatmosphäre beigetragen haben. Ganz herzlich bedanke ich mich auch bei Lisa Green.

In besonderem Maße möchte ich meiner Mutter Waltraut, meiner Schwester Mascha, sowie meinem Vater Michael und seiner Frau Rita danken, die mich seit jeher in jeglicher Hinsicht unterstützt und motiviert haben. Ohne Euch wäre diese Arbeit nicht möglich gewesen.

Zum Schluss möchte ich mich ganz herzlich bei meiner Freundin Stefanie Heinrichs und Ihren Eltern, Margret und Heinz-Leo, bedanken. Steffi war immer eine unglaublich tolle Unterstützung und hat mir bei allen Entscheidungen geholfen, die richtige zu treffen.

Contents

Summary	7
Zusammenfassung	11
1 Educational Performance Gaps in Eastern Europe	16
1.1 Introduction	17
1.2 Overview of the Educational Systems in Eastern Europe	19
1.3 Identification Strategy	20
1.4 Data	26
1.5 Estimation Results	27
1.6 Conclusion	33
Bibliography	35
Appendix 1.A Tables	38
Appendix 1.B Figures	41
2 Semiparametric Decomposition of the Gender Achievement Gap: An Application to Turkey	43
2.1 Introduction	44
2.2 Background and Literature	47
2.3 Data and Descriptive Statistics	50
2.4 Econometric Model	56
2.5 Results	60
2.6 Conclusion	66
Bibliography	67
Appendix 2.A Tables	73
Appendix 2.B Figures	81

3	A Simple and Successful Method to Shrink the Weight	84
3.1	Introduction	85
3.2	Propensity Score Methods	86
3.3	Shrunken Weights	89
3.4	Monte Carlo Study	91
3.5	Conclusion	103
	Bibliography	104
	Appendix 3.A Tables	107
	Appendix 3.B Figures	119
	Appendix 3.C Supplementary Proofs	122
	Complete Bibliography	124
	Eigenabgrenzung	133

List of Tables

1.1	Semiparametric decomposition result for science	28
1.2	Semiparametric decomposition result for reading	29
1.3	Semiparametric decompositions for science between Eastern European countries	31
1.4	Semiparametric decompositions for reading between Eastern European countries	31
1.A.1	Variables' description	38
1.A.2	Weighted means and standard deviations	38
1.A.3	Test score gaps at different quantiles for science	39
1.A.4	Test score gaps at different quantiles for reading	40
2.1	The standard BO decomposition of the gender test score gap in math	62
2.2	The standard BO decomposition of the gender test score gap in science	63
2.3	Semiparametric BO decomposition of the mean test score gap for the common support subpopulation	64
2.4	Semiparametric BO decomposition of the mean test score gap	65
2.A.1	The index of beliefs in own abilities in science	73
2.A.2	The index of motivation in science	74
2.A.3	Descriptive statistics by gender	75
2.A.4	OLS estimates of the gender test score gap in math	77
2.A.5	OLS estimates of the gender test score gap in science	78
2.A.6	Estimates of the responsiveness of test scores to covariates by gender	79
2.A.7	Semiparametric BO decomposition across the distribution for the common support subpopulation	80
3.1	Functional form for $m(q)$	92
3.2	Treated-to-control ratios.	92
3.3	Parameter combinations.	93

LIST OF TABLES

3.4	Descriptive statistics for the optimal λ with known ATE	94
3.5	Average percentage improvement in MSE for the ATE	95
3.6	Descriptive statistics for the optimal λ for known ATE with trimming rule 2 (tr 2)	97
3.7	Average percentage improvement in MSE, fixed valued λ	98
3.8	Average percentage improvement in MSE, $MSE(\hat{p}_i^s)$ -minimizing λ (MSE- min. λ)	99
3.9	Average values for λ obtained through the different methods.	100
3.10	Average percentage improvement in MSE, cross-validated λ	101
3.11	Average percentage improvement in MSE for the most realistic setting and different λ s.	102
3.12	Regression of the MSE improvement on different settings.	103
3.A.1	$MSE(ATE(\hat{p}^s$ with fixed valued $\lambda))$ vs. $MSE(ATE(\hat{p}))$	107
3.A.2	$MSE(ATE(\hat{p}^s$ with MSE-min. $\lambda))$ vs. $MSE(ATE(\hat{p}))$	108
3.A.3	$MSE(ATE(\hat{p}^s$ with data-driven $\lambda))$ vs. $MSE(ATE(\hat{p}))$	109
3.A.4	$MSE(ATE(\hat{p}^s$ with fixed valued $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p}))$	110
3.A.5	$MSE(ATE(\hat{p}^s$ with fixed valued $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p} + \text{tr } 1))$	111
3.A.6	$MSE(ATE(\hat{p}^s$ with fixed valued $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p} + \text{tr } 2))$	112
3.A.7	$MSE(ATE(\hat{p}^s$ with MSE-min. $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p}))$	113
3.A.8	$MSE(ATE(\hat{p}^s$ with MSE-min. $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p} + \text{tr } 1))$	114
3.A.9	$MSE(ATE(\hat{p}^s$ with MSE-min. $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p} + \text{tr } 2))$	115
3.A.10	$MSE(ATE(\hat{p}^s$ with data-driven $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p}))$	116
3.A.11	$MSE(ATE(\hat{p}^s$ with data-driven $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p} + \text{tr } 1))$	117
3.A.12	$MSE(ATE(\hat{p}^s$ with data-driven $\lambda + \text{tr } 2))$ vs. $MSE(ATE(\hat{p} + \text{tr } 2))$	118

List of Figures

1.B.1 Histogram estimates of the propensity score distributions	41
1.B.2 Quantile test score gaps in science	42
1.B.3 Quantile test score gaps in reading	42
2.B.1 The mean gender test score gap in science across OECD countries . .	81
2.B.2 The mean gender test score gap in mathematics across OECD countries	81
2.B.3 Histogram estimates of the propensity score distributions	82
2.B.4 Histogram estimates of the propensity score distributions	82
2.B.5 Histogram estimates of the propensity score distributions	82
2.B.6 Quantile test score gaps in science	83
2.B.7 Quantile test score gaps in math	83
3.B.1 Individual MSE minimizing λ s for $n = 100$	119
3.B.2 Individual MSE minimizing λ s for $n = 100$	119
3.B.3 Individual MSE minimizing λ s for $n = 500$	120
3.B.4 Monte Carlo ATE's for $n = 100$	120
3.B.5 Monte Carlo ATE's for $n = 200$	121
3.B.6 Monte Carlo ATE's for $n = 500$	121

Summary

This dissertation consists of three stand-alone research papers on the semiparametric estimation of treatment effects. Treatment effects refer to the causal effect of a variable on an outcome variable of interest and the semiparametric estimation avoids the parametric assumptions on the outcome equation. The thesis is organized as follows: In the first chapter we analyze the effect of individual characteristics on the test score gaps between different Eastern European countries and Finland. Additionally, we look at the test score gaps between Eastern European countries. The second chapter analyzes the gender test score gap in a within country study for Turkey. In this chapter we evaluate the effect of individual characteristics, family characteristics and school characteristics on the gender test score gap. The third chapter deals with inverse propensity score weighting estimators and double robust estimators. In this chapter a new estimation procedure is developed, which allows to estimate the treatment effects with a lower mean squared error.

The first chapter of this thesis analyzes the Programme for International Student Assessment (PISA) test score gaps between Finland and seven Eastern European countries as well as between Eastern European countries. This chapter is joint work with Alina Botezat and forthcoming in *The Economics of Transition*. Using data from the 2006 survey, we choose Finland as benchmark for our analysis. It is the best performing country in the PISA study and is considered to have the most effective and equitable school system. In the first step we analyze the contribution of individual characteristics to the test score gaps between Finland and Eastern European countries. In the second step we disentangle the PISA test score gaps between Eastern European countries, which had similar educational systems 20 years ago. The precondition that two countries belonged to the same country forms a natural experiment, that reveals how two countries develop over the subsequent years. We extend the semiparametric alternative of the twofold Blinder-Oaxaca decomposition to estimate a threefold decomposition. Our decomposition method has several advantages

over the parametric Blinder-Oaxaca decomposition usually applied in the literature. The semiparametric decomposition relaxes the parametric functional form assumption on the outcome equation. It provides useful information on the gender test score gap not only at the mean but also on the distribution of the gap over the entire test score distribution. In addition, the standard Blinder-Oaxaca decomposition ignores the common support problem. In the semiparametric decomposition, on the other hand, counterfactual outcomes are computed only for the common support subpopulation. Moreover, the semiparametric matching method allows to estimate the missing potential outcome for each individual separately, allowing us to account for arbitrary individual effect heterogeneity. We provide evidence that only a small part of the gap can be attributed to the fact that the Finnish students are better endowed with more favorable family background characteristics. The main part of the gap still remains after controlling for the individual background. The students from Southeastern Europe are those who have the largest potential outcome increase if they would have more of the unobserved factors like other individual characteristics, institutional aspects of the school system, resources, cultural factors and so forth. Moreover, we find that the average test score gaps between Finland and Eastern European countries are mainly due to the fact that the poorly performing students in Finland score much higher than the poorly performing students in the Eastern European countries. Among Eastern European countries our results show that the differences in individual and family background characteristics are highly significant and explain part of the test score gap in science and reading.

The second chapter studies the origin of the gender inequalities in educational performance. This chapter is joint work with Zahide E. Gevrek and has the status revise and resubmit at the journal *Labour Economics*. Gender inequalities in educational performance has been the subject of much research for many decades. Promoting gender equality in education is an important policy goal especially in developing countries as it is associated with greater equality in employment outcomes, lower infant mortality rates, a decrease in the number of early marriages and better investments in education and health of future generations. Using data from the 2006 Programme for International Student Assessment (PISA), this study explores the gender gap in mathematics and science achievement of 15-year-olds in Turkey. Turkey is an interesting case to study as it has the largest average gender test score gap in science and one of the smallest gap in mathematics among OECD

countries. The exploration of gender test score gap is important for the following reasons. First, recent research on the economic impact of human capital investment underlines the prime importance of educational quality over pure schooling attainment. Social scientists use international tests of students' performance in cognitive skills such as mathematics and science as a proxy for education quality. There exists a significant effect of the mathematics test score on annual earnings. Since math and science skills are highly valued in the labor market, understanding the gender patterns in these subject fields allows us to gain insight into the gender wage gap and differential education and labor market choices across genders. Second, using data from the international student achievement tests, empirical growth research documents a significant impact of the quality of education on economic growth. Moreover, educational quality leads to longer school attendance in the developing countries. Thus, educational policies aimed to improve quality of education also help meet goals for educational attainment. For our decomposition we use the parametric Blinder-Oaxaca decomposition as well as a robust kernel estimator, which is a linear combination of the local constant estimator and the local linear estimator. The semiparametric BO decomposition results can be summarized as follows. The mean test score gap is 15.1 points in favor of girls in science while it is not statistically significant in math. Girls possess more of the characteristics associated with high science scores. School characteristics are the most important observable characteristics in explaining the gap, followed by the family background. Our findings suggest that ignoring the common support problem causes the underestimation of the part of the gap attributable to observable characteristics. Moreover, the gender test score gap shows a heterogeneous pattern across the test score distribution. We find that in science, the gap favoring girls is statistically significant until the top quantile and the largest gap occurs at the median. In math, the gap is statistically significant only at the top quantile where boys outperform girls.

In the third chapter, which is joint work with Selver Derya Uysal and Winfried Pohlmeier, a simple way of improving propensity score weighting and double robust estimators in terms of mean squared error (MSE) in finite samples is introduced. The approach achieves a lower MSE by shrinking the propensity score towards the share of treated. This Stein-type simple shrinkage substantially mitigates the problems arising from propensity score estimates close to the boundaries. Even though shrinkage methods are very popular in other areas of statistics and econometrics,

they have not been combined with weighting estimators yet. The proposed shrinkage method is a linear combination of the conditional mean of the treatment variable and its unconditional mean. Like other shrinkage methods the degree of shrinkage is determined by a tuning parameter. We propose three different methods to choose this parameter such that certain optimality conditions are satisfied. First, we consider a simple fixed valued tuning parameter, which only depends on the sample size. Second, we minimize the MSE of our linear combination to choose the optimal value. Third, we propose a cross validation procedure to obtain the optimal tuning parameter. We demonstrate the mean squared error gains in finite samples via a comprehensive Monte Carlo study. We consider homogeneous and heterogeneous treatment, homoscedastic and heteroscedastic error terms as well as different ratios of treatment and control group. Moreover, the simulation design captures different functional forms. Since we construct the shrunk propensity scores in such a way that they converge to the conventional propensity scores our proposed method leads to the same results as the standard approaches in large samples. Therefore, we focus on sample sizes 100, 200 and 500 only. Additionally, we evaluate the finite sample performance with and without applying trimming rules. Our results show that the estimators based on the shrunk propensity scores have a lower MSE than the weighting estimators based on the unshrunk propensity scores in all of the settings if we use the fixed valued or the MSE minimizing tuning parameter. For the cross validated tuning parameter the MSE is reduced in 99.3% of the cases, respectively. If a trimming rule is applied to the proposed approach we are able to decrease the MSE of the ATE in 99.7% of the cases for the fixed valued tuning parameter. For the MSE minimizing and cross validated tuning parameter the MSE is reduced in 98.8% and 96.9% of the cases, respectively. In the rare cases where the MSE is not improved the increase is very small.

Zusammenfassung

Diese Dissertation besteht aus drei eigenständigen Aufsätzen zur semiparametrischen Schätzung von Behandlungseffekten. Behandlungseffekte bezeichnen den kausalen Effekt einer Variablen auf eine bestimmte Zielgröße und die semiparametrischen Schätzungen vermeiden die parametrischen Annahmen bezüglich ihres Zusammenhangs. Die Arbeit ist wie folgt gegliedert: Das erste Kapitel analysiert, in wie weit sich die Differenzen in den Testergebnissen des Programmes zur internationalen Schülerbewertung (PISA) zwischen verschiedenen osteuropäischen Ländern und Finnland auf die unterschiedlichen Charakteristiken der Schüler zurückführen lassen. Zusätzlich werden in diesem Kapitel die Differenzen in den Testergebnissen zwischen osteuropäischen Ländern, die bis vor 20 Jahren ähnliche Bildungssysteme hatten, untersucht. Das zweite Kapitel analysiert den geschlechtsspezifischen Unterschied in den Testergebnissen der PISA-Studie innerhalb der Türkei. In diesem Kapitel wird die Wirkung der Charakteristiken der Schüler, des Familienhintergrundes und der unterschiedlichen Schulcharakteristiken auf das geschlechtsspezifische Testergebnis untersucht. Das dritte und letzte Kapitel befasst sich mit so genannten “Inverse Propensity Score Weighting Estimators” sowie doppelt robusten Verfahren zur Schätzung eines Behandlungseffektes. Darin wird ein neues Schätzverfahren entwickelt, das es erlaubt, die Behandlungseffekte mit einem niedrigeren mittleren quadratischen Fehler zu schätzen.

Das erste Kapitel dieser Arbeit analysiert die Unterschiede in den Testergebnissen der PISA-Studie zwischen Finnland und sieben osteuropäischen Staaten sowie zwischen osteuropäischen Staaten untereinander. Dieses Kapitel ist eine gemeinsame Arbeit, die ich mit Alina Botezat geschrieben habe und die in Kürze in *The Economics of Transition* erscheinen wird. Anhand der Daten der PISA-Studie 2006 wählen wir Finnland als Maßstab für unsere Analyse, da es die höchste durchschnittlich Punktezahll aller teilnehmenden Staaten erzielte. Das finnische Schulsystem wird außerdem häufig als das beste, das effektivste und gerechteste Schul-

system der Welt bezeichnet. Im ersten Schritt analysieren wir, in wie weit sich die Unterschiede in den Testergebnissen zwischen Finnland und verschiedenen osteuropäischen Staaten auf die individuellen Charakteristika der Schüler zurückführen lassen. Im zweiten Schritt betrachten wir die Differenzen in den Testergebnissen zwischen Tschechien und der Slowakei, sowie zwischen Estland und Lettland, also Staaten, die bis vor 20 Jahren zum jeweils gleichen Staat gehört haben und somit ähnliche Bildungssysteme hatten. Die Voraussetzung, dass zwei Staaten zu dem gleichen Staat gehörten, bildet die Grundlage eines natürlichen Experiments, das es erlaubt, die unterschiedliche Entwicklung in beiden Staaten über die vergangenen Jahre zu betrachten. Unser Verfahren zur Zerlegung der Unterschiede hat mehrere Vorteile gegenüber der parametrischen Blinder-Oaxaca-Zerlegung, die in der Literatur üblicherweise angewandt wird. Das von uns verwendete semiparametrische Verfahren beruht nämlich nicht auf parametrischen Annahmen über die funktionale Form der PISA-Testgleichung. Des Weiteren werden, im Vergleich zur Standard Blinder-Oaxaca-Zerlegung, nur die Beobachtungen in Betracht gezogen, die, gegeben den Regressoren, auch vergleichbar sind. Darüber hinaus erlaubt die semiparametrische Matching-Methode individuelle Heterogenität in der kontrafaktischen Evidenz. Unsere Ergebnisse zeigen, dass nur ein kleiner Teil der Differenzen in den Testergebnissen auf die Tatsache zurückzuführen ist, dass die finnischen Schüler mit besseren Eigenschaften ausgestattet sind. Der Hauptteil der Differenzen bleibt auch bestehen, wenn man für die unterschiedlichen, individuellen Hintergrund kontrolliert. Die Studenten aus Südosteuropa sind diejenigen, die am meisten davon profitieren würden, wenn sie mehr der unbeobachteten Faktoren (wie andere individuelle Merkmale oder die institutionellen Aspekte des Schulsystems, Ressourcen, kulturelle Faktoren usw.) der Finnen haben würden. Außerdem finden wir, dass sich die durchschnittlichen Differenzen in den Testergebnissen zwischen Finnland und den osteuropäischen Staaten hauptsächlich auf die schlechteren Schüler zurückführen lassen. Relativ schlechte Schüler aus Finnland erzielen wesentlich höhere Punktzahlen als die schlecht abscheidenden Schüler aus den osteuropäischen Staaten. Wenn wir die Differenzen in den Testergebnissen zwischen osteuropäischen Staaten betrachten, sieht man, dass die Unterschiede in individuellen und familiären Hintergrundvariablen signifikant sind und somit einen Teil der Differenz in den Testergebnissen erklären. Dies gilt sowohl in dem naturwissenschaftlichen Test, als auch für die Ergebnisse im Lesetest.

Das zweite Kapitel befasst sich mit der Herkunft geschlechtsspezifischer Ungleichheiten in schulischen Leistungen. Es ist eine gemeinsame Arbeit mit Zahide E. Gevrek und hat den Status "revise and resubmit" bei der Zeitschrift *Labour Economics*. Forschungen über geschlechtsspezifische Ungleichheiten in schulischen Leistungen waren über viele Jahrzehnte Gegenstand zahlreicher Untersuchungen. Die Förderung der Gleichstellung der Geschlechter in der Bildung ist eines der wichtigsten politischen Ziele. Dies gilt insbesondere in Entwicklungsländern, da geschlechtsspezifische Ungleichheiten in schulischen Leistungen mit Gleichbehandlung in der Beschäftigung, niedrigerer Säuglingssterblichkeit, einem Rückgang der Zahl der Eheschließungen im Jugendalter und größeren Investitionen in Bildung und Gesundheit der zukünftigen Generationen korrelieren. Anhand der PISA-Daten aus dem Jahr 2006 untersucht diese Studie die geschlechtsspezifischen Unterschiede in Mathematik und den Naturwissenschaften der 15-Jährigen Schüler/-innen in der Türkei. Dieser Staat ist ein interessanter Fall, da es dort unter allen teilnehmenden OECD-Staaten den größten geschlechtsspezifischen Unterschied in den naturwissenschaftlichen Testergebnissen gibt, aber gleichzeitig auch der kleinste Unterschied bezüglich der Testergebnisse in Mathematik beobachtet werden kann. Die Erforschung der Unterschiede in den Testergebnissen von Jungen und Mädchen ist aus folgenden Gründen wichtig. Erstens unterstreicht die neuere Forschung auf dem Gebiet der wirtschaftlichen Auswirkungen der Investitionen in das Humankapital die vorrangige Bedeutung der pädagogischen Qualität im Vergleich zur reinen Schulbildung, also dem daraus resultierenden Signal. Da sozialwissenschaftliche Untersuchungen zeigen, dass in internationalen Tests die Leistungen der Schüler/-innen in Fächern wie Mathematik und Naturwissenschaften als Proxy für die Qualität der Bildung dienen, können Rückschlüsse auf diese Qualität gezogen werden. So gibt es beispielsweise einen signifikanten Effekt der Testergebnisse in Mathematik auf den Jahresverdienst. Da mathematische und naturwissenschaftliche Fähigkeiten auf dem Arbeitsmarkt besonders gut entlohnt werden, ermöglicht das Verständnis der geschlechtsspezifischen Muster in diesen Bereichen Einblicke in das geschlechtsspezifische Lohngefälle sowie die unterschiedlichen Bildungs- und Arbeitsmarktentscheidungen von Männern und Frauen. Zweitens wurde in der Wachstumsforschung anhand von internationalen Vergleichstests ein wesentlicher Einfluss der Bildungsqualität auf das Wirtschaftswachstum nachgewiesen, sodass durch die Schließung der geschlechter-spezifischen Unterschiede auch eine Auswirkung auf dieses Wachstum zu erwarten ist. Des Weiteren hat die Qualität der Schulbildung, insbesondere in Entwick-

lungsländern, einen direkten Einfluss auf die Bildungsabschlüsse der Schüler/-innen. Für die Untersuchung der geschlechtsspezifischen Ungleichheiten in schulischen Leistungen verwenden wir zusätzlich zur parametrischen Blinder-Oaxaca Zerlegung eine semiparametrische Alternative, die auf einem robusten Kerndichteschätzer beruht. Die Ergebnisse der semiparametrischen Zerlegung können wie folgt zusammengefasst werden: In den Naturwissenschaften erzielen die Mädchen im Durchschnitt ein Testergebnis, das 15.1 Punkte besser ist als das der Jungen, wohingegen wir in Mathematik keinen signifikanten Unterschied finden. Die Schulcharakteristiken sind die wichtigsten beobachtbaren Eigenschaften bei der Erklärung der Differenzen, gefolgt vom Familienhintergrund. Darüber hinaus zeigen unsere Ergebnisse, dass sich die Differenzen in den Testergebnissen über die Verteilung ändern. Im naturwissenschaftlichen Test ist die Differenz über alle Quantile signifikant und die größte Differenz ist am Median. In Mathematik ist die Differenz lediglich am obersten Quantil statistisch signifikant. Hier schneiden Jungen besser ab als Mädchen.

Das dritte Kapitel entstammt einer gemeinsamen Arbeit mit Selver Derya Uysal und Winfried Pohlmeier. Darin wird eine einfache Möglichkeit zur Verbesserung der "Propensity-Score Weighting" und doppelt robusten Schätzern hinsichtlich der mittleren quadratischen Fehler (MSE) in endlichen Stichproben vorgeschlagen. Dabei werden die Gewichte auf Basis einer linearen Kombination aus der konditionalen Wahrscheinlichkeit und der un konditionalen Wahrscheinlichkeit bestimmt. Dieser Stein-Typ-Ansatz der Schätzung verringert die Probleme, die durch konditionale Wahrscheinlichkeiten nahe Null und Eins entstehen. Obwohl diese Methoden sehr beliebt und in anderen Bereichen der Statistik und Ökonometrie weit verbreitet sind, wurden sie noch nie auf diese Art der Schätzer angewandt. Um die konditionalen Wahrscheinlichkeiten mit der un konditionalen Wahrscheinlichkeit kombinieren zu können, muss ein Komplexitätsparameter bestimmt werden. Wir entwickeln in unserem Artikel drei verschiedene Varianten, um diesen Parameter zu bestimmen. Alle drei Verfahren befriedigen bestimmte Optimalitätsbedingungen. Zunächst betrachten wir einen einfachen, festen Wert für die Wahl des Komplexitätsparameters, in dem wir ihn lediglich von der Stichprobengröße anhängig wählen. Die zweite Variante basiert auf der Minimierung des MSE der vorgeschlagenen linearen Kombination, um den optimalen Wert zu wählen. Drittens schlagen wir ein Verfahren vor, das nur von den Daten abhängig ist. Der optimale Komplexitätsparameter wird dabei durch eine Kreuzvalidierung bestimmt. Durch eine umfassende Monte Carlo Studie

zeigen wir die Verbesserungen des mittleren quadratischen Fehlers in endlichen Stichproben. Wir betrachten hierfür homogene und heterogene Behandlungseffekte, homoskedastische und heteroskedastische Fehlerterme, sowie unterschiedliche Verhältnisse der Behandelten- und der Kontrollgruppe. Darüber hinaus erfasst das Design unserer Simulation verschiedene funktionale Formen. Da die Linearkombination aus konditionaler Wahrscheinlichkeiten und den unkonditionalen Wahrscheinlichkeiten zu den ursprünglichen konditionalen Wahrscheinlichkeiten konvergiert, betrachten wir in unserer Monte Carlo Studie lediglich die Stichprobenumfänge 100, 200 und 500. Darüber hinaus bewerten wir die Schätzungen mit und ohne Anwendung so genannter Trimming-Regeln. Unsere Ergebnisse zeigen, dass die Verwendung der Linearkombination aus der konditionalen Wahrscheinlichkeit und den unkonditionalen Wahrscheinlichkeiten den mittleren quadratischen Fehler der Schätzer für die Behandlungseffekte in allen Fällen reduziert, wenn man den Komplexitätsparameter lediglich abhängig von der Stichprobengröße oder anhand der MSE-Minimierung wählt. Bestimmt man ihn durch Kreuzvalidierung, kann man den MSE in 99.3% der Fälle verbessern. Wird zusätzlich eine Trimming-Regel angewandt, verbessert sich der MSE der Schätzer für die Behandlungseffekte in 99.7% der Fälle, wenn man den Komplexitätsparameter lediglich abhängig von der Stichprobengröße wählt. Wird er anhand der MSE-Minimierung bestimmt, verbessert sich der MSE der Schätzer in 98.8% der Fälle und in 96.9% der Vergleiche, wenn man ihn mit Hilfe einer Kreuzvalidierung bestimmt.

CHAPTER 1

Educational Performance Gaps in Eastern Europe

1.1 Introduction

Over the past twenty years, Eastern European countries have gone through periods of transition and structural changes which also affected the educational system. Most Eastern European countries have adopted reforms to adapt the educational system to the new requirements of the job market. The success of these reforms in education can be assessed by analyzing the results of international standardized test scores such as PISA (Programme for International Student Assessment), TIMSS (Third International Mathematics and Science Study), or PIRLS (Progress in International Reading Literacy Study). The results from PISA 2006, for example, show that there is a high variation in performance of the Eastern European countries. Many of the Eastern European countries are still in a transition process and have not yet overcome the initial disadvantages compared to Western countries. Most of them perform statistically significantly below the OECD average and only Estonia, Slovenia and the Czech Republic perform in the upper part of the distribution (OECD (2007)).

The first aim of this paper is to analyze the PISA test score gaps between Finland and seven Eastern European countries (Estonia, Czech Republic, Hungary, Romania, Bulgaria, Latvia and Slovakia). Using data from the 2006 survey, we choose Finland as the benchmark for our analysis. It is the best performing country in the PISA study and is considered to have the most effective and equitable school system (Ammermüller (2007)). Our results help to understand how much of the gap can be attributed to individual and family background characteristics and how much is due to other factors.

The second aim is to disentangle the PISA test score gap between countries which had similar educational systems 20 years ago. Estonia as well as Latvia belonged to the Soviet Union until 1991, the Czech Republic and Slovakia together formed Czechoslovakia until the end of 1992. The precondition that two countries belonged to the same country forms a natural experiment, that reveals how two countries, which start from more or less the same point, develop over the subsequent years.

To achieve the two aims, we disentangle the effects that explain the gaps in order to show which factors contribute to the differences in school performance. More specifically, we look at the extent to which the differences in individual and family

background characteristics contribute to explaining the observable gaps in school performance. Thus, we should be able to answer the following questions: Which educational system manages to generate high returns to these important individual and family background characteristics? What would be the expected outcome of the students from one country, if, given their individual characteristics, they would attend the school system of a country that on average performs better than their home country?

This paper contributes to the previous literature in several ways: First of all, it makes an original contribution by introducing a semiparametric method to estimate a threefold decomposition into the educational literature. Thus far such a semiparametric method is only used to estimate a twofold decomposition, especially in explaining the gender differences in wages, but not in the research regarding the decomposition of differences in school outcomes. More exactly, the methodology applied here is a semiparametric version of the threefold Blinder-Oaxaca decomposition which disentangles the effects in an endowment, return and an interaction effect between these two. This is important in its own right since recent papers have demonstrated that the functional form assumptions of the parametric Blinder (1973) and Oaxaca (1973) decomposition can give misleading results (Barsky et al. (2002), Mora (2008)). The method is based on an approach proposed by Frölich (2007), who uses propensity score matching to compute the counterfactual mean. Furthermore, this is the first paper that decomposes the differences in PISA test scores between the best performing country in the study and several Eastern European countries as well as between some Eastern European countries.

The remainder of the paper is the following: The next section provides a general overview of the educational systems in Eastern Europe. The section 1.3 focuses on the identification strategy used to decompose the gap in school performance. Section 1.4 presents the PISA study 2006 and describes the data. Section 1.5 discusses the estimation results. The last section concludes.

1.2 Overview of the Educational Systems in Eastern Europe

According to Cerych (1997) and Radó (2001) the following issues of the school systems in post-communist Eastern Europe can be identified. In all countries, a depolitisation of education took place, implying the end of ideological control and orientation of the system. Furthermore, educational change led to the decentralization and liberalization in educational management by breaking down the state monopoly. Moreover, the pupils or their parents, respectively, now have freedom of choice concerning their educational path. Another issue of the reforms was redefining the quality in education. During communism, the most important indicators for quality was the participation rates and the achievement of the most talented students (Radó (2001)).

Our sample consists of following countries from Eastern Europe: Romania, Bulgaria, Hungary, Czech Republic and Slovakia, Estonia and Latvia. Even if these countries started reforms at the same time, their subsequent evolution was different, depending, especially, on the development and the speed of economic reforms. For example, countries, such as, Estonia, Czech Republic and Hungary went through a process of rapid privatization (Bjørnskov and Potrafke (2011)). They are also among the Eastern European countries performing the best in PISA test scores. Thus, with few exceptions, we cannot speak of continuity in educational reforms as long as they depend on factors outside the system itself. Only in the case of Hungary and Estonia were educational policies undivided, due to measures taken before 1989 (Radó (2001)). The Estonian schools already won a degree of autonomy regarding the content of curricula during the Soviet period when textbooks were predominantly written by Estonian authors (Kitsing (2008)).

Generally, previous empirical research on the school performance of Eastern European countries is quite limited, providing mixed results and inconclusive evidence. One reason was the lack of reliable data that can objectively describe the educational process in these countries. Before 1989, data reported on human capital stock (years of schooling, for example) were over-estimated (Beirne and Campos (2007)) and, after 1989, the participation at the international standardized tests (TIMSS, PIRLS, PISA) was not the same for all countries. Estonia, for example, participated

for the first time in the PISA Study in 2006. The existence of such comparative data and of cross-national individual-level survey has allowed the extension of research in the last years, promising to answer key questions concerning the quality of the educational system in Eastern Europe.

For the transition period, the paper by Ammermüller et al. (2005) provides evidence regarding the production of school quality in Eastern European countries. Even if these countries faced similar characteristics in the economic and political development, the impact of individual factors, school resources and institutional settings on school performance shows different patterns. Using TIMSS data from 1995, the authors show that the student's background has a lower impact in those countries which perform worse (Lithuania, Latvia and Romania) and which adopted reforms regarding the school system later than the other countries. The largest effects are obtained in Czech Republic and Hungary. The impact of school resources and teacher characteristics on school performance is low in magnitude and does not necessarily indicate a particular pattern. Only in some cases (Romania, Czech Republic and Hungary), better training and richer experience of the teachers can positively influence the test scores. The most favorable institutional setting is in Czech Republic, although the results show that the variation in test scores cannot be explained by institutional differences between countries. All in all, Ammermüller et al. (2005) show substantial effects of student background on educational performance and much lower impact of resources and institutional settings.

Based on these findings, our purpose is to quantify the gaps in cognitive skills of children from Eastern Europe, which is due to differences in individual and family background characteristics.

1.3 Identification Strategy

One of the central themes in economics of education is to measure the school achievement gaps. The analysis of disparities in school performance are focused either on the gender gap in different subjects (Fryer and Levitt (2010), Niederle and Vesterlund (2010)), on the differences between countries (McEwan and Marshall (2004), Ammermüller (2007)), and between different subgroups (Card and Rothstein (2007), Patacchini and Zenou (2009), Krieg and Storer (2006), Duncan and Sandy (2007),

Schneeweis (2011)).

All of these studies use a parametric approach and most of them used the Blinder-Oaxaca decomposition or a modified parametric version of it. The traditional Blinder-Oaxaca decomposition determines the source of the differences at the means and breaks down a gap into two parts by estimating one counterfactual mean. The first part, the characteristics effect, can be explained by the differences in the characteristics of individuals and the second part, commonly known as the unexplained gap, is a structure effect, which reflects the differences in slope coefficients. A comprehensive overview of the Blinder-Oaxaca decomposition is provided by Fortin et al. (2010). The main disadvantages of the Blinder-Oaxaca decomposition are the ignorance of the common-support problems and the functional form assumptions.

To avoid these drawbacks, we apply a semiparametric method, which does not assume a specific functional form of the outcome equations. Moreover, the counterfactual mean is computed using only those individuals who are actually comparable. The semiparametric matching method also accounts for arbitrary individual effect heterogeneity (Heckman et al. (1999), Imbens (2004)).

This semiparametric method identifies the counterfactual mean as it is done in the evaluation literature. There, the interest usually lies in the estimation of the effect of a program. To isolate the true effect of the program, the observed outcome has to be compared to the outcome that would have resulted had the individuals not been treated (not participated in the program). To estimate this counterfactual mean, information on the non-participants is used. One possibility is to match treatment with comparison units that are similar in terms of their observable characteristics. Generally, matching directly on the vector of characteristics would be computationally demanding and, due to the curse of dimensionality, it would become hard to find good matches if the number of covariates is large.

To overcome this problem, Rosenbaum and Rubin (1983) demonstrate that matching can be done on a single-index variable, namely the propensity score. Frölich (2007) is the first to use such a matching procedure outside the treatment evaluation literature. He shows that mean independence is sufficient for consistency of propensity score matching and uses it to decompose the gender wage gap analogously to the

Blinder-Oaxaca decomposition into a characteristics and return effect. In this paper, we will extend this procedure to estimate a threefold decomposition.

To obtain the propensity score, we estimate the probability that an individual belongs to the better performing country ($D = 1$) by a logit regression, i.e.

$$p = \Pr[D = 1|X = x] = F(x'\beta) \quad (1.1)$$

where $F(x'\beta)$ represents the cumulative logistic distribution. In the next step, the density of this propensity score is estimated using a Gaussian Kernel estimator. Kernel matching then uses all members of one group to generate a match for each observation in the other group. The contribution of each member is thereby determined by the bandwidth and is smaller, the poorer the match is. Following Frölich (2004), we select the bandwidths by leave-one-out cross-validation to minimize the least-squares criterion and choose as bandwidth search grid $0.01\sqrt{1.2^{g-2}}$ for $g = 1, \dots, 59$ and ∞ .

To apply propensity score matching, we only use data at the individual level. We refer here to measures for the students' characteristics (age and gender) and for family background (number of books at home, parents' education). These variables are commonly used to measure the (in)equality of educational opportunities (Wößmann (2008), Schütz et al. (2008), Martins and Veiga (2010)). From these indicators, the number of books is preferable, being the most important measure of family background, which best predicts the student performance (Wößmann (2003), Fuchs and Wößmann (2007), Wößmann (2008)). As pointed out in the literature (Schütz et al. (2008)), due to the heterogeneity in the structure of school systems, a certain level of parents' education in one country may correspond to a different level in another country. This may affect the comparability of the impact that the parental education has on children's school performance across countries. Despite this drawback, we nonetheless use information on parents' education in order to capture the intergenerational genetic transmission of abilities that are also associated with the educational achievement of children (Plug and Vijverberg (2003)).

Under these considerations, we intend to measure precisely how much of the total gap can be explained by differences in the distributions of observable individual and family background characteristics and how much of the gap is due to other factors,

such as school resources and different institutional features of the school system.

We decide not to include school variables in the estimation of the propensity score for the following two main reasons. First of all, the matches become poor when including school and educational resources variables as some of them - like comprehensive schooling, for example - are almost perfect predictors for the respective country. Secondly, the educational resources are not randomly allocated into schools (Schneeweis (2011)) and, thus, may distort the impact they have on school achievement.

Let $f_1(p)$ be the distribution of the propensity score $p = p(X)$ among those from country $D = 1$ (the better performing country) and $f_0(p)$ the distribution among those pupils from country $D = 0$ (the worse performing country). In such a way, the test score gap

$$\Delta = E[Y^1|D = 1] - E[Y^0|D = 0] \tag{1.2}$$

where Y^d indicates the outcome of those from country $D = d$, for $d \in \{0, 1\}$, can be expressed as

$$\Delta = \int E_1[Y|p(x) = p]f_1(p) dp - \int E_0[Y|p(x) = p]f_0(p) dp \tag{1.3}$$

where $E_1[Y|p(x) = p] = E[Y|p(x) = p, D = 1]$ and $E_0[Y|p(x) = p] = E[Y|p(x) = p, D = 0]$

The common support is evaluated by comparing the distributions (histograms) of the estimated propensity scores by the treatment variable as suggested in Lechner (2010). Figure 1.B.1 of Appendix 1.B shows that for each country comparison there are individuals with similar propensity scores from both countries. Thus, the histograms do not indicate overlap problems and, therefore, we estimate the counterfactual means without applying any common support correction.¹

¹If we follow Dehejia and Wahba (1999) and use only those observations for the estimation that have a propensity score which is lower than the maximum propensity score of the control group and higher than the minimum propensity score in the treated group, the estimation results do not change. These results are available upon request.

Moreover, we assume mean independence given x . If $E[Y|D = 0, X = x] = E[Y|D = 1, X = x]$ holds, Frölich (2007) shows that the counterfactual means are identified by estimating

$$E[Y^1|D = 0] = \int E_1[Y|p(x) = p]f_0(p) dp \text{ and} \quad (1.4)$$

$$E[Y^0|D = 1] = \int E_0[Y|p(x) = p]f_1(p) dp \quad (1.5)$$

where the counterfactual mean for $p(x) = p$ can be estimated by the Nadaraya-Watson estimator

$$\hat{E}_d [Y|p(x) = p] = \frac{\sum_i^n \mathbb{1}\{D_i = d\}K\left(\frac{p-p_i}{h}\right)Y_i}{\sum_i^n \mathbb{1}\{D_i = d\}K\left(\frac{p-p_i}{h}\right)}, \text{ for } d \in \{0, 1\} \quad (1.6)$$

Thereby, K is the kernel function, h the bandwidth and n the number of observations. The first counterfactual $E[Y^1|D = 0]$ gives the expected outcome those from country $D = 0$ would have in country $D = 1$.²

In order to disentangle the effects of the gap, we extend the procedure applied by Frölich (2007) by decomposing the gap into three parts, where $D = 1$ always denotes the better performing country:

$$\begin{aligned} & \int E_1[Y|p(x) = p]f_1(p) dp - \int E_0[Y|p(x) = p]f_0(p) dp \\ &= \underbrace{\int E_0[Y|p(x) = p][f_1(p) - f_0(p)] dp}_{\Delta_c} \\ &+ \underbrace{\int [E_1[Y|p(x) = p] - E_0[Y|p(x) = p]] f_0(p) dp}_{\Delta_r} \\ &+ \underbrace{\int [E_1[Y|p(x) = p] - E_0[Y|p(x) = p]][f_1(p) - f_0(p)] dp}_{\Delta_{cr}} \end{aligned} \quad (1.7)$$

In terms of the Blinder-Oaxaca decomposition, the first term can be attributed to differences in the distributions of individual characteristics and is, therefore, the characteristics effect (Δ_c). It captures the difference of the test scores that would vanish if the characteristics of the students from the worse performing country would follow

²Note that the problem of self-selection does not occur in our context as the treatment is the attendance of a school system in another country. Since we only use natives and second generation immigrants (see section 1.4), this cannot be influenced by the individuals.

the same distribution as those of the students from the better performing country. The second summand is the part of the gap that can be explained by those factors, other than the few individual characteristics described above, that determine the school performance (e.g. other individual characteristics, institutional aspects of the school system, resources, cultural factors etc.). It is analogous to the return effect (Δ_r) in the Blinder-Oaxaca decomposition. The term in the last brackets (Δ_{cr}) is the interaction effect between the characteristics and the return effect, reflecting the fact that the gap could also be determined by the simultaneous existence of differences in the distributions of individual characteristics and in the returns.

We decide to apply the threefold decomposition, used for the first time in decomposing the gap in test score by Ammermüller (2007), for the following reason. When we have to decompose a gap in test score, we should take into account that individuals can be better endowed with characteristics that, at the same time, are better rewarded by their school systems than by the other school system.

In our case, the interaction term (if positive) expresses how much better the students from the worse performing country would score on average if the students from the better performing country did not have the advantage of being better endowed with those characteristics that are also better rewarded in terms of test scores in their country, or less endowed with those characteristics that are better rewarded in the worse performing country.

Compared to the parametric Blinder-Oaxaca decomposition, the approach applied here does not specify the regression function as linear.

To analyze the heterogeneous pattern of the test score gaps across the test score distribution we additionally look at the gaps at different quantiles:

$$\Delta^\tau = F_{y^1|D=1}^{-1}(\tau) - F_{y^0|D=0}^{-1}(\tau)$$

where $F_{y^1|D=1}^{-1}(\tau)$ ($F_{y^0|D=0}^{-1}(\tau)$) is the τ -quantile of the test score distribution among country 1 (country 0).

All standard errors of our estimates are obtained by bootstrapping, using 1000 bootstrap iterations.

1.4 Data

The following analysis is based on data from PISA 2006. PISA assesses the achievement of 15-year-olds in mathematics, reading and science literacy. Apart from test scores, data on pupils' social and cultural background were collected as well as information about the school environment of students (OECD (2007)).

The data contain information on more than 35 000 students and more than 2000 schools. For comparison reasons, the scores have been standardized to a mean of 500 and a standard deviation of 100. Our sample consists of data from Finland and seven Eastern European countries: Estonia, Czech Republic, Hungary, Romania, Bulgaria, Latvia and Slovakia. A general description of the variables used in this study is given in Table 1.A.1 of Appendix 1.A. Since the performance of the immigrants from the first-generation could also reflect the influence of other school systems than the one they currently attend, we decide to drop these students from the samples. Moreover, the share of first generation immigrants was quite different for the countries in our sample.

Having to deal with a high volume of data, the problem of missing data in PISA study is inevitable. As Ammermüller (2007) noted, dropping individuals with missing information could lead to an upward bias in test scores, since the missing data are not missing at random, being predominant among students who have low test scores. One solution to overcome this problem is to predict the values of these data using the complete information available from all students. Thus, we decide to impute all the missing values by applying a method suggested by Wößmann et al. (2009).

Table 1.A.2 of Appendix 1.A presents the weighted means and standard deviations for the variables used in our study.

The descriptive statistics reported in Table 1.A.2 show some differences in observable characteristics between students from different countries. Looking at the number of books, more than a third of students from Bulgaria and Romania have less than 25 books at home, while the corresponding percentage in the other countries is between 16 and 20. In all countries, the parents are well educated, but some differences can still be noticed. In Finland, the majority of the parents have a tertiary education whereas the majority in the Eastern European countries have upper secondary edu-

cation. Among the Eastern European countries, the parents in Czech Republic and Slovakia are best educated. In both countries, more than 75 percent of the students have parents who completed upper secondary education.

According to data from Table 1.A.2, the range of differences in test scores between Finland and countries from Eastern Europe is very large. It is between 152 points (Finland - Romania in reading) and 32 points (Finland - Estonia in science). Also, the spread of the test scores in countries from Eastern Europe is very different: higher in Bulgaria and in Czech Republic, lower in Estonia, Latvia and Romania.

1.5 Estimation Results

To estimate the different components of the PISA test score gap, we include the individual and family background variables explained above in the estimation of the propensity score. Since the estimation results are similar for math and science from the point of view of the magnitude and sign effects, we only report the science results. All of our decompositions are formulated from the point of view of the worse performing country ($D = 0$).

Results for the Decompositions of the Science score gaps between Finland and Eastern European countries

Table 1.1 shows the results of the semiparametric decompositions for the science PISA test scores between Finland and seven Eastern European countries. The first striking result is that, for all seven countries, the return effect is significantly positive and the effect with the largest magnitude. This indicates that, given their average characteristics, the students from each of the seven Eastern European countries would have on average higher test scores in science if they attended the Finnish school system.

1. EDUCATIONAL PERFORMANCE GAPS IN EASTERN EUROPE

Table 1.1: Semiparametric decomposition result for science

Countries	Δ_c	Δ_r	Δ_{cr}	Δ
FIN-EST	3.04 (1.96)	30.39*** (2.07)	-1.42 (2.34)	32.02*** (1.73)
FIN-CZE	5.05 (3.02)	54.74*** (3.10)	-8.43** (3.85)	51.35*** (1.85)
FIN-HUN	0.76 (1.98)	59.04*** (2.20)	1.05 (2.36)	60.85*** (1.82)
FIN-LTV	-1.05 (2.41)	69.05*** (2.22)	5.94** (2.81)	73.94*** (1.84)
FIN-SLK	4.66** (3.44)	63.34*** (2.31)	8.15* (3.78)	76.15*** (1.89)
FIN-BUL	29.86*** (1.96)	115.55*** (2.25)	-16.47*** (2.19)	128.94*** (2.02)
FIN-ROM	10.76*** (1.65)	134.10*** (2.04)	2.33 (1.81)	147.20*** (1.92)

Note: The country which has worse performance is always the reference country. Standard errors are in brackets and simulated with 1000 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

Particularly the pupils from Bulgaria and Romania would profit from a such school system, making it possible to increase their score in science by more than 100 points on average.

The characteristics effect is smaller in magnitude and only significant for three country comparisons that include the poorer performing countries. It is positive for six countries, but only significant for the comparisons Finland-Slovakia, Finland-Bulgaria and Finland Romania. This reveals that the Finnish students tend to have, on average, slightly more favorable characteristics than the students from Eastern European countries. We only obtain a negativ characteristics effect when we compare Finland with Latvia, but this effect is insignificant.

The interaction effect is significantly positive only when we compare Finland with Latvia and Finland with Slovakia, showing that the gap would be smaller if the Finnish students did not have the advantage of being better endowed with those characteristics which are also better rewarded by the Finnish school system compared to the other school system. For Finland-Bulgaria and Finland-Czech Republic, the interaction effects are significantly negative.

All in all and under the assumption of mean independence given the covariates, our estimation results suggest that the higher average score in science in Finland is not due to a better individual and family background of the Finnish students, but rather to the fact that the Finnish school system is more efficient in transforming the given inputs into PISA test score points.

Results for the Decompositions of the Reading score gaps between Finland and Eastern European countries

Table 1.2 contains the results for the PISA reading scores.

Table 1.2: Semiparametric decomposition result for reading

Countries	Δ_c	Δ_r	Δ_{cr}	Δ
FIN-EST	-5.48** (2.12)	42.38*** (2.28)	7.93*** (2.66)	44.82*** (1.70)
FIN-CZE	5.98* (3.45)	62.04*** (3.09)	-4.38 (4.01)	63.64*** (2.03)
FIN-HUN	-0.28 (2.29)	61.41*** (2.18)	4.06 (2.59)	65.19*** (1.84)
FIN-LTV	-2.58 (2.65)	61.75*** (2.90)	8.51** (3.58)	67.41*** (1.82)
FIN-SLK	2.10 (3.97)	67.58*** (2.32)	11.47** (4.23)	81.15*** (1.92)
FIN-BUL	30.50*** (2.11)	127.59*** (2.46)	-15.11*** (2.39)	142.98*** (2.14)
FIN-ROM	9.42*** (1.91)	138.15*** (2.05)	4.92** (2.02)	152.49*** (1.95)

Note: The country which has worse performance is always the reference country. Standard errors are in brackets and simulated with 1000 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

It can be seen that, except for the comparison of Finland and Latvia, the gaps for reading scores are larger than the gaps for the science results. Moreover, the results yield more or less the same interpretation as the results for the PISA science test scores. Again, all return effects are significantly positive and by far the largest in magnitude. The characteristics effects are only negative for Estonia, Latvia and Hungary. It is statistically significant only for the first one, indicating that the Estonian students are slightly better endowed with those characteristics which yield higher reading scores. For the other four countries, the characteristics effect is positive indi-

cating that, on average, the Finnish students are slightly better endowed with more favorable characteristics or less endowed with less favorable characteristics. For the reading scores, four of the interaction effects are positive and significantly different from zero, which suggest that the Finnish students have a slight advantage due to the fact that they are better endowed with those characteristics that also yield a higher return in Finland. Only the interaction effect for the country comparison with Bulgaria is significantly negative.

Results for the Decompositions of the Science and Reading score gaps among Eastern European countries

As indicated before, the results from the PISA study show that there is a significant variation in the performance, not only between Finland and Eastern European countries, but also between countries from Eastern Europe, which shared the same educational system for decades. We refer here to Czech Republic and Slovakia as well as Estonia and Latvia. Since each pair of countries also share a common history with respect to their religion, culture and the influence of other countries, we expect them to be more similar than the students in the previous decompositions. Given these considerations, it is interesting to have a look at the gap of each of these two pair of countries that were more common twenty years ago but have developed differently since the early 1990's, in order to explain their test scores gaps at PISA study 2006.

The decomposition results are presented in Tables 1.3 and 1.4. The two country comparisons reveal interesting results. The characteristics effect is highly significant for science and reading. The return effect is also high in magnitude and statistically significant, explaining almost the whole gap between Estonia and Latvia, both in science, as well as in reading. The interaction effect is negative for all four decompositions and highly significant.

1. EDUCATIONAL PERFORMANCE GAPS IN EASTERN EUROPE

Table 1.3: Semiparametric decompositions for science between Eastern European countries

Countries	Δ_c	Δ_r	Δ_{cr}	Δ
EST-LTV	7.08*** (0.96)	40.68*** (1.77)	-5.84*** (1.04)	41.92*** (1.90)
CZE-SLK	19.67*** (1.19)	27.19*** (2.15)	-22.06*** (1.79)	24.80*** (1.96)

Note: The country which has worse performance is always the reference country. Standard errors are in brackets and simulated with 1000 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

Table 1.4: Semiparametric decompositions for reading between Eastern European countries

Countries	Δ_c	Δ_r	Δ_{cr}	Δ
EST-LTV	7.01*** (0.94)	22.20*** (1.84)	-6.62*** (1.07)	22.59*** (1.95)
CZE-SLK	21.84*** (1.37)	24.37*** (2.41)	-28.70*** (2.06)	17.51*** (2.30)

Note: The country which has worse performance is always the reference country. Standard errors are in brackets and simulated with 1000 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

In the case of the Czech Republic and Slovakia, the magnitude of this effect is especially large and works in favor of Slovakian students. Even if the students from the Czech Republic have the advantage of higher returns, they are less endowed with those characteristics that are better rewarded by their school system than by the Slovakian system or more endowed with those characteristics that are better rewarded by the Slovakian school system, as reflected by the negative values of the interaction effects.

Results for the Science and Reading total score gaps at different quantiles

In this part we will look at the PISA test score gaps at different quantiles to understand whether the students performing well or poorly drive the differences in the average test scores.

Figure 1.B.2 of Appendix 1.B displays the distribution of the total gaps at different percentiles for the science test score, showing that the distributions of the gaps

are quite different for various country comparisons. To get a better understanding Table 1.A.3 of Appendix 1.A additionally presents the science test score gap at five quantiles.

For the Finland-Estonia comparison, the gap is relatively low for very small percentiles and then increases to around 30 points at the 5th percentile. Afterwards, it is approximately constant between 30 and 34 points. For the Finland-Czech Republic, Finland-Slovakia as well as Finland-Bulgaria comparisons, the test score gap is decreasing over the whole distribution, indicating that the poorly performing students are the driving forces behind the average gap. If we look at the distribution of the test score gap for Finland-Hungary we see that it is increasing until approximately the 12th percentile and then decreasing up to the 96th percentile. For Finland-Latvia and Estonia-Latvia, the test score gaps are constant over the distribution. For Finland-Romania, it is inverse *u*-shaped and for the Czech Republic-Slovakia comparison, it is decreasing for very small percentiles and increasing afterwards.

Figure 1.B.3 of Appendix 1.B displays the distribution of the total gaps at different percentiles for the reading test score. Table 1.A.4 of Appendix 1.A additionally presents the test score gaps at five quantiles. It shows that the differences in gaps between the two extremes of the distribution (p_5-p_{95}) are higher for the reading than for the science test score. Thus, these results show that there is a higher heterogeneity in students' performance not only between students from different countries at the respective percentile, but also along the same distribution of the reading test scores. Moreover, the gaps in reading are higher than the total gaps in science at the lower part of the distribution.

Figure 1.B.3 also shows that for all country comparisons except Czech Republic-Slovakia, the test score gap in reading is decreasing over the distribution of the test score gaps. This result, which holds for all comparisons between Finland and the seven Eastern European countries, indicates that the large average test score gaps in reading are mainly due to the poorly performing students.

This result gives further insight as to why the Finnish students perform best in the PISA 2006 study. The Finnish school achieves that the poorly performing students perform much better than the poorly performing students of the other countries.

If we look at the comparison between Estonia and Latvia, as well as, the Czech Republic and Slovakia, we find this pattern only for the first pair. In the case of the Czech Republic and Slovakia, the total gap is smallest in the lower part of the distribution and increases steadily afterwards.

1.6 Conclusion

This paper analyzes how much of the average differences in PISA test scores between the seven Eastern European countries and Finland as well as the differences between the countries from Eastern Europe, can be attributed to a small set of important individual and family characteristics.

Moreover, we contribute to the literature by introducing a semiparametric matching procedure to estimate a threefold decomposition. Most important, our procedure relaxes the functional form assumptions of the usual Blinder-Oaxaca decomposition.

Applying this method in decomposing the gaps in PISA test scores provides interesting insights. We provide evidence that only a small part of the gap can be attributed to the fact that the Finnish students are better endowed with more favorable family background characteristics. The main part of the gap still remains after controlling for the individual background. The students from South Eastern Europe are those who have the largest potential outcome increase if they would have more of the unobserved factors like other individual characteristics, institutional aspects of the school system, resources, cultural factors and so forth.

Estonia, the country which adapted its school system to the Finnish school system more than any other country in our sample, has the lowest estimated return effect. Estonia not only performs best out of our Eastern European countries but also among the best of all participating countries.

For science, we find different patterns of the test score distributions. For the Finland-Estonia comparison, the gap is increasing at low percentiles and stays relatively constant afterwards. For the Finland-Latvia comparison, the test score gaps are constant over the distribution, whereas for Finland-Romania it is u-shaped. For the Finland-Czech Republic, Finland-Slovakia as well as Finland-Bulgaria comparisons,

the test score gap is decreasing over the whole distribution and the distribution of the test score gap for the comparison Finland-Hungary is increasing at low percentile and then decreasing. For reading, we find for all comparisons between Finland and the seven Eastern European countries that the test score gap is decreasing over the whole distribution.

All in all and especially for reading, this indicates that the average test score gaps between Finland and the Eastern European countries are mainly due to the fact that the poorly performing students in Finland score much higher than the poorly performing students in the Eastern European countries.

Moreover, our paper exploits the fact that some Eastern European countries had a very similar school system 20 years ago. Estonia and Latvia both belonged to the Soviet Union and share a similar history and the Czech Republic and Slovakia composed Czechoslovakia until 1992. This provides us with a situation similar to a natural experiment. In both cases, the countries started from a very similar point but then developed differently over the past years. For these countries, the differences in individual and family background characteristics are highly significant and explain part of the test score gap in science and reading.

If we look at the distribution of the gap, we find that it is constant over the distribution if we compare the science test score between Estonia and Latvia. It is slightly decreasing if we focus on the reading test score, indicating that the poorly performing Estonian students perform better than the poorly performing Latvia students, whereas the better performing students score closer together in the two countries. Compared to that, we find that the gap is increasing over the distribution for both subjects if we compare the Czech Republic to Slovakia. Hence, this is the only pair of countries where the better performing country scores, on average, higher due to the fact that the best performing students perform much better than the best performing students of the worse performing country and the weak performing students from the better performing country perform only slightly better than the weak performing students from the other country.

Bibliography

- AMMERMÜLLER, A. (2007): “PISA: What makes the Difference? Explaining the Gap in Test Scores between Finland and Germany,” *Empirical Economics*, 33, 263–287.
- AMMERMÜLLER, A., H. HEIJKE, AND L. WÖSSMANN (2005): “Schooling Quality in Eastern Europe: Educational Production during Transition,” *Economics of Education Review*, 24, 579–599.
- BARSKY, R., J. BOUND, K. CHARLES, AND J. LUPTON (2002): “Accounting for the Black-White Wealth Gap: A Nonparametric Approach,” *Journal of the American Statistical Association*, 97, 663–673.
- BEIRNE, J. AND N. F. CAMPOS (2007): “Educational Inputs and Outcomes before the Transition from Communism,” *The Economics of Transition*, 15, 57–76.
- BJØRNSKOV, C. AND N. POTRAFKE (2011): “Politics and Privatization in Central and Eastern Europe: A Panel Data Analysis,” *The Economics of Transition*, 19, 201–230.
- BLINDER, A. (1973): “Wage Discrimination: Reduced Form and Structural Estimates,” *Journal of Human Resources*, 8, 436–455.
- CARD, D. AND J. ROTHSTEIN (2007): “Racial Segregation and the Black-White Test Score Gap,” *Journal of Public Economics*, 91, 2158–2184.
- CERYCH, L. (1997): “Educational Reforms in Central and Eastern Europe: Processes and Outcomes,” *European Journal of Education*, 32, 75–96.
- DEHEJIA, R. H. AND S. WAHBA (1999): “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053–1062.
- DUNCAN, K. C. AND J. SANDY (2007): “Explaining the Performance Gap Between Public and Private School Students,” *Eastern Economic Journal*, 33, 177–191.
- FORTIN, N., T. LEMIEUX, AND S. FIRPO (2010): “Decomposition Methods in Economics,” in *Handbook of Labour Economics*, ed. by O. Ashenfelter and D. Card, Amsterdam: North-Holland.

- FRÖLICH, M. (2004): “Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators,” *Review of Economics and Statistics*, 86, 77–90.
- (2007): “Propensity Score Matching without Conditional Independence Assumption—With an Application to the Gender Wage Gap in the United Kingdom,” *Econometrics Journal*, 10, 359–407.
- FRYER, R. G. AND S. D. LEVITT (2010): “An Empirical Analysis of the Gender Gap in Mathematics,” *American Economic Journal: Applied Economics*, 2, 210–240.
- FUCHS, T. AND L. WÖSSMANN (2007): “What accounts for International Differences in Student Performance? A Re-Examination using PISA Data,” *Empirical Economics*, 32, 433–464.
- HECKMAN, J., J. LALONDE, AND J. SMITH (1999): “The Economics and Econometrics of Active Labor Market Programs,” in *Handbook of Labor Economics*, Vol. 3A, ed. by O. Ashenfelter and D. Card, Elsevier Science, 1865–2097.
- IMBENS, G. (2004): “Nonparametric Estimation of Average Treatment Effects Under Exogeneity,” *Review of Economics and Statistics*, 86, 4–29.
- KITSING, M. (2008): “PISA 2006–Estonian Results,” Technical Report, Ministry of Education and Research, External Evaluation Department, Tartu, Estonia.
- KRIEG, J. M. AND P. STORER (2006): “How Much Do Students Matter? Applying the Oaxaca Decomposition to Explain Determinants of Adequate Yearly Progress,” *Contemporary Economic Policy*, 24, 563–581.
- LECHNER, M. (2010): “A Note on the Common Support Problem in Applied Evaluation Studies,” *Annals of Economics and Statistics*, 91–92, 217–234.
- MARTINS, L. AND P. VEIGA (2010): “Do Inequalities in Parents’ Education Play an Important Role in PISA Students’ Mathematics Achievement Test Score Disparities?” *Economics of Education Review*, 29, 1016–1033.
- MCEWAN, P. AND J. MARSHALL (2004): “Why does academic achievement vary across countries? Evidence from Cuba and Mexico,” *Education Economics*, 12, 205–217.
- MORA, R. (2008): “A Nonparametric Decomposition of the Mexican American Average Wage Gap,” *Journal of Applied Econometrics*, 23, 463–485.

- NIEDERLE, M. AND L. VESTERLUND (2010): “Explaining the Gender Gap in Math Test Scores: The Role of Competition,” *The Journal of Economic Perspectives*, 24, 129–144.
- OAXACA, R. (1973): “Male-Female Wage Differentials in Urban Labor Markets,” *International Economic Review*, 14, 693–709.
- OECD (2007): “PISA 2006 Science Competencies for Tomorrow’s World,” Technical Report, OECD, Paris, France.
- PATACCHINI, E. AND Y. ZENOU (2009): “On the Sources of the Black-White Test Score Gap in Europe,” *Economics Letters*, 102, 49–52.
- PLUG, E. AND W. VIJVERBERG (2003): “Schooling, Family Background, and Adoption: Is it Nature or is it Nurture?” *Journal of Political Economy*, 111, 611–641.
- RADÓ, P. (2001): *Transition in Education*, The Open Society Institute, Institute for Educational Policy, Budapest, Hungary.
- ROSENBAUM, P. R. AND D. B. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.
- SCHNEEWEIS, N. (2011): “Educational Institutions and the Integration of Migrants,” *Journal of Population Economics*, 24, 1281–1308.
- SCHÜTZ, G., H. W. URSPRUNG, AND L. WÖSSMANN (2008): “Education Policy and Equality of Opportunity,” *Kyklos*, 61, 279–308.
- WÖSSMANN, L. (2003): “Schooling Resources, Educational Institutions and Student Performance: The International Evidence,” *Oxford Bulletin of Economics and Statistics*, 65, 117–170.
- (2008): “How Equal are Educational Opportunities? Family Background and Student Achievement in Europe and the United States,” *Zeitschrift für Betriebswirtschaft*, 78, 45–70.
- WÖSSMANN, L., E. LÜDEMANN, G. SCHÜTZ, AND M. R. WEST (2009): *School Accountability, Autonomy and Choice around the World*, Elgar, Cheltenham.

Appendix 1.A Tables

Table 1.A.1: Variables' description

Variable	Min	Max	Description
Test scores			
Reading score	5.67	781.96	mean of five plausible values for reading
Math score	40.61	819.05	mean of five plausible values for math
Science score	93.56	820.52	mean of five plausible values for science
Student Background			
Student's sex	0	1	1 for male
Student's age (in months)	182.04	195.96	Student's age in month
Books Cat.1	0	1	1 if less than 11 books at home
Books Cat.2	0	1	1 if 11-25 books
Books Cat.3	0	1	1 if 26-100 books
Books Cat.4	0	1	1 if 101-200 books
Books Cat.5	0	1	1 if 201-500 books
Books Cat.6	0	1	1 if more than 500 books
<i>Mother's education</i>			
No secondary	0	1	1 if completed at most ISCED 1
Lower secondary	0	1	1 if completed ISCED 2
Upper secondary	0	1	1 if completed ISCED 3A,3B,3C or 4
Tertiary	0	1	1 if completed ISCED 5B or higher
<i>Father's education</i>			
No secondary	0	1	1 if completed at most ISCED 1
Lower secondary	0	1	1 if completed ISCED 2
Upper secondary	0	1	1 if completed ISCED 3A,3B,3C or 4
Tertiary	0	1	1 if completed ISCED 5B or higher

Source: PISA 2006 data, own calculations.

Table 1.A.2: Weighted means and standard deviations

Variable	Finland		Estonia		Czech R.		Hungary	
	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
Reading Score	548.50	75.86	506.68	80.18	484.87	105.76	483.31	90.05
Math Score	550.26	75.15	516.48	75.80	511.14	98.41	491.81	86.84
Science Score	565.56	80.66	533.55	79.67	514.21	94.84	504.71	84.70
Male	0.49	0.50	0.51	0.50	0.56	0.50	0.52	0.50
Age	187.77	3.40	189.66	3.44	190.53	3.43	188.49	3.43
Books Cat. 1	0.05	0.22	0.05	0.21	0.05	0.23	0.06	0.24
Books Cat. 2	0.11	0.31	0.11	0.31	0.10	0.30	0.11	0.31
Books Cat. 3	0.34	0.48	0.29	0.45	0.35	0.48	0.28	0.45
Books Cat. 4	0.24	0.43	0.24	0.43	0.23	0.42	0.21	0.41
Books Cat. 5	0.19	0.39	0.20	0.40	0.17	0.38	0.18	0.39
Books Cat. 6	0.07	0.26	0.12	0.32	0.09	0.29	0.16	0.37
<i>Mother's education</i>								
No Secondary	0.04	0.19	0.00	0.03	0.01	0.08	0.01	0.09
Lower Secondary	0.06	0.24	0.03	0.17	0.03	0.17	0.14	0.35
Upper Secondary	0.33	0.47	0.62	0.49	0.77	0.42	0.57	0.50
Tertiary	0.58	0.49	0.35	0.48	0.20	0.40	0.28	0.45
<i>Father's education</i>								
No Secondary	0.06	0.24	0.00	0.06	0.00	0.05	0.01	0.07
Lower Secondary	0.09	0.28	0.04	0.21	0.02	0.14	0.09	0.29
Upper Secondary	0.42	0.49	0.77	0.42	0.76	0.43	0.68	0.47
Tertiary	0.43	0.50	0.18	0.39	0.22	0.41	0.22	0.42
Number of obs.	4609		4703		5813		4395	

Source: PISA 2006 data, own calculations.

1. EDUCATIONAL PERFORMANCE GAPS IN EASTERN EUROPE

Table 1.A.2 (cont'd): Weighted means and standard deviations

Variable	Latvia		Slovakia		Bulgaria		Romania	
	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
Reading Score	481.09	84.09	467.35	99.67	405.52	110.95	396.02	86.35
Math Score	487.98	77.43	493.07	89.85	415.70	95.60	414.92	79.31
Science Score	491.63	80.02	489.41	89.49	436.63	102.63	418.37	77.68
Male	0.48	0.50	0.51	0.50	0.52	0.50	0.50	0.50
Age	189.61	3.42	188.64	3.38	188.87	3.42	188.91	3.31
Books Cat. 1	0.06	0.24	0.08	0.27	0.20	0.40	0.15	0.36
Books Cat. 2	0.12	0.33	0.12	0.32	0.16	0.37	0.21	0.41
Books Cat. 3	0.31	0.46	0.40	0.49	0.28	0.45	0.33	0.47
Books Cat. 4	0.24	0.43	0.23	0.42	0.16	0.37	0.15	0.36
Books Cat. 5	0.17	0.38	0.12	0.33	0.12	0.33	0.10	0.30
Books Cat. 6	0.10	0.30	0.05	0.22	0.08	0.27	0.06	0.24
<i>Mother's education</i>								
No Secondary	0.00	0.05	0.00	0.04	0.02	0.12	0.03	0.16
Lower Secondary	0.02	0.14	0.04	0.21	0.11	0.31	0.09	0.29
Upper Secondary	0.69	0.46	0.81	0.40	0.59	0.49	0.54	0.50
Tertiary	0.29	0.45	0.15	0.36	0.29	0.45	0.34	0.48
<i>Fathers's education</i>								
No Secondary	0.00	0.06	0.00	0.05	0.01	0.12	0.03	0.16
Lower Secondary	0.03	0.18	0.03	0.16	0.08	0.27	0.07	0.25
Upper Secondary	0.77	0.42	0.82	0.39	0.74	0.44	0.60	0.49
Tertiary	0.20	0.40	0.16	0.36	0.16	0.37	0.30	0.46
Number of obs.	4542		4675		4255		5102	

Source: PISA 2006 data, own calculations.

Table 1.A.3: Test score gaps at different quantiles for science

Quantile	5%	25%	50%	75%	95%
FIN-EST	30.49*** (4.49)	32.36*** (2.50)	33.66*** (2.26)	31.70*** (2.22)	29.65 (3.39)
FIN-CZE	69.56*** (4.38)	66.77*** (2.74)	52.68*** (2.48)	38.98*** (2.56)	26.30*** (2.96)
FIN-HUN	63.59*** (3.53)	65.93*** (2.62)	62.48*** (2.43)	57.81*** (2.11)	51.29*** (3.08)
FIN-LTV	71.33*** (4.25)	75.44*** (2.74)	75.16*** (2.35)	74.50*** (2.35)	70.22*** (2.99)
FIN-SLK	86.25*** (4.26)	83.36*** (2.47)	77.95*** (2.52)	67.51*** (2.87)	61.17*** (2.87)
FIN-BUL	152.93*** (4.92)	151.06*** (2.50)	134.84*** (3.10)	113.48*** (3.00)	84.39*** (3.26)
FIN-ROM	130.17*** (4.04)	149.57*** (2.78)	152.09*** (2.94)	152.09*** (2.57)	144.16*** (3.60)
EST-LTV	40.84*** (4.55)	43.08*** (3.04)	41.50*** (2.39)	42.80*** (2.38)	40.56*** (3.44)
CZE-SLK	16.69*** (4.55)	16.60*** (2.82)	25.27*** (2.57)	28.53*** (3.12)	34.87*** (2.94)

Note: The country which has worse performance is always the reference country. Standard errors are in brackets and simulated with 1000 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

1. EDUCATIONAL PERFORMANCE GAPS IN EASTERN EUROPE

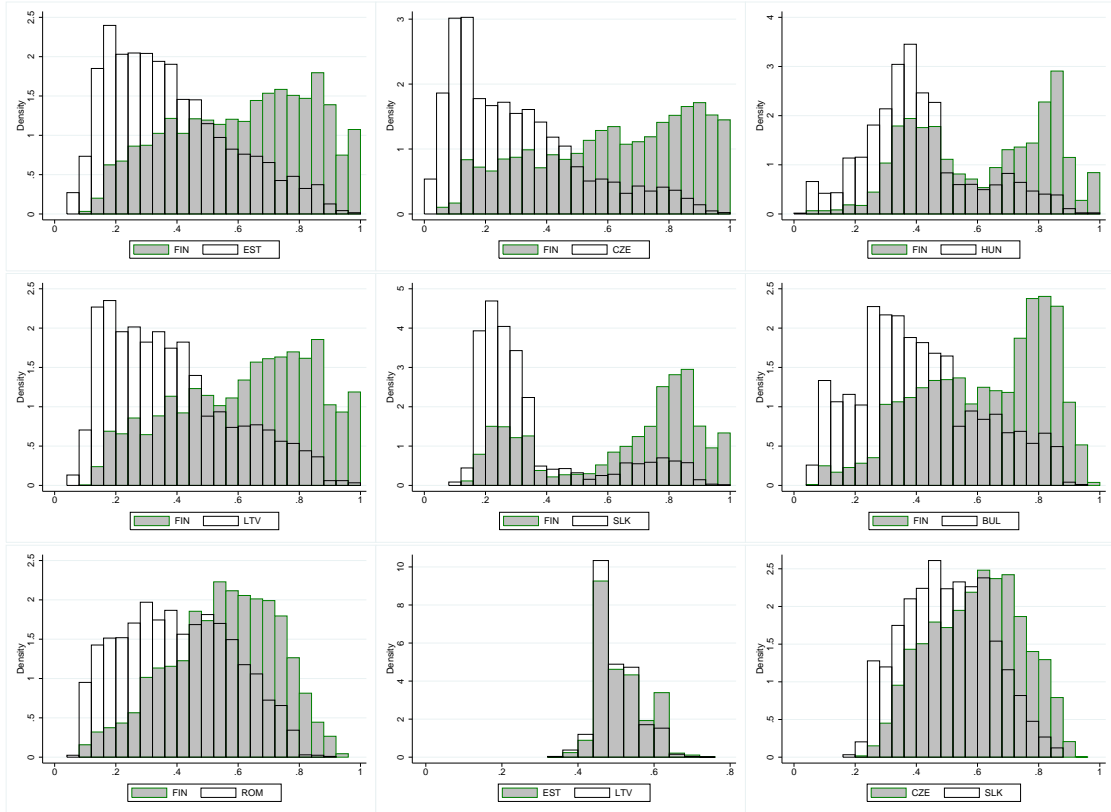
Table 1.A.4: Test score gaps at different quantiles for reading

Quantile	5%	25%	50%	75%	95%
FIN-EST	50.85*** (5.42)	46.79*** (2.66)	44.13*** (2.03)	41.55*** (2.18)	41.63*** (2.76)
FIN-CZE	113.59*** (5.45)	87.63*** (3.16)	60.86*** (2.54)	38.93*** (2.46)	23.12*** (3.06)
FIN-HUN	89.28*** (4.76)	74.54*** (3.02)	60.49*** (2.32)	54.02*** (2.10)	51.67*** (2.85)
FIN-LTV	81.97*** (3.79)	75.51*** (2.80)	65.70*** (2.26)	60.49*** (2.33)	57.44*** (2.84)
FIN-SLK	127.89*** (3.78)	98.27*** (2.68)	77.01*** (2.38)	62.29*** (2.24)	48.78*** (3.66)
FIN-BUL	192.82*** (4.24)	172.32*** (2.94)	142.68*** (3.01)	115.67*** (2.61)	84.53*** (3.64)
FIN-ROM	166.03*** (3.88)	163.58*** (3.00)	152.85*** (2.27)	143.26*** (3.15)	135.89*** (3.67)
EST-LTV	31.12*** (5.70)	28.71*** (3.12)	21.58*** (2.57)	18.94*** (2.52)	15.81*** (2.71)
CZE-SLK	14.30** (5.89)	10.65*** (3.63)	16.16*** (3.07)	23.36*** (2.58)	25.66*** (3.64)

Note: The country which has worse performance is always the reference country. Standard errors are in brackets and simulated with 1000 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

Appendix 1.B Figures

Figure 1.B.1: Histogram estimates of the propensity score distributions



1. EDUCATIONAL PERFORMANCE GAPS IN EASTERN EUROPE

Figure 1.B.2: Quantile test score gaps in science

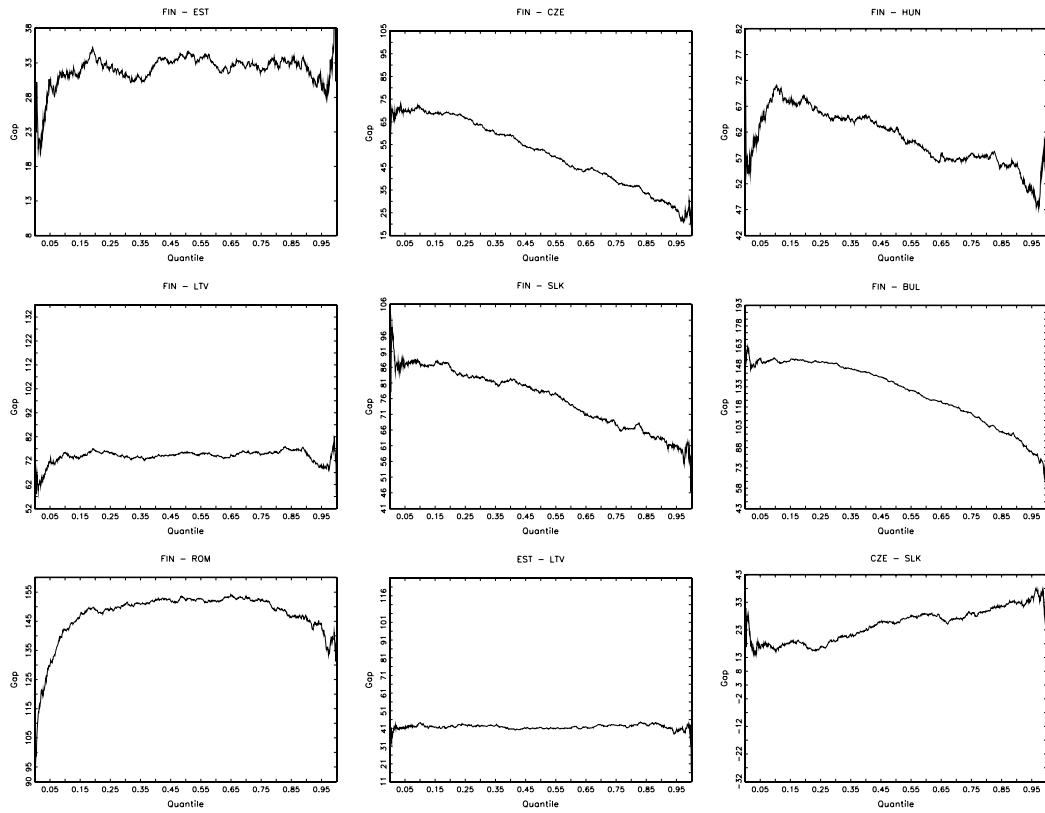
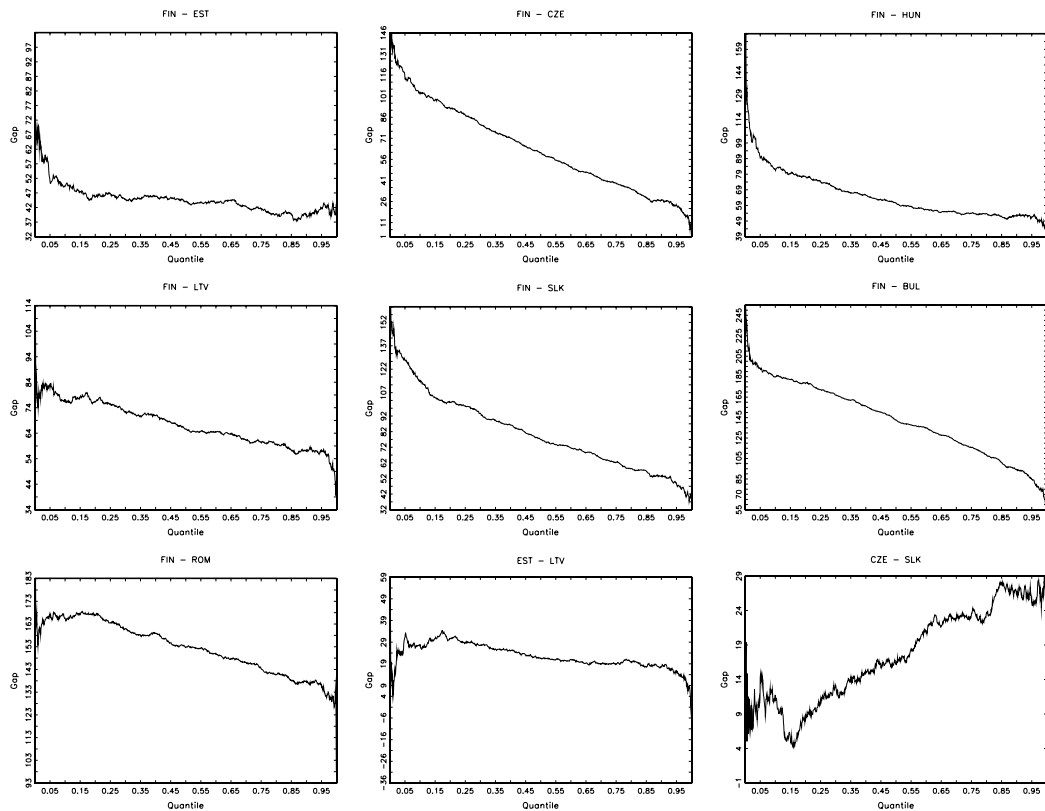


Figure 1.B.3: Quantile test score gaps in reading



CHAPTER 2

Semiparametric Decomposition of the Gender Achievement Gap: An Application to Turkey

2.1 Introduction

Gender inequalities in educational performance have been the subject of much research for many decades. Promoting gender equality in education is an important policy goal especially in developing countries as it is associated with greater equality in employment outcomes, low infant mortality rates, a decrease in the number of early marriages and better investments in education and health of future generations (OECD (2010)). Using the data from the 2006 Programme for International Student Assessment (PISA), this study explores the gender gap in mathematics and science achievement of 15-year-olds in Turkey. We apply a semiparametric Blinder-Oaxaca (BO) decomposition to investigate the gap.

The exploration of gender test score gap is important for the following reasons. First, the recent research on the economic impact of human capital investment underlines the prime importance of educational quality over pure schooling attainment (Hanushek et al. (2009)). Social scientists use international tests of students' performance in cognitive skills such as mathematics and science as a proxy for education quality. Mulligan (1999) and Murnane et al. (2000) show that mathematics test score in high school has a significant effect on annual earnings. Since math, and science skills are highly valued in the labor market, understanding the gender patterns in these subject fields allows us to gain insight into the gender wage gap and differential education and labor market choices across genders. For example, if girls lag behind boys in terms of the accumulation of math skills in childhood and adolescence, they are less likely than boys to choose science, technology and engineering as a field of study at tertiary level, promoting gender inequality in employment opportunities such as the underrepresentation of women in math-intensive fields. Second, using data from the international student achievement tests, empirical growth research documents a significant impact of the quality of education on economic growth (Hanushek and Kimko (2000); Jamison et al. (2007)). Moreover, Hanushek and Hitomi (2008) provide evidence that educational quality leads to longer school attendance in the developing countries. Thus, educational policies aimed to improve quality of education also help meet goals for educational attainment.

The contribution of this paper to the literature is threefold. First, although Blinder-Oaxaca decomposition (Oaxaca (1973); Blinder (1973)) has been widely used to examine discrimination in the labor market, the application of this methodology in the economics of education is quite recent. It has been applied to examine the test score gap between countries (Ammermüller (2007)), schools (private versus public) (Duncan and Sandy (2007); Krieg and Storer (2006)) and ethnic groups (indigenous

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

versus non-indigenous) (Sakellariou (2008); McEwan (2004)). There are only two studies that use the decomposition to analyze the gender test score gap. Sohn (2012) uses an aggregate quantile decomposition to analyze the gender mathematics gap in primary school in the USA while Hille (2011) use a detailed decomposition at mean to study the gender gap in mathematics in French primary school.

Our decomposition method has several advantages over the standard BO decomposition. The semiparametric decomposition relaxes the parametric functional form assumption of the standard BO decomposition. In addition, the standard BO decomposition ignores the common support problem. Ñopo (2008) provides evidence that failure to account for the problem of lack of common support leads to systematically upward-biased estimates of the unexplained part. In the semiparametric decomposition, on the other hand, counterfactual mean is computed only for the common support subpopulation. The rationale behind this empirical strategy ensures that female and male observations that are actually comparable in terms of their observed characteristics are matched. The semiparametric matching method also accounts for arbitrary individual effect heterogeneity (Heckman et al. (1999), Imbens (2004)).

Second, there are a number of studies that examine gender gap in educational attainment in Turkey¹ However, studies on the quality of education, which is measured by achievement on standardized tests, basically analyze the determinants of academic achievement without paying sufficient attention to the causes of the gender test score gap.² To the best of our knowledge, this is the first study that rigorously examines the gender test score gap in Turkey using a semiparametric BO decomposition. Moreover, Turkey is an interesting case to study as it has the largest average gender test score gap in science and one of the smallest gap in mathematics among

¹Tansel (2002) uses data from the 1994 Household Budget Survey to investigate determinants of the gender gap in educational attainment. Utilizing data from the 1988 and 2006 Household Labor Force surveys, Hisarcıklılar et al. (2010) examine how the gender gap in educational attainment changed over an 18-year period during which Turkey launched the educational modernization program. Focusing on undergraduate students in a large public university in Turkey, Dayioğlu and Türüt-Aşık (2007) examine gender gaps in university entrance exam scores and academic performance. Smits and Hosgor (2006) study the impact of family background variables on participation in primary and secondary education of children and point to the importance of mother's education especially in primary participation of girls. Dayioğlu et al. (2009) investigate the effect of sibling composition on the gender gap in school enrollment in urban Turkey.

²Dincer and Uysal (2010) examine determinants of student achievement in science using data from the 2006 PISA. Aypay et al. (2007) aim to answer the same research question by utilizing data from the 1999 Trends in Mathematics and Science Study (TIMMS). Erberber (2010) uses the 2007 TIMMS to investigate factors associated with Turkey's regional differences in science achievement.

OECD countries.³

Third, as science literacy was the subject area assessed in depth in PISA 2006, students were asked about different aspects of how they view science. The PISA 2006 contains questions looked at students' general and personal value of science, their interest and enjoyment of science, plus their self-concept of their own abilities in science and whether they are motivated to use science in the future. Taking advantage of this information, we construct two indexes, namely the index of students' general level of beliefs in their academic abilities in science and the index of motivation in science. We expect that the higher levels on each index, the higher the student's performance in science and math. Although, these indexes are subjective measures of motivation and ability, they allow us to control for potentially endogenous effects at least to some extent.

The semiparametric BO decomposition results can be summarized as follows. The mean test score gap is 15.1 points in favor of girls in science while it is not statistically significant in math. Girls possess more of the characteristics associated with high science scores. School characteristics are the most important observable characteristics in explaining the gap. Our findings suggest that ignoring the common support problem causes the underestimation of the part of the gap attributable to observable characteristics. Moreover, gender test score gap shows heterogeneous pattern across the test score distribution. We find that in science, the gap favoring girls is statistically significant until the top quantile and the largest gap occurs at the 50th percentile. In math, the gap is statistically significant only at the top quantile where boys outperform girls.

This study is organized as follows. The next section reviews studies exploring the gender test score gap and provides background information on the education system in Turkey. Section 2.3 describes the data and variables used in the empirical analysis. Section 2.4 introduces the econometric model and discusses the identification strategy. Section 2.5 presents results while section 2.6 concludes.

³According to the PISA 2006 test results, the mean gender test score gap in science across OECD countries ranges between 11.9 score points in favor of girls in Turkey and 10.06 score points in favor of boys in the UK. In math, boys outscore girls in all countries except Iceland. Turkey with 4.48 score points on the low end while Austria has the highest gap with 22.61 score points.

2.2 Background and Literature

Factors Influencing Gender Test Score Gap: Nature versus Nurture

There has been much interest in examining the link between the structure of brain and gender differences in educational outcomes. The proponents of biological theories argue that gender differences in brain composition, hormone levels and spatial ability produce a gap in achievement. Uncovering anatomical, chemical and functional differences between the brains of men and women, Cahill (2012) provides evidence how these gender-based variations relate to differences in male and female cognition. Davison and Susman (2001) look at the relationship between cognitive ability and hormone levels. They investigate whether higher levels of testosterone are associated with better spatial skills by assessing testosterone levels of boys and girls aged between 9 and 14 years old at three test session every six months. The results show positive relations between spatial scores and testosterone for boys at all three sessions whereas for girls at the third session. Brosnan (2006) points out that the effects of testosterone on spatial performance cannot be generalized at the right tail of the spatial distribution. Kucian et al. (2005) examine whether males and females differ in brain activation and performance patterns when solving number-related tasks. They observe different activation patterns between males and females in tasks that require the use of complex problem solving strategies. Ceci et al. (2009) give an extensive review of studies focusing on biological explanations and underscore that biological evidence provided by research in this domain is inconsistent and sometimes contradictory. In our study, we cannot test hypotheses based on biological explanations due to data limitations.

The other strand of the literature emphasizes sociological and environmental factors as the cause of the gap. One of the most important sociological arguments attributes the gender gap in educational achievement to stereotypical thinking such as believing that girls just cannot do math, language is for girls and math is for boys. Investigating geographic differences in gender test score gap in the U.S., Pope and Sydnor (2010) find that there is a large and statistically significant variation in the gender ratios of 8th graders scoring at the top percentiles of the National Assessment Educational Process (NAEP) across states and census divisions.⁴ They also find that in areas where men and women are viewed as more equal, gender disparities are smaller in both stereotypically male-dominated tests of math and science and

⁴The NAEP is a series of standardized tests in subjects such as math, science and reading. The test is taken by a sample of public school students in grades 4, 8, and 12 throughout the United States.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

stereotypically female-dominated tests of reading. Fryer and Levitt (2010) show that parents have lower math expectations for their daughters. They test whether these parental gender stereotypes can explain why girls fall behind in math. To capture the parental expectation effect, they also include an indicator variable for having a mother who works in math-related professions in their regression. The findings indicate that parental expectations do not have a significant impact on the gender math gap. However, Tiedemann (2000) and Jacobs and Eccles (1992) provide evidence that parents' gender stereotyped beliefs about their children's competence in math (i.e. math is a male domain) influence children's self-perceptions of ability in math and hence their math achievement. Moreover, Carrell et al. (2009) examine the impact of a teacher's gender on students' performance in math and science classes. Having a female math or science teacher has a significant effect on female students' performance in math and science while teacher's gender has little impact on male students. The effect is more pronounced for female students with very strong math skills.

Guiso et al. (2008) examine gender test score gaps in mathematics and reading across 40 countries using data from the 2003 PISA. They find that boys, on average, outperform girls by 2 percent in mathematics. Moreover, the gap virtually disappears in more gender-equal countries such as Norway and Sweden. The gender gap is reversed in reading. Girls, on average, score 6.6 percent higher than boys. However, more gender-equal cultures are associated with an increase in the reading gap in favor of girls. For example, girls' reading score average is 25.1 higher than that of boys in Turkey while in Iceland it is 61 higher.⁵ Therefore, their findings point to the importance of culture in explaining the gender test score gap.

There exists a substantial body of research on the importance of family background variables such as parental education, family income and educational resources available at home on student performance (Lara-Cinisomo et al. (2004); Carneiro and Heckman (2003); Hanushek (2003); Wößmann (2003)). In the literature, there is a long-lasting debate over the relative importance of family background and school-related factors on students' achievement (Nonoyama-Tarumi and Willms (2010)). After taking account of family background factors, Fuchs and Wößmann (2007) and Wößmann et al. (2009) examine the impact of school-related factors on student achievement across countries using the PISA data sets. They show that both family

⁵They use several measures for the gender equality of a country such as The World Economic Forum's Gender Gap Index (GGI). Larger values of GGI correspond to a better average position of women in society. Turkey with 0.59 GGI score is on the low end while Iceland with 0.78 is on the high end.

socioeconomic status variables and school institutional factors such as whether the school is public or private and school academic selectivity have significant impacts on student achievement. However, evidence on school resources such as class size and student-teacher ratio is mixed. Hanushek (2003) points to inconsistency of the estimated effect of school resources in the literature.

In sum, it is difficult to address the question of nature versus nurture related to the gender test score gap due to interactions between biological and environmental factors.

Secondary Education in Turkey

Formal education in Turkey consists of the institutions of preschool education, primary education, secondary education and higher education. Preschool education is voluntary. Eight years of primary education, which used to be five years until 1997, is compulsory for all Turkish citizens and is free of charge in public schools. There are also private schools under state control. Students who completed the compulsory level successfully could proceed to secondary education where most of the schools are public. Private secondary schools make up only 8,2 percent of the secondary education system in Turkey (Eurydice (2010)). The duration of the secondary education is minimum four years. Some schools in which the medium of instruction is a foreign language in most of the courses, have a foreign language preparation grade. Thus, in these schools the secondary education lasts 5 years.

The public secondary education consists of two categories: general secondary education and vocational-technical secondary education. There are many different types of institutions in both categories.⁶ Vocational and technical education institutions aim not only prepare students for tertiary education but also to educate them as manpower for business and professional branches.

There are schools such as Anatolian high schools and science high schools that admit students with entrance examination. Students are accepted to those schools according to secondary education placement score (Eurydice (2010)). This score is determined by the entrance exam and competence grades at the end of the 6th, 7th, and 8th class. The centralized nationwide entrance exam consists of multiple choice questions focusing on the curriculum. Year end competence grade is the mean of the two semesters' end averages of all courses.

⁶For further information on these institutions, see <http://oogm.meb.gov.tr>.

2.3 Data and Descriptive Statistics

In the empirical analysis, we use the 2006 PISA survey which is a standardized achievement test in reading, mathematics and science among 15-year-olds enrolled in grades 7 or above.⁷ The implementation of PISA is coordinated by the Organization for Economic Cooperation and Development (OECD) at three-year intervals. The first PISA survey took place in 2000. Every period of assessment focuses on one particular subject. In the 2006 PISA, the focus is on science. Thus, the survey consists of questions that assess students' general and personal value of science, their interest and enjoyment of science and their self-beliefs as science learners (OECD (2007)). In addition to test scores in these three subjects, the 2006 PISA data set provides information on student, family and institutional factors that could help to explain differences in performance.

PISA 2006 implemented a two-stage stratified sample design. The first stage sampling units comprised individual schools having 15-year-old students. In the second stage, 35 students were randomly drawn with equal probability within each school. Final student weights were constructed to account for the probabilities of selection for individual nonresponse, or for errors in estimating the size of the school or the number of 15-year-olds in the school at the time of sampling. In our analysis, we used the sample weights provided in the PISA data set.⁸

Figures 2.B.1 and 2.B.2 of Appendix 2.B present the mean gender test score gaps in science and mathematics for all OECD countries. The figures indicate that the gender gaps in cognitive skills show a similar trend. In mathematics, boys tend to outperform girls in all the countries except for Iceland. Gender differences are more pronounced in mathematics than in science. Only in six OECD countries (the United Kingdom, Luxembourg, Denmark, the Netherlands, Mexico, and Switzerland), boys on average have significantly higher science achievement than girls and the opposite is true in two OECD countries (Turkey and Greece). In addition, across OECD countries Turkey has the largest average gender test score gap in science while it has one of the smallest gap in mathematics across OECD countries.⁹

⁷Turkey participated in PISA in 2003, 2006 and 2009. PISA 2006 started to provide a parents questionnaire which is an important resource for the socioeconomic backgrounds of students. However, this questionnaire was optional and therefore not carried out in all the participating countries. As Turkey did not take part in the parents questionnaire in 2009. Our analysis is based only on the 2006 PISA survey.

⁸For detailed information on the technical characteristics of the 2006 PISA, see OECD (2009).

⁹In Turkey, girls, on average, outscore boys in science by 12 score points. In mathematics, however, boys outscore girls by 6 score points.

PISA is performed on a representative sample of the national population of 15-year-olds enrolled in school. However, in Turkey, most of the 15-year-old students are not subject to compulsory school attendance and 40-45 percent of 15-year-old children are not in school (Blancy and Sasmaz (2011)). Therefore, when evaluating the PISA results, one should take into account the fact that the selection into enrollment may not be random. The majority of children who are not enrolled in school are more likely to have lower academic achievement levels than those enrolled in school.

In PISA 2006, there were thirteen test booklets, each of which contained a slightly different subset of items. As each student was administered different items, the classic test scores, such as the percent of correct answers a student gets on a test, are not accurate measures of student performance. Thus, PISA 2006 used item response theory (IRT) to summarize the performance of a sample of students in a subject area with a simple scale or series of scales.¹⁰ For each test and each student, PISA 2006 reported five plausible values to present students' achievement. We use the average of those five plausible mathematics (science) values as a measure of the student's mathematics (science) performance in PISA 2006.

The variables used to analyze the gender test score gap can be classified into the following three categories.

Student Characteristics

We control for a student's grade as it may have something to do with students' cognitive development. In order to capture a student's attitudes regarding the importance of studying math, we use students' responses to the following question: 'In general, how important do you think it is for you to do well in mathematics?' A four-point scale with the response categories recoded as 'very important' (=4); 'important' (=3); 'little important' (=2); and 'not important at all' (=1) is used. The 2006 PISA data set contains information on attitudinal measures consisting of self-efficacy, self-concept, interest in science, enjoyment of science, instrumental motivation to learn science and career intentions (OECD (2009)).

We use two indexes to measure students' general level of belief in their academic abilities in science. The index of self-concept measures how students felt about their academic abilities in science. Table 2.A.1 of Appendix 2.A shows six questions in the 2006 PISA designed to measure how good students felt they were at science. To assess self-efficacy in science, students were asked about their level of confidence in

¹⁰For more information on item response theory scaling methodology, see OECD (2009).

tackling specific scientific tasks. The notion of self-efficacy differs from self-concept as it includes not only a student's confidence in their ability to do science but also a student's belief in their ability to overcome difficulties when attempting scientific tasks (Marshall et al. (2008)). The list of tasks presented to students is given in Table 2.A.1. We create the index of beliefs in own abilities in science by simply adding up indexes of self-efficacy and self-concept. The higher the index is, the more confident a student is about her/his academic abilities in science. To control for students' interest, enjoyment, and motivation with respect to science, we utilize three indexes. The index of general interest in science is based on a series of questions designed to gauge their interest in learning about science topics. The list of topics is shown in Table 2.A.2 of Appendix 2.A. The index of enjoyment of science measures how much students enjoy learning science topics and acquiring new knowledge in science. Students were asked questions about the usefulness of science for them and their future careers. The index of importance of learning science for future career is created by using students' responses to those questions that are listed in Table 2.A.2. The index of motivation in science is the simple sum of these three indexes. The higher the index is, the more motivated a student is to do well in science. In PISA 2006 students were also asked whether they think that they will have a science-related career when they are about 30 years old. We create an indicator variable that takes the value of one if the student expected to have a science-related career at age 30 and zero otherwise. We expect that students expecting to pursue a science-related career have higher science/math achievement than those who are not.

Family Background Characteristics

The 2006 PISA provides researchers with a number of variables that summarize socioeconomic status of family. We use parents' educational attainment, occupational status and family income as measures of family socioeconomic status. Parental education is measured by the highest level completed and classified into 3 categories: i) at most primary education; ii) secondary education; and iii) tertiary education. We utilize the index of the highest parental occupational status which is based on the International Socio-Economic Index of Occupational Status (ISEI). The value of the index ranges from 16 to 90. The higher values of the index are associated with occupations that have higher returns to education.¹¹ We also include two binary variables indicating whether mother/father has a science-related career.

We use the number of books at home as an indicator of cultural capital. We create

¹¹For detailed information on the construction of ISEI index, see Gazeboom et al. (1992).

four dummy variables based on the following categories: i) 0 to 10 books; ii) 11 to 25 books; iii) 26 to 100 books; and iv) more than 100 books. We also control for the index of home education resources. This index is derived from the availability of various household items at home such as a study room, technical reference books, a computer that students can use for schoolwork and educational software.¹²

School Characteristics

We create an indicator variable taking the value of one if the school is in a rural area which is defined as a geographical unit with less than 15,000 inhabitants. We also create the following seven regional dummies as a set of controls for school location: Marmara region, Central Anatolian region, Aegean region, Mediterranean region, Blacksea region, Eastern Anatolian region and Southeastern Anatolian region. We divide schools into three types, namely general high schools, Anatolian high schools, and vocational high schools.¹³ To control for possible differences between public and private schools, we construct a variable indicating whether the school is public or private. Examining the effects and mechanisms of gender peer effects in elementary, middle, and high schools, Lavy and Schlosser (2011) find that an increase in the proportion of girls is associated with improvement in boys and girls' cognitive outcomes. To capture the potential favorable interaction between genders, we use sex ratio in the school that measures the proportion of girls enrolled at school. The variables that account for school resources are average class size, and the index of the quality of the school's educational resources. The computation of the index is based on the school principal's perceptions on potential factors hindering instruction at school. For example, the school principals were asked whether their schools' capacity to provide instruction is hindered by shortage of science laboratory equipment or computers for instruction.¹⁴

¹²See PISA 2006 Technical Report (OECD (2009), p. 316) for the construction of the index.

¹³Schools are categorized into the following eight groups in the data set: primary schools, general high schools, Anatolian high schools, high schools with intensive foreign language teaching, science high schools, vocational high schools, Anatolian vocational high schools, secondary and vocational high schools (also called multi-program high schools). It is important to note that when we construct school type indicator variables, we group Anatolian vocational high schools and high schools with intensive foreign language teaching as Anatolian high schools. As there is only one science high school in the data set, it is treated as an Anatolian high school. Moreover, secondary and vocational high schools are classified as vocational high schools. Primary schools are classified as general high schools.

¹⁴See PISA 2006 Technical Report (OECD (2009), p. 340) for detailed information on the construction of the index.

Descriptive Statistics

Table 2.A.3 of Appendix 2.A shows summary statistics for the final sample by gender. The final sample excludes cases with missing values on the variables. In mathematics, males have significantly higher average achievement than females. There is also a gender difference in perceived importance of doing well in Mathematics. Interestingly, females, on average, attribute more importance to mathematics. The mean test scores in science indicate a significant gender gap in favor of girls. However, there is no significant gender difference on the index of motivation in science. 27% of females expect to have a science-related career at age 30, compared to 25% of males. 94% of the students are enrolled in the 9th grade or above. Students in the 7th or the 8th grades who did not complete their compulsory education make up only 4% of the sample.¹⁵

The distribution of the educational attainment of parents does not differ considerably between male and female samples. However, there is a significant difference between the average scores of males and females on the index of the highest parental occupational status. 5% of fathers and 2% of mothers have a science-related career. The average annual family income does not vary significantly between male and female samples. 91% of students are from families whose annual income is less than the median annual income (24.000 TL) in Turkey. The index of home education resources does not show a significant gender difference whereas females seem to have more books at home than males.

98% of students attend public schools. Almost half of the sample has a classroom size of at least 30 students. On average, females attend schools with a more balanced sex ratio than males do. Females are overrepresented among students who are enrolled in general high school and Anatolian high school and underrepresented among those who are enrolled in vocational high school.

The mean value of the index of the quality of the school's educational resources is -0.81. The negative value provides evidence that instruction in schools, on average is hindered by a lack of adequate educational resources. The index does not differ significantly between male and female samples.¹⁶

¹⁵In Turkey, the compulsory education ends after eighth grade. Therefore, our sample largely consists of students who are not subject to compulsory school attendance.

¹⁶It is important to note that the construction of the index relies on the judgment of school principals rather than on external observations or the views of students and teachers. Principals may not provide objective measures of the condition of physical infrastructure.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

One drawback of the PISA assessments is that they are based on the performance of 15-year-olds that are enrolled in formal education. Therefore, differential drop-out rates across genders may considerably influence the results. In the specific case of Turkey, there are significant gender differences in enrollment across regions. Akkoyunlu-Wigley and Wigley (2008) point out that females residing in rural areas and eastern provinces are less likely to attend school than their male counterparts.¹⁷ Consistent with the evidence provided in the literature, in our sample females are underrepresented among schools that are in rural areas and among those that are in Southeastern Anatolian and Blacksea regions. Given lower female enrollment rate, the sample of girls is likely to be positively selected sample of 15-year-old girls, which causes any gender gap favorable to boys to understate. One should take into account the possible selection bias when interpreting the results.

¹⁷Table IV in Akkoyunlu-Wigley and Wigley (2008) shows that in Turkey, the gender gap in secondary education enrollment is 6.1% in 2003, ranging from 13.3% in Southeastern Anatolian region to 0.4% in Marmara region.

2.4 Econometric Model

As an alternative to the standard BO decomposition¹⁸, we apply propensity score matching method to decompose the gender test score gap into the composition effect and the return effect. This method estimates the counterfactual mean only for the individuals who are on the common support. Decomposing the gender wage gap among college graduates in the UK, Frölich (2007) is the first to use such a matching procedure outside the treatment evaluation literature. Botezat and Seiberlich (2013) extend this procedure to estimate the threefold Blinder-Oaxaca decomposition semiparametrically and analyze the PISA test score gap between several Eastern European countries and Finland.

To obtain the propensity score, we estimate the probability that an individual is female ($D = 1$) by a logit regression, i.e. $F(x'\beta) = \Pr[D = 1|X = x] = p(x)$,

¹⁸In the standard BO decomposition the test score production function is assumed to be linear for both genders, i.e. $Y_1 = X\beta_1 + \varepsilon_1$ for females and $Y_0 = X\beta_0 + \varepsilon_0$ for males. Under the zero conditional mean assumption $E[\varepsilon_1|X] = E[\varepsilon_0|X] = 0$. Let D be a dummy variable indicating whether the student is female ($D=1$) or not ($D=0$). After taking the expectations over X , the overall mean gender test score gap Δ can be written as follows:

$$\begin{aligned}\Delta &= E[Y_1|D = 1] - E[Y_0|D = 0] \\ &= (E[X|D = 1]\beta_1 + E[\varepsilon_1|D = 1]) - (E[X|D = 0]\beta_0 + E[\varepsilon_0|D = 0])\end{aligned}$$

where $E[\varepsilon_1|D = 1] = E[\varepsilon_0|D = 0] = 0$. After adding and subtracting the counterfactual test score for females, $E[X|D = 1]\beta_0$, which asks what would girls' mean test score be if they had the same returns to educational inputs as boys, the gender test score gap becomes:

$$\Delta = (E[X|D = 1] - E[X|D = 0])\beta_0 + E[X|D = 1](\beta_1 - \beta_0)$$

where the first term is called the composition effect which can be attributed to differences in average characteristics between females and males. The second term is due to differences in average returns to those characteristics and called the return effect.

The estimated gender test score gap is obtained by replacing the expected values of the covariates by the sample averages and the coefficients by their OLS estimates.

The detailed decomposition can be written as follows:

$$\hat{\Delta} = \underbrace{(\bar{X}_1 - \bar{X}_0)'\hat{\beta}_0}_{\hat{\Delta}_c} + \underbrace{\bar{X}_1'(\hat{\beta}_1 - \hat{\beta}_0)}_{\hat{\Delta}_r} = \underbrace{\sum_{k=1}^K (\bar{X}_{1k} - \bar{X}_{0k})\hat{\beta}_{0k}}_{\hat{\Delta}_c} + \underbrace{(\hat{\beta}_{10} - \hat{\beta}_{00}) + \sum_{k=1}^K \bar{X}_{1k}(\hat{\beta}_{1k} - \hat{\beta}_{0k})}_{\hat{\Delta}_r}$$

where K is the number of regressors without constant, $\hat{\beta}_1$ is the vector of coefficients for females while $\hat{\beta}_{1k}$ is the coefficient of explanatory variable k for females. The same distinction applies between $\hat{\beta}_0$ and $\hat{\beta}_{0k}$ for males. $\bar{X}_{1k} = \frac{1}{n_1} \sum_i \mathbb{1}\{D_i = 1\}X_{ik}$ and $\bar{X}_{0k} = \frac{1}{n_0} \sum_i \mathbb{1}\{D_i = 0\}X_{ik}$, $\bar{X}_1 = (1 \ \bar{X}_{11} \ \bar{X}_{12} \ \cdots \ \bar{X}_{1K})'$, $\bar{X}_0 = (1 \ \bar{X}_{01} \ \bar{X}_{02} \ \cdots \ \bar{X}_{0K})'$, n_1 is the number of females and n_0 the number of males in the samples.

where $F(x'\beta)$ represents the cumulative logistic distribution. Next, we estimate the density of this propensity score using a Kernel estimator. Let $f_1(p)$, $f_0(p)$ be the distributions of the propensity score for females ($D = 1$) and males ($D = 0$) respectively.

The common support is evaluated by comparing the distributions (histograms) of the estimated propensity scores by the treatment variable as suggested in Lechner (2010). Figures 2.B.3 - 2.B.5 of Appendix 2.B show the histograms for our different specifications. If we control for students, family and school characteristics (Figure 2.B.5) there are some females with very high propensity scores, whereas we do not have male students in our sample with equally high propensity scores. Thus, the histogram indicates overlap problems and we do the following common support correction.

We define the common support as $\{S : \hat{p}_i \in [\hat{p}^{\min_M}, \hat{p}^{\max_M}]\}$, i.e. all observations with an estimated propensity score that is smaller than the maximum propensity score of males (\hat{p}^{\max_M}) and larger than the minimum estimated propensity score of males (\hat{p}^{\min_M}) belong to the common support subpopulation. Let $f_1^S(p)$ and $f_0^S(p)$ denote the distributions of the propensity score $P = P(X)$ for this common support subpopulation S for females ($D = 1$) and males ($D = 0$) respectively.¹⁹

The gender test score gap for the common support subpopulation can then be written as follows:

$$\begin{aligned} \Delta_S &= E_S[Y^1|D = 1] - E_S[Y^0|D = 0] \\ &= \int_S E_1[Y|P(X) = p]f_1^S(p) dp - \int_S E_0[Y|P(X) = p]f_0^S(p) dp, \end{aligned} \quad (2.1)$$

where Y^0 and Y^1 denote the potential outcomes, $E_1[Y|P(X) = p] = E[Y|P(X) = p, D = 1]$ and $E_0[Y|P(X) = p] = E[Y|P(X) = p, D = 0]$.

Frölich (2007) shows that the counterfactual mean is identified as follows:

$$E_S[Y^0|D = 1] = \int_S E_0[Y|P(X) = p]f_1^S(p) dp \quad (2.2)$$

The counterfactual represents the expected test score that females ($D = 1$) would

¹⁹ $f_d^S(p) = \frac{f_d(p)}{\mu_{S|D=d}}$ is scaled such that the integral integrates to one, where $\mu_{S|D=d}$ is the empirical probability of being on the common support conditional on having gender d .

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

have if they had the same returns to educational inputs as males ($D = 0$).

We estimate the expected outcome for $P(X) = p$ by the ridge regression proposed by Seifert and Gasser (1996). This estimator is a convex combination of the local constant and local linear estimators. Several Monte Carlo studies show that this estimator has a better performance than other matching estimators (see Frölich (2004), Busso et al. (2009)). The ridge regression takes the following form:

$$\hat{E}[Y|P(X) = p, D = 0] = (1 - \bar{R}) \frac{T_0}{S_0} + \bar{R} \left(\frac{T_0}{S_0} + \frac{T_1(p - \bar{p})}{S_2} \right) \quad (2.3)$$

where $\bar{R} = \frac{S_2}{S_2 + R}$, $\bar{p} = \sum_i^{n_0} \frac{K\left(\frac{p_i^0 - p}{h}\right) p_i^0}{K\left(\frac{p_i^0 - p}{h}\right)}$, $S_j = \sum_i^{n_0} K\left(\frac{p_i^0 - p}{h}\right) (p_i^0 - \bar{p})^j$ and $T_j = \sum_i^{n_0} K\left(\frac{p_i^0 - p}{h}\right) (p_i^0 - \bar{p})^j Y_i^0$. Thereby K is the kernel function, h the bandwidth, n_0 the number of observations and p^0 the propensity score of those from group 0. \bar{p} is chosen such that $S_1 = 0$ and R is the ridge parameter. Seifert and Gasser (2000) develop a rule of thumb for choosing R , according to their rule of thumb the ridge parameter for local linear regressions is $R = rh|p - \bar{p}|$. Thus, this ridge parameter depends on the point of evaluation p the bandwidth h and $r = \frac{\max_u(K(u))}{4 \int K^2(u) du}$, e.g. as we use a Gaussian Kernel, $r = 0.3535$.

The bandwidths are selected by leave-one-out cross-validation and are chosen to minimize the least-squares criterion: $h^* = \arg \min_{h \in H} \sum_{j \in I_0} (Y_j - \hat{E}_{-j}[Y_j|D_j = 0, P_j(X) = p_j])^2$, where $\hat{E}_{-j}[Y_j|D_j = 0, P_j(X) = p_j]$ is the out of sample predicted outcome for observation j that is obtained from the data sample without observation j . Following Frölich (2004) we choose as bandwidth search grid $0.01\sqrt{1.2^{g-2}}$ for $g = 1, \dots, 59$ and ∞ .

After adding and subtracting the counterfactual mean in (2), we can decompose the gender test score gap for the common support subpopulation in (1), into two parts:

$$\begin{aligned} \Delta_S &= \underbrace{\int_S E_0[Y|P(X) = p] [f_1^S(p) - f_0^S(p)] dp}_{\Delta_c} \\ &+ \underbrace{\int_S [E_1[Y|P(X) = p] - E_0[Y|P(X) = p]] f_1^S(p) dp}_{\Delta_r} \end{aligned} \quad (2.4)$$

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

The first term can be attributed to differences in the distribution of propensity scores between females and males and is called the composition effect (Δ_c). It would vanish if females had the same characteristics as males. The second term is due to differences in returns to these characteristics and thus called the return effect (Δ_r). It would vanish if females had the same returns to educational inputs as males.

The identification of the composition and return effects crucially relies on two assumptions. The first one is the conditional independence assumption (CIA) which guarantees that conditional on confounding variables the potential outcomes are stochastically independent of the treatment: $Y_i^0 \perp D_i | P_i(X)$, where $P_i(X)$ denotes the propensity score based on the confounding variables of individual i . The CIA requires that all confounding factors associated with the potential outcomes as well as the treatment status are observed.

To justify the CIA, we control for a rich set of covariates available in the PISA data set, including student, family background, and school characteristics. Although we do not have information on innate ability, the data set allows us to construct two indexes that provide subjective measures of ability and motivation in science. Those indexes at least partially account for potentially endogenous effects. Moreover, Fortin et al. (2010) point out that the aggregate decomposition would even be valid in the presence of the correlation between unobserved and observed characteristics under the condition that the correlation is the same for both genders.

The second assumption is the overlap assumption. It requires that the probability of being female is smaller than one, i.e. $\Pr(D = 1|X) < 1$. This type of overlap assumption is standard in the literature (e.g. Rosenbaum and Rubin (1983), Heckman et al. (1997), Hahn (1998), Wooldridge (2002), Imbens (2004)).²⁰ To guarantee that $\text{supp}(X|D = 1) \subseteq \text{supp}(X|D = 0)$ we restrict the estimation of the composition and return effect to the common support subpopulation.

To account for the observations in the sample that cannot be matched, we follow

²⁰There is a stronger version of the overlap assumption called strict overlap (e.g. Robins et al. (1994), Abadie and Imbens (2006), Crump et al. (2009)). Strict overlap requires that the probability of being female is strictly smaller than $1 - \xi$ for some $\xi > 0$. Khan and Tamer (2010) point out that a comparable assumption to the strict overlap assumption is needed for \sqrt{N} -convergence of some semiparametric estimators. Busso et al. (2009) provide further evidence on the importance of (strict) overlap assumption.

$\tilde{\text{Nopo}}$ (2008) and decompose the whole gap Δ into three parts: $\Delta = \Delta_1 + \Delta_c + \Delta_r$.²¹ In addition to the composition and return effects, we have Δ_1 which represents the part of the test score gap that can be explained by differences between two groups of females: those who can be matched with males and those who remain out of the common support, weighted by the empirical fraction of females who are out of the common support. A positive value of Δ_1 indicates that female students, who are out of the common support, perform better than their counterparts, who are in the common support.

To analyze the heterogeneous pattern of the gender test score gap across the test score distribution we additionally look at the gaps of the common support subpopulation at different quantiles:

$$\Delta_S^\tau = F_{y^1|D=1,S}^{-1}(\tau) - F_{y^0|D=0,S}^{-1}(\tau) \quad (2.5)$$

where $F_{y^1|D=1,S}^{-1}(\tau)$ ($F_{y^0|D=0,S}^{-1}(\tau)$) is the τ -quantile of the test score distribution among females (males) who are on the common support.

2.5 Results

The Standard BO Decomposition

Gender Test Score Gap in Mathematics

Table 2.A.4 of Appendix 2.A presents OLS estimates of the gender test score gap in math. As one moves to right in the table, the number of covariates steadily

²¹The gap can be written explicitly as follows:

$$\begin{aligned} \Delta = & \underbrace{\mu_{\bar{S}|D=1} \left[\int_{\bar{S}} E_1[Y|P(X)=p] f_1^{\bar{S}}(p) \, dp - \int_S E_1[Y|P(X)=p] f_1^S(p) \, dp \right]}_{\Delta_1} \\ & + \left[\int_S E_1[Y|P(X)=p] f_1^S(p) \, dp - \int_S E_0[Y|P(X)=p] f_0^S(p) \, dp \right] \\ & + \underbrace{\mu_{\bar{S}|D=0} \left[\int_{\bar{S}} E_0[Y|P(X)=p] f_0^{\bar{S}}(p) \, dp - \int_S E_0[Y|P(X)=p] f_0^S(p) \, dp \right]}_{\Delta_0} \end{aligned}$$

where \bar{S} denotes the non-common support and $\mu_{\bar{S}|D=d}$ the empirical probability of being unmatched conditional on having gender d . As before, the second summand can be decomposed into Δ_c and Δ_r . Due to our definition of the common support $\mu_{\bar{S}|D=0} = 0$.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

increases. The raw gap is 8.6 score points in favor of males. Column 2 adds controls for student's characteristics. The coefficients of grade indicator variables are positive and statistically significant. Students in the 9th, 10th and 11th grades have significantly higher test scores compared to those who are in the 7th or 8th grade. The higher levels on the index of beliefs in own abilities in science is associated with higher achievement in mathematics. Students who expected to pursue a scientific career and placed a higher value on mathematics score significantly higher on the test. These results are robust across all specifications. The index of motivation in science has an unexpected negative sign in column 2. However, the coefficient of the index becomes statistically insignificant when we control for family background and school characteristics. The test score gap remains negative and significant when the controls for family background variables are included in column 3. Almost all these controls enter with the expected sign. The number of books is significantly positively associated with test score on math. Parents' education and occupational status and family income are important predictors of math test score. Students from families with higher socioeconomic status score better. The coefficient on the index of home educational resources is positive and statistically significant. Students with mothers who have a science-related career score better. However, it does not matter whether the father is in science-related occupation.

The final specification in Table 2.A.4 also adds a set of covariates capturing school characteristics. As expected, students who attend schools in an rural area and those with low quality of educational resources score worse. A higher percentage of girls is associated with a higher math test score. Compared to students attending schools in Marmara region, those attending schools in Aegean, Mediterranean, and Blacksea regions score better while those attending in Eastern Anatolian and Southeastern regions score worse. School type matters. Students from Anatolian high school score better but, those from vocational high school perform worse than those from general high school. It is worth noting that in Table 2.A.4, the gender test score gap becomes larger than the raw gap when the number of covariates increases. The estimates suggest that controlling for other factors, females score worse than males in math.

Because females and males may not be equally responsive to changes to covariates, we perform the analysis separately for males and females. The results presented in columns 1 and 2 of Table 2.A.6 of Appendix 2.A suggest that the responsiveness to motivation and ability indexes varies across genders. The effect of motivation index on math achievement is statistically significant only for males while the effect

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

of the ability index is statistically significant only for females. Father's education is more important for daughters. Moreover, coefficients on school characteristics show important differences between males and females.

Table 2.1 below presents the results from the standard BO decomposition for a reduced set of three factors: student characteristics, family characteristics and school characteristics. The composition effect is positive and statistically significant at the 1% level, implying that gender differences in observable characteristics predict an advantage for females over males in the average mathematics scores. School characteristics are by far the most important explanatory factors contributing to the composition effect. Students and family characteristics account for 11.3 percent and 8.7 percent of the composite effect respectively.

Table 2.1: The standard BO decomposition of the gender test score gap in math

		Student Characteristics	Family Characteristics	School Characteristics	Constant
Total Gap	-8.669*** (3.072)				
Composition Effect	17.668*** (2.765)	2.011*** (0.743)	1.538** (0.734)	14.120*** (2.388)	
Return Effect	-26.337*** (3.136)	-11.026 (27.869)	-9.948 (18.127)	-55.596*** (20.413)	50.233 (39.638)

Note: Males are treated as the reference group. Robust standard errors are given in parentheses. The estimations are carried out using sample weights provided in the data set. ***, ** and * indicate that the estimated coefficients are statistically significant at the 1%, 5% and 10% levels respectively.

The return effect is negative and statistically significant at the 1% level, suggesting that males are more able to convert educational inputs into higher math test scores. The contributions of student and family characteristics to the return effect are negative but statistically insignificant, indicating that there is no discernible gender difference in transforming student and family inputs into math test scores. The negative and statistically significant contribution of school characteristics to the return effect suggests that males appear to have a particular advantage with converting school inputs into better math test scores.

Gender Test Score Gap in Science

OLS estimates of the gender test score gap in science is presented in Table 2.A.5 of Appendix 2.A. The raw test score gap in science is 10 score points in favor of females. As expected, students with a more positive view of their abilities in science and those expecting to pursue a science-related career at age 30 tend to have higher scores. The more motivated a student is to do well in science, the higher achievement in science. Moreover, students who attribute more importance to math score better. When we include family background characteristics, the gender test score

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

gap decreases in magnitude but remains statistically significant. Parental education, income, and occupational status, whether mother has a science-related career, the number of books at home, and the index of home education resources are statistically significant in explaining scientific literacy achievement. The next specification in column 4 adds the school input measures. The gender test score gap diminishes further in magnitude and loses statistical significance when these variables are included. School location and type, proportion of girls enrolled at school, and average class size are statistically significant predictors of achievement in science.

The results presented in columns 3 and 4 of Table 2.A.6 suggest that variables associated with statistically significant estimated coefficients are nearly the same for both gender, implying that there is no substantial difference between male and female education production function. However, the responsiveness to different covariates changes across genders. For instance, as in the case of math test results, father's education is more important for daughters and males respond to school characteristics differently than females do.

The standard BO Decomposition results are presented in Table 2.2.

Table 2.2: The standard BO decomposition of the gender test score gap in science

		Student Characteristics	Family Characteristics	School Characteristics	Constant
Total Gap	10.098*** (2.731)				
Composition Effect	16.517*** (2.498)	1.643** (0.720)	0.678 (0.631)	14.196*** (2.122)	
Return Effect	-6.419** (2.868)	-5.751 (22.252)	-11.871 (16.576)	-54.154*** (18.481)	65.357** (34.159)

Note: Males are treated as the reference group. Robust standard errors are given in parentheses. The estimations are carried out using sample weights provided in the data set. ***, ** and * indicate that the estimated coefficients are statistically significant at the 1%, 5% and 10% levels respectively.

The composition effect is positive and statistically significant, suggesting that gender differences in observable characteristics predict an advantage for girls over boys in the science test score. Consistent with math test score results, school characteristics is the most important factor contributing to the composition effect. The contribution of family characteristics to the composition effect is not statistically significant while differences in student characteristics account for only 4.1 percent of the composition effect. The return effect is negative and statistically significant, providing evidence that males are more efficient in transforming educational inputs into higher science test scores. Similar to the decomposition results for math score presented in Table 2.1, the most important advantage for males results from higher returns to school inputs and the contributions of student and family characteristics to the return effect are negative but statistically insignificant.

The Semiparametric BO Decomposition

Table 2.3 presents the results from the semiparametric BO decomposition of the mean test score gap that restricts the comparison to the common support. The upper panel of Table 2.3 shows the results for science while the lower part shows those for mathematics. As one moves down in both lower and upper panels of the table, the set of covariates steadily grows. In the final specification, the percentage of females who are out of the common support is 12.5 percent.²²

Table 2.3: Semiparametric BO decomposition of the mean test score gap for the common support subpopulation

	Characteristics	Δ_c	Δ_r	Δ_S
Science	Student	0.458 (2.124)	9.739*** (2.731)	10.197*** (2.687)
	Student + Family	4.306** (1.741)	5.757** (2.404)	10.063*** (2.674)
	Student + Family + School	19.526*** (2.146)	-4.380 (2.557)	15.146*** (2.883)
Math	Student	2.047 (2.482)	-10.694*** (3.273)	-8.647*** (3.053)
	Student + Family	6.839*** (1.955)	-15.537*** (2.647)	-8.698*** (3.049)
	Student + Family + School	22.169*** (2.408)	-25.290*** (2.768)	-3.121 (3.223)

Note: Males are treated as the reference group. The estimations are carried out using sample weights provided in the data set. Standard errors are given in parentheses and simulated with 500 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

The following conclusions can be drawn from Table 2.3. In science, except the first specification which controls for only student characteristics, the composition effect is positive and statistically significant, implying that gender differences in observable characteristics predict an advantage for girls over boys. As expected, the composition effect increases as we control for more covariates in the model. The contribution of school characteristics to the composition effect is very important. The return effect becomes statistically insignificant when we control for school characteristics. The final specification presented in the third row of Table 2.3 indicates that girls outperform boys in science by 15.1 points. In math, the mean test score gap is 8.7 points in favor of boys in the first two specifications, however it turns out to be statistically insignificant in the final specification. A comparison of the results presented in Table 2.3 and those in Tables 2.1 and 2.2 reveals that the standard BO

²²In Table 2.3, the percentage of females who are out of the common support changes across specifications. In the first specification which only controls for student characteristics, 0.29 percent of females are out of the common support while this rate is 0.04 percent for the second specification which controls for student and family characteristics and 12.47 percent for the final specification which also adds school characteristics to the second specification.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

decomposition tends to underestimate the component of the gender test score gap attributable to the composition effect.

Table 2.4 presents the results from the semiparametric BO decomposition that also accounts for the out-of-common-support observations.

Table 2.4: Semiparametric BO decomposition of the mean test score gap

Characteristics		Δ_1	Δ_c	Δ_r	Δ
Science	Student	-0.100 (0.139)	0.458 (2.124)	9.739*** (2.731)	10.098*** (2.679)
	Student + Family	0.034 (0.119)	4.306** (1.741)	5.757** (2.404)	10.098*** (2.679)
	Student + Family + School	-5.048*** (0.735)	19.526*** (2.146)	-4.380 (2.557)	10.098*** (2.679)
Math	Student	-0.022 (0.125)	2.047 (2.482)	-10.694*** (3.273)	-8.669*** (3.046)
	Student + Family	0.029 (0.108)	6.839*** (1.955)	-15.537*** (2.647)	-8.669*** (3.046)
	Student + Family + School	-5.548*** (0.753)	22.169*** (2.408)	-25.290*** (2.768)	-8.669*** (3.046)

Note: Males are treated as the reference group. The estimations are carried out using sample weights provided in the data set. Standard errors are given in parentheses and simulated with 500 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

The mean test score gap in science (math) in the full sample is 10.1 (-8.7) points and thus very different from the gap in the common support subpopulation presented in Table 2.3. This finding points to the importance of restricting the comparison only to those individuals with comparable characteristics. Δ_r and Δ_c are the same as those in Table 2.3 and they are computed only over the common support. Δ_1 , the difference between females who can be matched with males and those who cannot, is statistically significant only in the final specification which controls for student, family and school characteristics. In both science and math, the negative value of Δ_1 indicates that females who are in the common support perform better than those who are out of the common support.

It is worth noting that the outperformance of girls over boys in science could be partially explained by the fact that there exist gender differences in secondary education enrollment rate in Turkey where boys have higher enrolment rates than girls. Therefore, our sample is likely to be composed of a positively selected sample of girls, causing the overestimation of the gender gap favorable to girls.²³

²³Table 2.A.3 shows that most of the 15-year-old students in our sample are not subject to compulsory education which ends after 8th grade.

Table 2.A.7 of Appendix 2.A presents the results from the semiparametric BO decomposition at different quantiles, allowing us understand the heterogeneous pattern across the distribution. The top panel of Table 2.A.7 shows that in science, the gap is in favor of girls and statistically significant until the top quantile. The largest gap takes place at the 50th percentile. The bottom panel of Table 2.A.7 indicates that in math, the gap is statistically significant only at the top of the distribution. At the top end, girls lag significantly behind boys.

More insight give Figures 2.B.6 and 2.B.7 of Appendix 2.B, which plot the test score gaps at the percentiles. Figure 2.B.6 shows that the largest gaps in science are at very low quantiles and around the median. From the 70th percentile onwards the gap rapidly decreases and is negative at the 95th percentile. Afterwards the gap starts to increase again. In math the gap is positive for low quantiles and at becomes negative at the 10th percentile (see Figure 2.B.7). Between the 10th percentile and the 68th percentile the gap is very small, but slightly in favor of boys. From the 75th percentile onwards the gap is again decreasing. At the 99th percentile the gap amounts to -20.41.

2.6 Conclusion

In this paper, we use a semiparametric Blinder-Oaxaca (BO) decomposition to investigate the gender PISA test score gap in mathematics/science in Turkey. Our semiparametric approach differs from the standard BO decomposition in several aspects. It decomposes the average test score gap for the common support population and relaxes the parametric assumptions of the standard BO decomposition.

The results for the semiparametric BO decomposition evaluated at the mean of test scores indicate that the gender test score gap is 15.1 points in favor of girls in science while it is not statistically significant in math. The positive and statistically significant composition effect suggest that girls possess more of characteristics associated with high science test scores. School characteristics plays an important role in explaining the gap. Our findings provide evidence that the failure to recognize the common support problem leads to an underestimation of the composition effect. We also find that the gender test score gap changes across the test score distribution. In math, the gap is statistically significant only at the top end of the distribution suggesting that high-achieving boys perform better than high-achieving girls in math. In science, the gap favoring girls is statistically significant until the top quantile and the largest gap occurs at the 50th percentile. For both, math and science, we

observe a strong decline in the test score gap between the 70th percentile and the 95th percentile.

Bibliography

- ABADIE, A. AND G. W. IMBENS (2006): “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74, 235–267.
- AKKOYUNLU-WIGLEY, A. AND S. WIGLEY (2008): “Basic Education and Capability Development in Turkey,” in *Education in Turkey, Vol. 26*, ed. by A.-M. Nohl, A. Akkoyunlu-Wigley, and S. Wigley, Waxmann Publishing, New York/Münster.
- AMMERMÜLLER, A. (2007): “PISA: What makes the Difference? Explaining the Gap in Test Scores between Finland and Germany,” *Empirical Economics*, 33, 263–287.
- AYPAY, A., M. ERDOĞAN, AND M. SÖZER (2007): “Variation among Schools on Classroom Practices in Science based on TIMSS-1999 in Turkey,” *Journal of Research in Science Teaching*, 44, 1417–1435.
- BLANCY, N. K. AND A. SASMAZ (2011): “PISA 2009: Where does Turkey Stand?” *Turkish Policy Quarterly*, 10, 125–134.
- BLINDER, A. (1973): “Wage Discrimination: Reduced Form and Structural Estimates,” *Journal of Human Resources*, 8, 436–455.
- BOTEZAT, A. AND R. R. SEIBERLICH (2013): “Educational Performance Gaps in Eastern Europe,” *The Economics of Transition*, forthcoming.
- BROSNAN, M. (2006): “Digit Ratio and Faculty Membership: Implications for the Relationship between Prenatal Testosterone and Academia,” *British Journal of Psychology*, 97, 455–466.
- BUSSO, M., J. DINARDO, AND J. MCCRARY (2009): “Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects,” Unpublished manuscript, http://emlab.berkeley.edu/~jmccrary/BDM_JBES.pdf.
- CAHILL, L. (2012): “His Brain, Her Brain,” *Special Editions, Scientific American*, 1, 46–53.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

- CARNEIRO, P. AND J. HECKMAN (2003): "Human Capital Policy," *Working Paper 9495*, National Bureau of Economic Research.
- CARRELL, S., M. PAGE, AND J. WEST (2009): "Sex and Science: How Professor Gender Perpetuates the Gender Gap," *The Quarterly Journal of Economics*, 125, 1101–1144.
- CAYGILL, R. (2003): "PISA 2006: Student Attitudes to and Engagement with Science - How Ready are our 15-year-olds for Tomorrow's World?" Technical Report, Ministry of Education, Wellington, New Zealand.
- CECI, S., W. WILLIAMS, AND S. BARNETT (2009): "Women's Underrepresentation in Science: Sociocultural and Biological Considerations," *Psychological Bulletin*, 135, 218–261.
- CRUMP, R. K., V. J. HOTZ, G. W. IMBENS, AND O. A. MITNIK (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199.
- DAVISON, K. AND E. SUSMAN (2001): "Are Hormone Levels and Cognitive Ability Related during early Adolescence?" *International Journal of Behavioral Development*, 25, 416–428.
- DAYIOĞLU, M., M. KIRDAR, AND A. TANSEL (2009): "Impact of Sibship Size, Birth Order and Sex Composition on School Enrolment in Urban Turkey," *Oxford Bulletin of Economics and Statistics*, 71, 399–426.
- DAYIOĞLU, M. AND S. TÜRÜT-AŞIK (2007): "Gender Differences in Academic Performance in a Large Public University in Turkey," *Higher Education*, 53, 255–277.
- DINCER, M. AND G. UYSAL (2010): "The Determinants of Student Achievement in Turkey," *International Journal of Educational Development*, 30, 592–598.
- DUNCAN, K. C. AND J. SANDY (2007): "Explaining the Performance Gap Between Public and Private School Students," *Eastern Economic Journal*, 33, 177–191.
- ERBERBER, E. (2010): "Analyzing Turkey's Data from TIMSS 2007 to Investigate Regional Disparities in Eighth Grade Science Achievement," in *The Impact of International Achievement Studies on National Education Policymaking (International Perspectives on Education and Society, Vol. 13)*, ed. by A. W. Wiseman, Amsterdam: North-Holland.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

- EURYDICE (2010): “Organisation of the Education System in Turkey,” Technical Report, European Commission, Brussels, Belgium.
- FORTIN, N., T. LEMIEUX, AND S. FIRPO (2010): “Decomposition Methods in Economics,” in *Handbook of Labour Economics*, ed. by O. Ashenfelter and D. Card, Amsterdam: North-Holland.
- FRÖLICH, M. (2004): “Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators,” *Review of Economics and Statistics*, 86, 77–90.
- (2007): “Propensity Score Matching without Conditional Independence Assumption—With an Application to the Gender Wage Gap in the United Kingdom,” *Econometrics Journal*, 10, 359–407.
- FRYER, R. G. AND S. D. LEVITT (2010): “An Empirical Analysis of the Gender Gap in Mathematics,” *American Economic Journal: Applied Economics*, 2, 210–240.
- FUCHS, T. AND L. WÖSSMANN (2007): “What accounts for International Differences in Student Performance? A Re-Examination using PISA Data,” *Empirical Economics*, 32, 433–464.
- GUIO, L., F. MONTE, P. SAPIENZA, AND L. ZINGALES (2008): “Culture, Gender, and Math,” *Science*, 320, 1164–1165.
- HAHN, J. (1998): “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66, 315–331.
- HANUSHEK, E., J. F. KAIN, AND S. G. RIVKIN (2009): “New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement,” *Journal of Labor Economics*, 27, 349–383.
- HANUSHEK, E. A. (2003): “The Failure of Input-Based Schooling Policies,” *The Economic Journal*, 113, F64–F98.
- HANUSHEK, E. A. AND D. D. KIMKO (2000): “Schooling, Labor-Force Quality, and the Growth of Nations,” *The American Economic Review*, 90, 1184–1208.
- HANUSHEK, E.A., V. L. AND K. HITOMI (2008): “Do Students Care about School Quality? Determinants of Dropout Behavior in Developing Countries,” *Journal of Human Capital*, 2, 69–105.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

- HECKMAN, J., H. ICHIMURA, AND P. TODD (1997): “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *The Review of Economic Studies*, 64, 605–654.
- HECKMAN, J., J. LALONDE, AND J. SMITH (1999): “The Economics and Econometrics of Active Labor Market Programs,” in *Handbook of Labor Economics*, Vol. 3A, ed. by O. Ashenfelter and D. Card, Elsevier Science, 1865–2097.
- HILLE, A. (2011): “The Gender Gap in Mathematics in French Primary School,” Master’s thesis.
- HISARCIKLILAR, M., A. MCKAY, AND P. WRIGHT (2010): “Gender Based Differences in Educational Achievement in Turkey: What Has Changed Over Time?” *Working Paper*, presented at the 30th Annual Conference of the MEEA.
- IMBENS, G. (2004): “Nonparametric Estimation of Average Treatment Effects Under Exogeneity,” *Review of Economics and Statistics*, 86, 4–29.
- JACOBS, J. AND J. ECCLES (1992): “The Impact of Mothers’ Gender-Role Stereotypic Beliefs on Mothers’ and Children’s Ability Perceptions,” *Journal of Personality and Social Psychology*, 63, 932–944.
- JAMISON, E., D. JAMISON, AND E. HANUSHEK (2007): “The Effects of Education Quality on Mortality Decline and Income Growth,” *Economics of Education Review*, 26, 772–789.
- KHAN, S. AND E. TAMER (2010): “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” *Econometrica*, 78, 2021–2042.
- KRIEG, J. M. AND P. STORER (2006): “How Much Do Students Matter? Applying the Oaxaca Decomposition to Explain Determinants of Adequate Yearly Progress,” *Contemporary Economic Policy*, 24, 563–581.
- KUCIAN, K., T. LOENNEKER, T. DIETRICH, E. MARTIN, AND M. VON ASTER (2005): “Gender Differences in Brain Activation Patterns During Mental Rotation and Number Related Cognitive Tasks,” *Psychology Science*, 47, 112–131.
- LARA-CINISOMO, S., A. PEBLEY, M. VAIANA, E. MAGGIO, M. BERENDS, AND S. LUCAS (2004): “A Matter of Class,” *Rand Review*, 28, 10–5.
- LAVY, V. AND A. SCHLOSSER (2011): “Mechanisms and Impacts of Gender Peer Effects at School,” *American Economic Journal: Applied Economics*, 3, 1–33.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

- LECHNER, M. (2010): “A Note on the Common Support Problem in Applied Evaluation Studies,” *Annals of Economics and Statistics*, 91–92, 217–234.
- MARSHALL, N., R. CAYGILL, AND S. MAY (2008): “PISA2006: Reading Literacy: How Ready are Our 15-year-olds for Tomorrow’s World?” Technical Report, Ministry of Education, Wellington, New Zealand.
- MCEWAN, P. J. (2004): “The Indigenous Test Score Gap in Bolivia and Chile,” *Economic Development and Cultural Change*, 53, 157–190.
- MULLIGAN, C. (1999): “Galton versus the Human Capital Approach to Inheritance,” *Journal of Political Economy*, 107, S184–S224.
- MURNANE, R., J. WILLETT, Y. DUHALDEBORDE, AND J. TYLER (2000): “How Important are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?” *Journal of Policy Analysis and Management*, 19, 547–568.
- NONOYAMA-TARUMI, Y. AND J. WILLMS (2010): “The Relative and Absolute Risks of Disadvantaged Family Background and Low Levels of School Resources on Student Literacy,” *Economics of Education Review*, 29, 214–224.
- ÑOPO, H. (2008): “Matching as a Tool to Decompose Wage Gaps,” *Review of Economics and Statistics*, 90, 290–299.
- OAXACA, R. (1973): “Male-Female Wage Differentials in Urban Labor Markets,” *International Economic Review*, 14, 693–709.
- OECD (2007): “PISA 2006 Science Competencies for Tomorrow’s World,” Technical Report, OECD, Paris, France.
- (2009): “Education at a Glance 2009,” Technical Report, OECD, Paris, France.
- (2010): “PISA 2009 Results: What Students Know and Can Do. Student Performance in Reading, Mathematics and Science,” Technical Report, OECD, Paris, France.
- POPE, D. AND J. SYDNOR (2010): “Geographic Variation in the Gender Differences in Test Scores,” *The Journal of Economic Perspectives*, 24, 95–108.
- ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (1994): “Estimation of Regres-

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

- sion Coefficients when Some Regressors are not Always Observed,” *Journal of the American Statistical Association*, 89, 846–866.
- ROSENBAUM, P. R. AND D. B. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.
- SAKELLARIOU, C. (2008): “Peer Effects and the Indigenous/Non-Indigenous early Test-Score Gap in Peru,” *Education Economics*, 16, 371–390.
- SEIFERT, B. AND T. GASSER (1996): “Finite-Sample Variance of Local Polynomials: Analysis and Solutions,” *Journal of the American Statistical Association*, 91, 267–275.
- (2000): “Data Adaptive Ridging in Local Polynomial Regression,” *Journal of Computational and Graphical Statistics*, 9, 338–360.
- SMITS, J. AND A. HOSGOR (2006): “Effects of Family Background Characteristics on Educational Participation in Turkey,” *International Journal of Educational Development*, 26, 545–560.
- SOHN, K. (2012): “A new Insight into the Gender Gap in Math,” *Bulletin of Economic Research*, 64, 135–155.
- TANSEL, A. (2002): “Determinants of School Attainment of Boys and Girls in Turkey: Individual, Household and Community Factors,” *Economics of Education Review*, 21, 455–470.
- TIEDEMANN, J. (2000): “Parents’ Gender Stereotypes and Teachers’ Beliefs as Predictors of Children’s Concept of their Mathematical Ability in Elementary School,” *Journal of Educational Psychology*, 92, 144–151.
- WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.
- WÖSSMANN, L. (2003): “Schooling Resources, Educational Institutions and Student Performance: The International Evidence,” *Oxford Bulletin of Economics and Statistics*, 65, 117–170.
- WÖSSMANN, L., E. LÜDEMANN, G. SCHÜTZ, AND M. R. WEST (2009): *School Accountability, Autonomy and Choice around the World*, Elgar, Cheltenham.

Appendix 2.A Tables

Table 2.A.1: The index of beliefs in own abilities in science

The index	Construction of the Index	Interpretation
1- Index of self-concept in science:	<p>It was created by using students' responses to the following six statements:</p> <ol style="list-style-type: none"> 1. I can usually give good answers to test questions on science topics 2. When I am being taught science, I can understand the concepts very well 3. I can easily understand new ideas in science 4. I learn science topics quickly 5. Science topics are easy for me 6. Learning advanced science <p>Response options for each statement were: strongly agree, agree, disagree, and strongly disagree</p>	<p>Students who agreed with these statements were higher on the index, and students who reacted more negatively to the statements were lower on the index.</p>
2-Index of self-efficacy in science:	<p>It was created by using students' responses to these eight tasks.</p> <ol style="list-style-type: none"> 1. Explain why earthquakes occur more frequently in some areas than in others 2. Recognise the science question that underlies a newspaper report on a health issue 3. Interpret the scientific information provided on the labelling of food items 4. Predict how changes to an environment will affect the survival of certain species 5. Identify the science question associated with the disposal of rubbish 6. Describe the role of antibiotics in the treatment of disease 7. Identify the better of two explanations for the formation of acid rain 8. Discuss how new evidence can lead you to change your understanding about the possibility of life on Mars <p>Response options for each statement were: I could do this easily, I could do this with a bit of effort, I would struggle to do this on my own, and I couldn't do this.</p>	<p>Students who agreed they could do these tasks were higher on the index, and students who reacted more negatively were lower on the index.</p>

Source: Caygill (2003).

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

Table 2.A.2: The index of motivation in science

The Index	Construction of the Index	Interpretation
1- Index of general interest in science	<p>It was created by using students' responses to the following eight statements:</p> <ol style="list-style-type: none"> 1. Human biology 2. Topics in chemistry 3. Topics in astronomy 4. Topics in physics 5. The biology of plants 6. The ways scientists design experiments 7. Topics in geology 8. What is required for scientific explanations <p>Response options for each statement were: high interest, medium interest, low interest, and no interest.</p>	<p>Students who reported higher interest were higher on the index, and students who were less interested in the science topics were lower on the index.</p>
2- Index of enjoyment of science	<p>It was created by using students' responses to the following five statements:</p> <ol style="list-style-type: none"> 1. I enjoy acquiring new knowledge in science 2. I generally have fun when I am learning science topics 3. I am interested in learning about science 4. I like reading about science 5. I am happy doing science problems <p>Response options for each statement were: strongly agree, agree, disagree, and strongly disagree.</p>	<p>Students who agreed with these statements were higher on the index, and students who reacted more negatively to the statements were lower on the index.</p>
3- Index of instrumental motivation in science	<p>It was created by combining students' responses to the following five statements:</p> <ol style="list-style-type: none"> 1. I study science because I know it is useful for me 2. Making an effort in my science subject(s) is worth it because this will help me in the work I want to do later on 3. Studying my science subject(s) is worthwhile for me because what I learn will improve my career prospects 4. I will learn many things in my science subject(s) that will help me get a job 5. What I learn in my science subject(s) is worthwhile for me because I need this for what I want to study later on <p>Response options for each statement were: strongly agree, agree, disagree, and strongly disagree. The proportions shown in this table combine those who agreed and those who strongly agreed.</p>	<p>Students who agreed with these statements were higher on the index, and students who reacted more negatively to the statements were lower on the index.</p>

Source: Caygill (2003).

Table 2.A.3: Descriptive statistics by gender

Variable	Description	Full Sample		Male		Female		t-Stat.
		Mean	St. Dv.	Mean	St. Dv.	Mean	St. D.	
<u>Test Scores</u>								
science	science test score	432.05	78.42	427.41	80.19	437.51	75.94	-4.00
math	math test score	432.35	87.68	436.33	90.43	427.66	84.11	3.07
<u>Student Characteristics</u>								
8 th grade	=1 if the student is in 7 th or 8 th grade	0.04	0.19	0.04	0.20	0.03	0.18	1.28
9 th grade	=1 if the student is in 9 th grade	0.40	0.49	0.39	0.49	0.40	0.49	-0.49
10 th grade	=1 if the student is in 10 th grade	0.54	0.50	0.53	0.50	0.54	0.50	-0.42
11 th grade	=1 if the student is in 11 th grade	0.03	0.16	0.03	0.17	0.02	0.15	1.28
science career	=1 if the student is expected to have a science-related career at 30	0.26	0.44	0.25	0.43	0.27	0.44	-1.66
motivation index	index of motivation in science	1.07	2.45	1.01	2.43	1.13	2.47	-1.54
ability index	index of belief in own ability in science	0.21	1.66	0.20	1.73	0.22	1.58	-0.42
math is important	How important is math, 4='very important' 1='not important at all'	3.62	0.66	3.58	0.69	3.67	0.62	-4.34
<u>Family Background Characteristics</u>								
mother-primaryeduc	=1 if the mother has at most primary education	0.72	0.45	0.72	0.45	0.71	0.45	0.91
mother-secondaryeduc	=1 if the mother has secondary education	0.22	0.42	0.21	0.41	0.23	0.42	-1.81
mother-tertiaryeduc	=1 if the mother has tertiary education	0.06	0.24	0.07	0.25	0.06	0.23	1.42
father-primaryeduc	=1 if the father has at most primary education	0.55	0.50	0.55	0.50	0.54	0.50	0.91
father-secondaryeduc	=1 if the father has secondary education	0.31	0.46	0.30	0.46	0.32	0.47	-1.80
father-tertiaryeduc	=1 if the father has tertiary education	0.14	0.35	0.15	0.36	0.14	0.34	1.08
books \leq 10	=1 if the number of books at home \leq 10	0.23	0.42	0.27	0.44	0.18	0.39	6.08
11 \leq books \leq 25	=1 if 11 \leq the number of books at home \leq 25	0.28	0.45	0.27	0.44	0.30	0.46	-2.56
26 \leq books \leq 100	=1 if 26 \leq the number of books at home \leq 100	0.30	0.46	0.28	0.45	0.31	0.46	-2.27
books $>$ 100	=1 if the number of books at home $>$ 100	0.19	0.39	0.19	0.39	0.20	0.40	-0.81
parents' occupational status	the index of the highest parental occupational status	39.84	15.71	39.11	15.86	40.70	15.50	-3.14
home education resources	the index of home education resources	-0.64	1.30	-0.67	1.34	-0.61	1.25	-1.40
mother-science career	=1 if the mother has a science-related career	0.02	0.13	0.02	0.13	0.02	0.12	0.69
father-science career	=1 if the father has a science-related career	0.05	0.21	0.05	0.21	0.05	0.21	-0.26
income $<$ 0.5 median	=1 if the family income \leq 0.5 median annual income	0.35	0.48	0.35	0.48	0.34	0.47	1.10
0.5 median \leq income $<$ 0.75 median	=1 if 0.5 median \leq the family income $<$ 0.75 median	0.36	0.48	0.35	0.48	0.38	0.48	-1.57
0.75 median \leq income $<$ median	=1 if 0.75 median \leq the family income $<$ median	0.20	0.40	0.20	0.40	0.20	0.40	-0.26
median \leq income $<$ 1.25 median	=1 if median \leq the family income $<$ 1.25 median	0.06	0.24	0.06	0.24	0.05	0.23	1.05
income \geq 1.25 median	=1 if the family income \geq 1.25 median	0.03	0.17	0.03	0.17	0.03	0.16	0.55

Note: PISA 2006 data, own calculations. A detailed description of the variables can be received upon request. The last column presents the test statistics of a two sample, weighted t-test using Welch's approximation.

Table 2.A.3 (cont'd): Descriptive Statistics by gender

Variable	Description	Full Sample		Male		Female		t-Stat.
		Mean	St. Dv.	Mean	St. Dv.	Mean	St. D.	
School's Characteristics								
percentage of girls	percentage of girls enrolled at school	0.43	0.21	0.35	0.19	0.53	0.20	-28.63
public	=1 if the school is public	0.98	0.15	0.97	0.17	0.98	0.14	-1.97
class size	=1 if the average class size is more than 30 at school	0.51	0.50	0.50	0.50	0.53	0.50	-1.57
rural	=1 if the school is in a rural area	0.79	0.40	0.82	0.39	0.77	0.42	4.01
school education resources	the index of the quality of the school's educational resources	-0.81	0.92	-0.83	0.91	-0.80	0.93	-0.91
general high school	=1 if the school is a general high school	0.44	0.50	0.42	0.49	0.47	0.50	-2.72
anatolian high school	=1 if the school is an anatolian high school	0.20	0.40	0.18	0.39	0.22	0.41	-2.42
vocational high school	=1 if the school is a vocational high school	0.36	0.48	0.39	0.49	0.32	0.47	4.88
marmara region	=1 if the school is in Marmara region	0.28	0.45	0.26	0.44	0.29	0.46	-2.21
central anatolian region	=1 if the school is in Central Anatolian region	0.21	0.41	0.21	0.41	0.21	0.41	0.16
aegean region	=1 if the school is in Aegean region	0.14	0.34	0.13	0.34	0.14	0.35	-1.15
mediterranean region	=1 if the school is in Mediterranean region	0.11	0.32	0.10	0.30	0.13	0.33	-2.45
blacksea region	=1 if the school is in Blacksea region	0.13	0.33	0.15	0.35	0.11	0.31	3.85
eastern anatolian region	=1 if the school is in Eastern Anatolian region	0.07	0.25	0.06	0.24	0.07	0.26	-0.95
southeastern anatolian region	=1 if the school is in Southeastern Anatolian region	0.07	0.25	0.08	0.28	0.05	0.21	4.38
N	number of observations	3832		2044		1788		

Notes: The last column presents the test statistics of a two sample, weighted t-test using Welch's approximation. The median annual income is 24.000 TL in Turkey.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

Table 2.A.4: OLS estimates of the gender test score gap in math

	(1)	(2)	(3)	(4)
female	-8.669*** (2.839)	-11.318*** (2.565)	-13.583*** (2.317)	-19.589*** (2.274)
9 th grade		59.802*** (6.791)	19.969*** (6.305)	15.273** (6.112)
10 th grade		31.552*** (6.676)	-1.730 (6.208)	16.870*** (6.003)
11 th grade		59.378*** (10.173)	26.042*** (9.305)	47.482*** (8.658)
motivation index		-1.641** (0.775)	-0.056 (0.700)	0.993 (0.628)
ability index		8.758*** (1.084)	4.519*** (0.986)	3.545*** (0.881)
science career		56.017*** (3.125)	40.328*** (2.858)	32.127*** (2.559)
math is important		11.368*** (2.040)	11.084*** (1.832)	10.715*** (1.635)
mother-primaryeduc			-37.292*** (5.956)	-20.397*** (5.336)
mother-secondaryedu			-30.143*** (5.734)	-21.399*** (5.113)
father-primaryedu			-18.431*** (4.677)	-10.626** (4.202)
father-secondaryedu			-13.264*** (4.250)	-7.969** (3.801)
11 ≤ books ≤ 25			15.371*** (3.280)	7.363** (2.932)
26 ≤ books ≤ 100			24.962*** (3.464)	12.986*** (3.119)
books > 100			26.833*** (4.119)	12.627*** (3.699)
parents' occupational status			0.554*** (0.095)	0.473*** (0.085)
educational resources			9.972*** (1.063)	7.284*** (0.968)
mother-science career			22.069** (9.643)	17.407** (8.600)
father-science career			2.919 (5.830)	0.216 (5.192)
0.5 median ≤ income < 0.75 median			5.315* (2.836)	4.704* (2.531)
0.75 median ≤ income < median			15.951*** (3.537)	8.280*** (3.158)
median ≤ income < 1.25 median			11.383** (5.620)	8.428* (5.022)
1.25 median ≤ income			-1.901 (7.168)	-2.735 (6.384)
percentage of girls				10.803* (5.786)
public				-2.533 (7.161)
class size				-4.111 (2.703)
rural				-15.708*** (2.725)
school's educational resources				3.353*** (1.244)
anatolian high school				58.920*** (3.619)
vocational high school				-31.021*** (2.538)
central anatolian region				2.950 (3.276)
aegean region				18.365*** (3.469)
mediterranean region				12.627*** (3.833)
blacksea region				10.008*** (3.770)
eastern anatolian region				-38.106*** (4.655)
southeastern anatolian region				-15.885*** (4.695)
<i>N</i>	3832	3832	3832	3832
<i>R</i> ²	0.002	0.193	0.353	0.491

Notes: The dependent variable is the PISA math score. Standard errors are given in parentheses. The estimations are carried out using sample weights provided in the data set. ***, ** and * indicate that the estimated coefficients are statistically significant at the 1%, 5% and 10% levels respectively. Constants are not reported. The reference categories for grade, mother's education, father's education, the number of books at home, family income, school type and region are students who are in the 7th grade or in the 7th grade, mothers with tertiary education, fathers with tertiary education, the number of books at home ≤ 10, the family income < 0.5 median family income, general high school and marmara region respectively.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

Table 2.A.5: OLS estimates of the gender test score gap in science

	(1)	(2)	(3)	(4)
female	10,098*** (2.537)	7,922*** (2.240)	6,434*** (2.054)	0.373 (2.061)
9 th grade		61,430*** (5.932)	28,423*** (5.589)	28,946*** (5.538)
10 th grade		26,425*** (5.832)	-1,266 (5.502)	17,269*** (5.438)
11 th grade		42,320*** (8.885)	13,932* (8.248)	34,363*** (7.844)
motivation index		-0,720 (0.677)	0,500 (0.621)	1,304** (0.569)
ability index		9,817*** (0.947)	6,338*** (0.874)	5,730*** (0.798)
science career		44,044*** (2.729)	31,080*** (2.533)	24,539*** (2.319)
math is important		6,819*** (1.782)	6,581*** (1.624)	6,270*** (1.481)
mother-primaryeduc			-20,482*** (5.280)	-7,708 (4.835)
mother-secondaryedu			-17,524*** (5.083)	-11,185** (4.633)
father-primaryedu			-20,103*** (4.146)	-13,737*** (3.807)
father-secondaryedu			-16,770*** (3.767)	-12,850*** (3.443)
11 ≤ books ≤ 25			6,918** (2.907)	0,283 (2.656)
26 ≤ books ≤ 100			18,213*** (3.071)	8,050*** (2.826)
books > 100			25,429*** (3.651)	13,462*** (3.351)
parents' occupational status			0,400*** (0.084)	0,331*** (0.077)
educational resources			7,713*** (0.942)	5,453*** (0.877)
mother-science career			18,784** (8.548)	15,621** (7.791)
father-science career			4,889 (5.167)	3,711 (4.704)
0.5 median ≤ income < 0.75 median			6,703*** (2.514)	6,066*** (2.293)
0.75 median ≤ income < median			15,998*** (3.135)	9,511*** (2.861)
median ≤ income < 1.25 median			11,342** (4.981)	8,665* (4.550)
1.25 median ≤ income			-1,631 (6.353)	-2,185 (5.784)
percentage of girls				14,989*** (5.243)
public				-2,727 (6.488)
class size				-5,586** (2.449)
rural				-11,876*** (2.469)
school's educational resources				1,799 (1.127)
anatolian high school				40,860*** (3.279)
vocational high school				-28,882*** (2.300)
central anatolian region				0,189 (2.968)
aegean region				8,909*** (3.143)
mediterranean region				9,038*** (3.473)
blacksea region				10,041*** (3.416)
eastern anatolian region				-29,894*** (4.218)
southeastern anatolian region				-22,642*** (4.254)
<i>N</i>	3832	3832	3832	3832
<i>R</i> ²	0.004	0.230	0.365	0.477

Notes: The dependent variable is the PISA science score. Standard errors are given in parentheses. The estimations are carried out using sample weights provided in the data set. ***, ** and * indicate that the estimated coefficients are statistically significant at the 1%, 5% and 10% levels respectively. Constants are not reported. The reference categories for grade, mother's education, father's education, the number of books at home, family income, school type and region are students who are in the 7th grade or in the 7th grade, mothers with tertiary education, fathers with tertiary education, the number of books at home ≤ 10, the family income < 0.5 median family income, general high school and marmara region respectively.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

Table 2.A.6: Estimates of the responsiveness of test scores to covariates by gender

	<u>Math</u>		<u>Science</u>	
	Male	Female	Male	Female
9 th grade	13.070 (8.412)	19.108** (8.827)	26.956*** (7.501)	32.031*** (8.233)
10 th grade	25.616*** (8.214)	11.429 (8.686)	22.858*** (7.324)	14.485* (8.102)
11 th grade	53.498*** (11.610)	41.318*** (12.831)	35.537*** (10.353)	34.935*** (11.969)
motivation index	2.072** (0.890)	-0.780 (0.872)	1.503** (0.794)	0.866 (0.814)
ability index	1.663 (1.213)	6.094*** (1.278)	5.259*** (1.081)	6.219*** (1.192)
science career	31.131*** (3.622)	33.842*** (3.553)	25.028*** (3.230)	23.573*** (3.314)
math is important	11.569*** (2.162)	10.490*** (2.457)	6.752*** (1.928)	6.116*** (2.292)
mother-primaryedu	-20.024*** (7.164)	-21.125*** (7.978)	-3.684 (6.388)	-14.222* (7.442)
mother-secondaryedu	-22.187*** (6.904)	-18.496** (7.582)	-10.496* (6.156)	-12.325* (7.073)
father-primaryedu	-3.797 (5.682)	-16.628*** (6.181)	-8.866* (5.066)	-18.037*** (5.766)
father-secondaryedu	-0.962 (5.150)	-15.643*** (5.570)	-8.881* (4.592)	-17.000*** (5.196)
11 ≤ books ≤ 25	6.854* (3.947)	6.835 (4.330)	-0.837 (3.519)	0.447 (4.039)
26 ≤ books ≤ 100	11.097*** (4.251)	12.566*** (4.563)	4.881 (3.790)	9.670** (4.256)
books > 100	11.201** (5.068)	13.336** (5.390)	11.288** (4.519)	14.580*** (5.028)
parents' occupational status	0.404*** (0.120)	0.508*** (0.118)	0.277*** (0.107)	0.358*** (0.110)
educational resources	5.211*** (1.332)	8.941*** (1.395)	4.510*** (1.188)	5.796*** (1.301)
mother-science career	21.681* (11.648)	12.002 (12.588)	21.008** (10.387)	8.526 (11.743)
father-science career	4.239 (4.441)	-2.433 (7.151)	8.490 (6.635)	-1.127 (6.671)
0.5 median ≤ income < 0.75 median	7.866** (3.609)	2.349 (3.501)	7.020** (3.218)	5.000 (3.266)
0.75 median ≤ income < median	4.934 (4.499)	12.044*** (4.337)	7.694* (4.012)	11.944*** (4.046)
median ≤ income < 1.25 median	8.457 (6.869)	8.598 (7.283)	12.098** (6.125)	4.380 (6.793)
1.25 median ≤ income	-0.134 (8.792)	-3.681 (9.175)	5.836 (7.839)	-11.285 (8.559)
percentage of girls	44.294*** (10.679)	-23.107*** (8.683)	54.040*** (9.522)	-11.817 (8.099)
public	6.788 (9.365)	-8.953 (11.143)	1.867 (8.350)	-4.927 (10.395)
class size	4.412 (3.990)	-6.026 (3.772)	2.576 (3.558)	-7.052** (3.519)
rural	-16.494*** (4.042)	-14.361*** (3.622)	-13.711*** (3.604)	-9.183*** (3.379)
school's educational resources	5.954*** (1.777)	4.204** (1.766)	4.724*** (1.584)	1.609 (1.647)
anatolian high school	79.456*** (5.342)	47.990*** (4.963)	60.012*** (4.763)	30.539*** (4.630)
vocational high school	-29.851*** (4.241)	-17.610*** (3.777)	-21.365*** (3.782)	-22.869*** (3.523)
central anatolian region	9.906** (4.911)	5.843 (4.463)	11.329*** (4.379)	-2.860 (4.163)
aegean region	13.320*** (4.985)	28.164*** (4.844)	8.175* (4.445)	14.073*** (4.518)
mediterranean region	13.843** (5.573)	14.655*** (5.211)	13.763*** (4.970)	5.620 (4.861)
blacksea region	10.490** (5.268)	19.278*** (5.523)	16.139*** (4.697)	10.722** (5.152)
eastern anatolian region	-27.476*** (6.832)	-37.781*** (6.423)	-12.019** (6.092)	-37.011*** (5.992)
southeastern anatolian region	-11.273* (6.177)	-21.796*** (7.437)	-12.527** (5.508)	-33.479*** (6.937)
<i>N</i>	2044	1788	2044	1788
<i>R</i> ²	0.513	0.499	0.507	0.466

Notes: In the first two columns, the dependent variable is the PISA math score while in the last two columns, it is the PISA science score. Standard errors are given in parentheses. The estimations are carried out using sample weights provided in the data set. ***, ** and * indicate that the estimated coefficients are statistically significant at the 1%, 5% and 10% levels respectively. Constants are not reported. The reference categories for grade, mother's education, father's education, the number of books at home, family income, school type and region are students who are in the 7th grade or in the 7th grade, mothers with tertiary education, fathers with tertiary education, the number of books at home ≤ 10, the family income < 0.5 median family income, general high school and marmara region respectively.

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

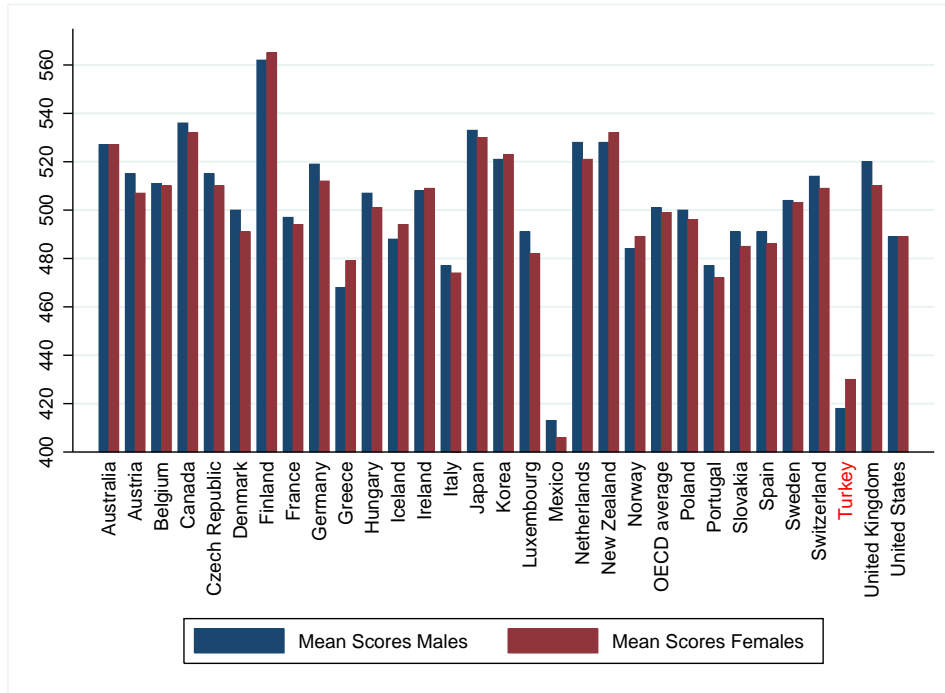
Table 2.A.7: Semiparametric BO decomposition across the distribution for the common support subpopulation

		5 %	25 %	50 %	75 %	95 %
		Quantile	Quantile	Quantile	Quantile	Quantile
Science	$\hat{\Delta}_S$	12.775** (4.158)	18.277*** (3.572)	20.887*** (3.485)	16.505*** (4.256)	-2.424 (6.710)
Math	$\hat{\Delta}_{\Delta_S}$	-0.467 (5.469)	-2.181 (4.170)	-0.156 (4.227)	0.312 (5.895)	-16.124* (8.355)

Notes: The estimations are carried out using sample weights provided in the data set. Standard errors are given in parentheses and simulated with 500 bootstrap replications. * if the 5% and 95% quantile of the bootstrap distribution have the same signs, ** if the 2.5% and 97.5% quantile of the bootstrap distribution have the same signs, *** if the 0.5% and 99.5% quantile of the bootstrap distribution have the same signs.

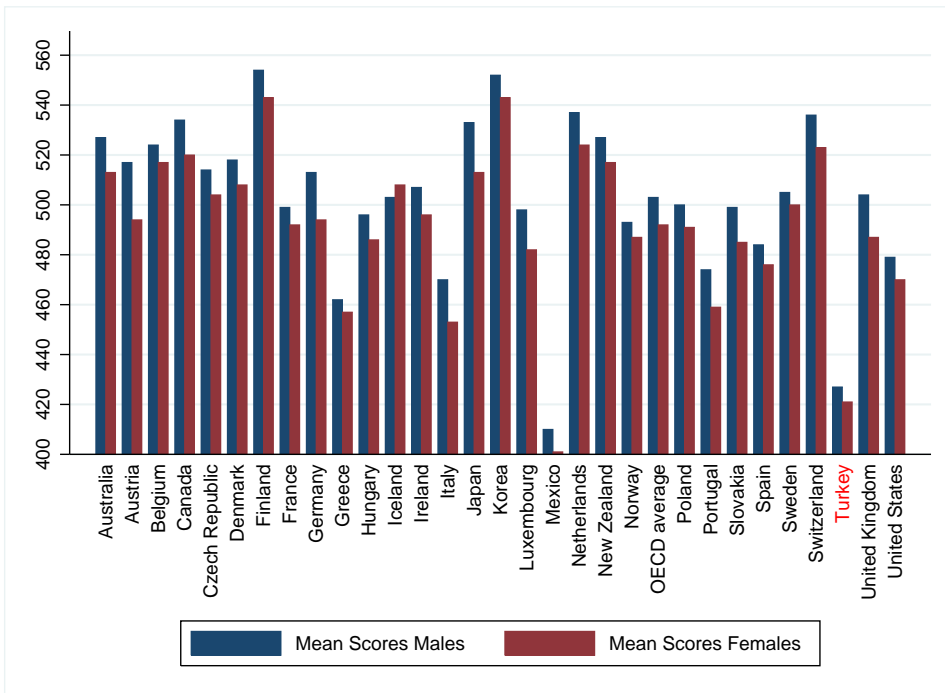
Appendix 2.B Figures

Figure 2.B.1: The mean gender test score gap in science across OECD countries



Source: OECD (2010).

Figure 2.B.2: The mean gender test score gap in mathematics across OECD countries



Source: OECD (2010).

2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

Figure 2.B.3: Histogram estimates of the propensity score distributions using student characteristics

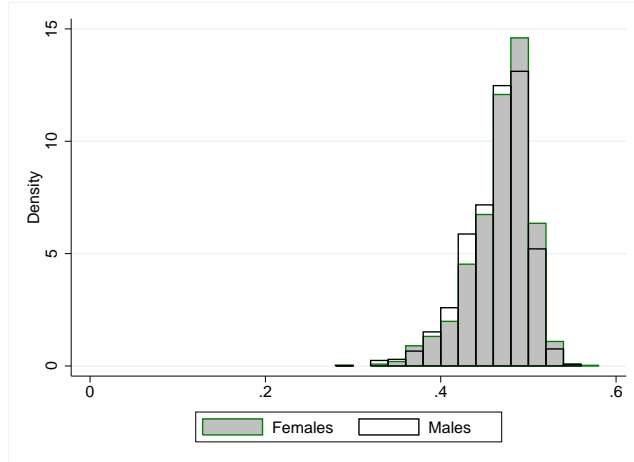


Figure 2.B.4: Histogram estimates of the propensity score distributions using student and family characteristics

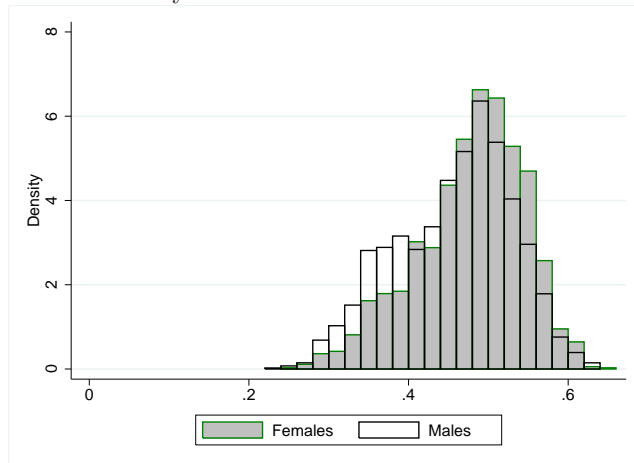
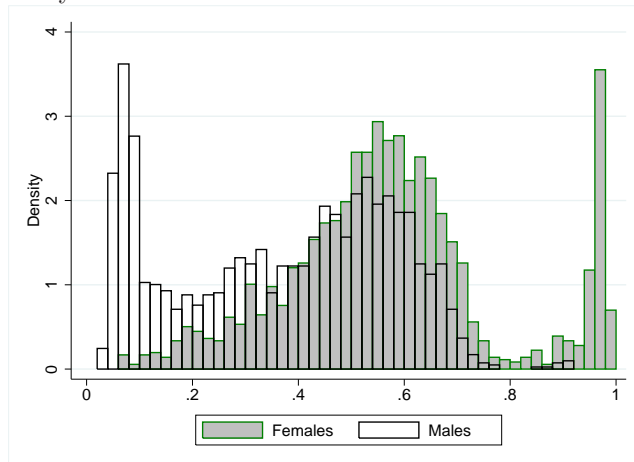
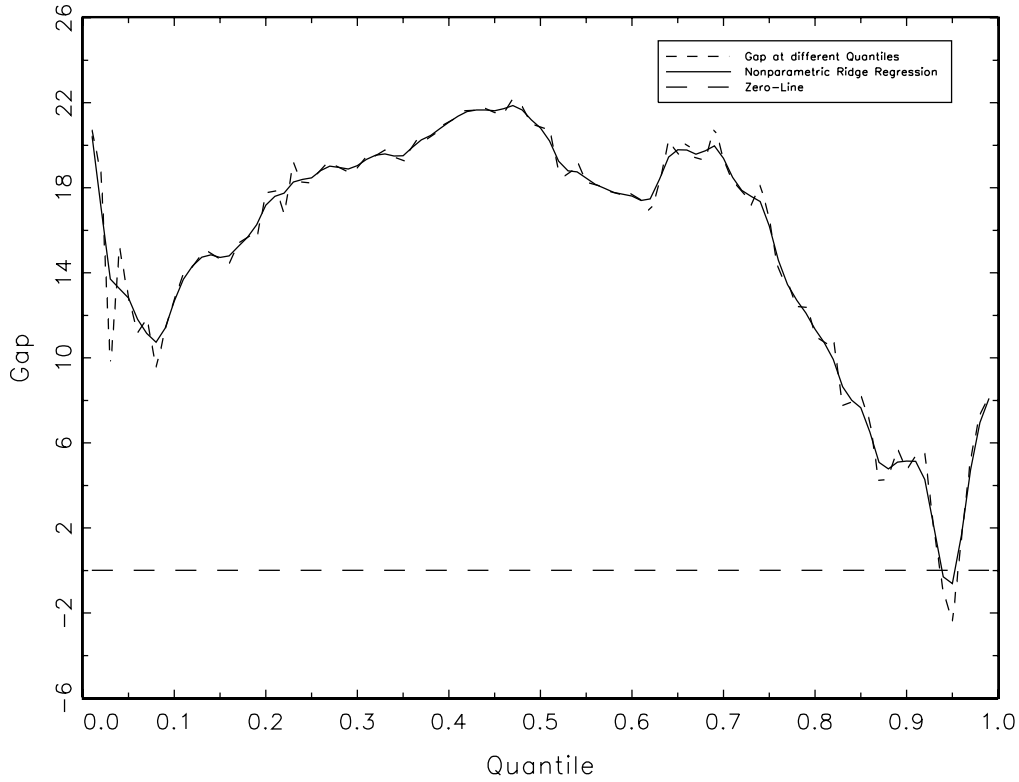


Figure 2.B.5: Histogram estimates of the propensity score distributions using student, family and school characteristics



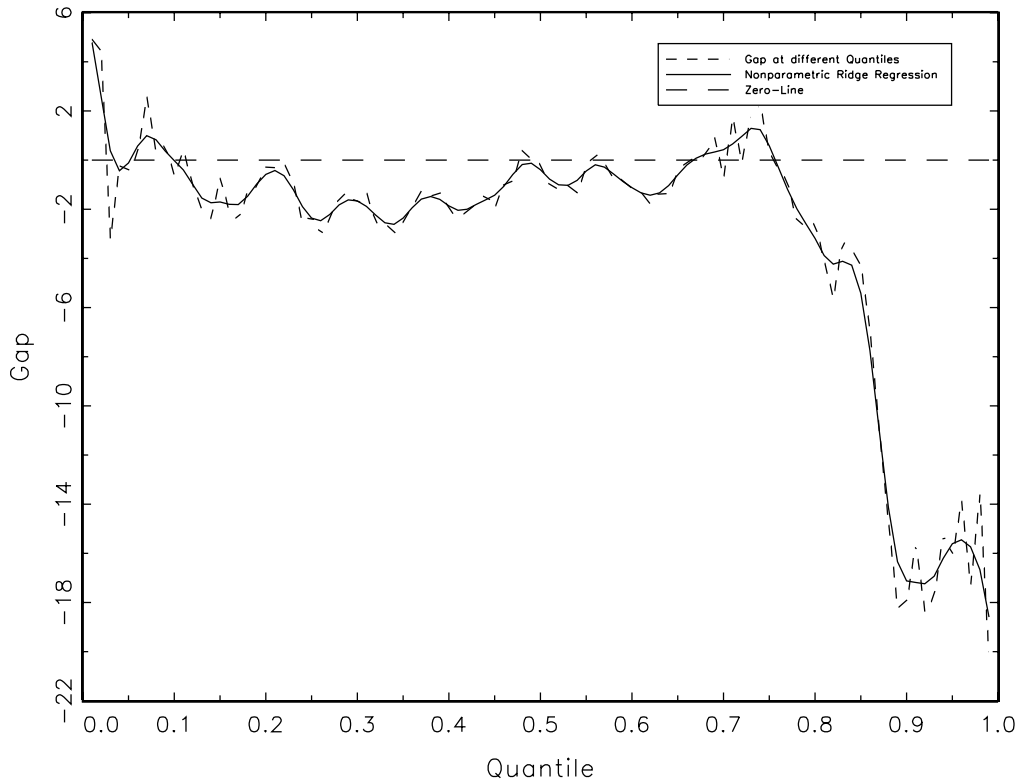
2. SEMIPARAMETRIC DECOMPOSITION OF THE GENDER ACHIEVEMENT GAP: AN APPLICATION TO TURKEY

Figure 2.B.6: Quantile test score gaps in science



Note: Nonparametric ridge regression with optimal cross validated bandwidth. The estimations are carried out for the common support subpopulation using sample weights provided in the data set.

Figure 2.B.7: Quantile test score gaps in math



Note: Nonparametric ridge regression with optimal cross validated bandwidth. The estimations are carried out for the common support subpopulation using sample weights provided in the data set.

CHAPTER 3

A Simple and Successful Method to Shrink the Weight

3.1 Introduction

In this paper we introduce a simple way of improving propensity score weighting and double robust estimators in terms of mean squared error (MSE) in finite samples. Our approach achieves a lower MSE by shrinking the propensity score towards the share of treated. This Stein-type simple shrinkage substantially mitigates the problems arising from propensity score estimates close to the boundaries. This reduces the variance of the weights and, therefore, the variance of the average treatment effect (ATE) estimators based on propensity score weighting. The shrinkage approach proposed is an attractive alternative to the popular trimming strategies applied to reduce the impact of large weights and can also be used jointly with trimming.

Even though shrinkage methods are very popular in other areas of statistics and econometrics, they have not been combined with weighting estimators yet. A notable exception is Frölich (2004), who uses the ridging method of Seifert and Gasser (1996) for matching estimators of the average treatment effect on the treated. They propose ridging of local polynomials to overcome the problems that arise in estimating a regression function when the conditional variance is unbounded.

The proposed shrinkage method is a linear combination of the conditional mean of the treatment variable and its unconditional mean. Like other shrinkage methods, the degree of shrinkage is determined by a tuning parameter. We propose three different methods to choose this parameter such that certain optimality conditions are satisfied. First, we consider a simple fixed valued tuning parameter, which only depends on the sample size. Second, we minimize the MSE of our linear combination to choose the optimal value. Third, we propose a pure cross validation procedure to obtain the optimal tuning parameter.

We demonstrate the MSE gains in finite samples via a comprehensive Monte Carlo study. To make our results comparable, we design our Monte Carlo study as in the settings of Busso et al. (2009) for poor overlap. We construct 72 settings to capture several possible issues when estimating the treatment effects and consider homogeneous and heterogeneous treatment, homoscedastic and heteroscedastic error terms as well as different ratios of treatment and control group. Moreover, the simulation design captures different functional forms. Since the shrunken propensity scores are constructed in such a way that they converge to the conventional propensity scores, our proposed method is asymptotically equivalent to standard approaches without shrinkage. Therefore, we focus on sample sizes 100, 200 and 500 only. Additionally, we evaluate the finite sample performance with and without applying trimming

rules.

Our results show that the estimators based on the shrunken propensity scores have a lower MSE than the weighting estimators based on the unshrunken propensity scores in all of the settings if we use the fixed valued or the MSE minimizing tuning parameter. For the cross validated tuning parameter, the MSE is reduced in 99.3% of the cases, respectively. If a trimming rule is applied to the proposed approach, we are able to decrease the MSE of the ATE in 99.7% of the cases for the fixed valued tuning parameter. For the MSE minimizing and cross validated tuning parameter, the MSE is reduced in 98.8% and 96.9% of the cases, respectively. In the rare cases where the MSE is not improved, the efficiency loss is very small.

The paper is organized as follows. Section 3.2 reviews the different weighting and double robust estimators. Section 3.3 introduces the shrunken propensity score and derives its properties in finite samples. In section 3.4, we present a Monte Carlo study and compare the MSE of the estimators based on the shrunken and classical propensity score weights. Section 3.5 concludes.

3.2 Propensity Score Methods

Consider the case of a binary treatment within Rubin’s (1974) potential outcome model.¹ Let Y_{1i} and Y_{0i} be the two potential outcomes for person i if she takes the treatment and if she does not take the treatment, respectively. D_i denotes the binary treatment indicator indicating whether person i participates in the program ($D_i = 1$) or not ($D_i = 0$). The observed outcome variable, Y_i , can than be written as a function of potential outcomes and the treatment variable:

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i} \quad \text{for } i = 1, \dots, n. \quad (3.1)$$

The difference between two potential outcomes of an individual, $Y_{1i} - Y_{0i}$, denotes the individual’s treatment effect. Depending on the realized treatment status, we only observe one of the two potential outcomes. Hence, the individual treatment effect cannot be identified from observed data. Under certain assumptions, however, we can still identify various average treatment effects. In this paper, we focus on the average treatment effect (ATE) defined as

$$\Delta_{\text{ATE}} = E[Y_{1i} - Y_{0i}], \quad (3.2)$$

¹See Imbens and Wooldridge (2009) for advantages of potential outcome model over observed outcome models.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

which measures the expected treatment effect if individuals are randomly assigned to treatment and control groups.

The identification of the ATE crucially depends on two assumptions. The first one is that, conditional on confounding variables, the potential outcomes are stochastically independent of the treatment: $Y_{0i}, Y_{1i} \perp D_i | X_i$, where X_i denotes the confounding variables of individual i . This assumption, known as the Conditional Independence Assumption (CIA), requires that all confounding factors associated with the potential outcomes as well as the participation decision are observed. If the CIA is satisfied, various estimation methods (e.g. weighting, regression and matching methods) are feasible to estimate the ATE.

The second assumption is the overlap assumption. It requires that the probability of receiving the treatment, the so-called propensity score, lies strictly between zero and one. In other words, each unit in a defined population has a positive probability of being treated and of not being treated. Although this type of overlap assumption is standard in the literature (e.g. Rosenbaum and Rubin (1983), Heckman et al. (1997), Hahn (1998), Wooldridge (2002), Imbens (2004)), there is a stronger version of the overlap assumption called “strict overlap” (e.g. Robins et al. (1994), Abadie and Imbens (2006), Crump et al. (2009)). Strict overlap requires that the probability of being treated is strictly between ξ and $1 - \xi$ for some $\xi > 0$. Khan and Tamer (2010) point out that a comparable assumption to the strict overlap assumption is needed for \sqrt{N} -convergence of some semiparametric estimators. Busso et al. (2009) provide further evidence on the importance of the (strict) overlap assumption.

Under the assumptions listed above, the ATE can be identified and estimated. There are several estimation methods proposed in the literature. We, however, focus only on the methods which use the propensity scores as weights. The propensity score, i.e. the probability of being treated conditional on the characteristics X_i , is given by

$$p_i = \Pr [D_i = 1 | X_i]. \quad (3.3)$$

As the propensity score is an unknown probability, it has to be estimated. Conventionally, standard parametric maximum likelihood methods are used to obtain the estimated propensity score and are denoted by \hat{p}_i .

Following Busso et al. (2009), we write the weighting type estimator for the ATE as

follows:

$$\hat{\Delta}_{ATE} = \frac{1}{n_1} \sum_{i=1}^n D_i Y_i \hat{\omega}_{i1} - \frac{1}{n_0} \sum_{i=1}^n (1 - D_i) Y_i \hat{\omega}_{i0}, \quad (3.4)$$

where n_1 is the number of treated observations and n_0 is the number of controls. $\hat{\omega}_{i0}$ and $\hat{\omega}_{i1}$ are defined differently for different types weighting estimators. We consider here three different weighting schemes proposed in the literature. The first one (IPW1) uses the following weighting functions:

$$\hat{\omega}_{i0}^{(1)} = \frac{n_0}{n} / (1 - \hat{p}_i) \quad (3.5)$$

$$\hat{\omega}_{i1}^{(1)} = \frac{n_1}{n} / \hat{p}_i, \quad (3.6)$$

where n is the total number of observations. The second weighting function (IPW2) results from an adjustment to force the weights to add up to one and is advocated by Imbens (2004). Formally, they are given by

$$\hat{\omega}_{i0}^{(2)} = \frac{1}{1 - \hat{p}_i} / \frac{1}{n_0} \sum_{i=1}^n \frac{1 - D_i}{1 - \hat{p}_i} \quad (3.7)$$

$$\hat{\omega}_{i1}^{(2)} = \frac{1}{\hat{p}_i} / \frac{1}{n_1} \sum_{i=1}^n \frac{D_i}{\hat{p}_i}. \quad (3.8)$$

The third weighting function (IPW3), which is not so common in the literature, is a combination of the first two methods, where the asymptotic variance of the resulting estimator is minimized for a known propensity score (see Lunceford and Davidian (2004) for details).

$$\hat{\omega}_{i0}^{(3)} = \frac{1}{1 - \hat{p}_i} (1 - C_{i0}) / \frac{1}{n_0} \sum_{i=1}^n \frac{(1 - D_i)}{1 - \hat{p}_i} (1 - C_{i0}) \quad (3.9)$$

$$\hat{\omega}_{i1}^{(3)} = \frac{1}{\hat{p}_i} (1 - C_{i1}) / \frac{1}{n_1} \sum_{i=1}^n \frac{D_i}{\hat{p}_i} (1 - C_{i1}) \quad (3.10)$$

with

$$C_{i0} = \frac{\frac{1}{1 - \hat{p}_i} \frac{1}{n} \sum_{i=1}^n \left(\frac{1 - D_i}{1 - \hat{p}_i} \hat{p}_i - D_i \right)}{\frac{1}{n} \sum_{i=1}^n \left(\frac{1 - D_i}{1 - \hat{p}_i} \hat{p}_i - D_i \right)^2} \quad (3.11)$$

$$C_{i1} = \frac{\frac{1}{\hat{p}_i} \frac{1}{n} \sum_{i=1}^n \left(\frac{D_i}{\hat{p}_i} (1 - \hat{p}_i) - (1 - D_i) \right)}{\frac{1}{n} \sum_{i=1}^n \left(\frac{D_i}{\hat{p}_i} (1 - \hat{p}_i) - (1 - D_i) \right)^2} \quad (3.12)$$

In all three cases, $\hat{\omega}_{i0}$ depends on $\frac{1}{1 - \hat{p}_i}$ and $\hat{\omega}_{i1}$ on $\frac{1}{\hat{p}_i}$. If the estimated propensity

score for individual i is close to one $\hat{\omega}_{i0}$, the weight for individual i , is large compared to the weights of the other observations. Therefore, the estimates are mainly determined by individual i . If the estimated propensity score is close to zero, $\hat{\omega}_{i1}$ is large, which again leads to an ATE estimator which exhibits high variance.

The doubly robust estimator of the ATE we consider here is derived from a weighted regression of the outcome model where the weights are inversely related to the propensity scores. The advantage of doubly robust estimation is that it stays consistent even if the outcome model or the propensity score model is specified incorrectly. It has been shown that doubly robust methods are more efficient than weighting methods (see, for example, Robins and Rotnitzky (1995), Wooldridge (2007)). Here, we consider the doubly robust method used by Hirano and Imbens (2001). They estimate the ATE by a weighted least square regression of the following outcome model with weights based on Equation (3.5):

$$Y_i = \alpha_0 + \Delta_{ATE}D_i + X_i'\alpha_1 + D_i(X_i - \bar{X})'\alpha_2 + \varepsilon_i \quad (3.13)$$

$$\hat{\omega}_i^{dr} = \sqrt{\frac{D_i}{\hat{p}_i} + \frac{1 - D_i}{1 - \hat{p}_i}}, \quad (3.14)$$

where \bar{X} is the sample average of X_i .

The weight $\hat{\omega}_i^{dr}$ again depends on $\frac{1}{1-\hat{p}_i}$ and $\frac{1}{\hat{p}_i}$, such that propensity scores close to one and zero have a similar effect as for the weighting estimators.

3.3 Shrunk Weights

A major drawback of the weighting and double robust estimators is that they can exhibit a high variance if the weights of some observations are very large. For the ATE, this is the case if the propensity score is close to one or zero. We propose different variants of Stein-type simple shrinkage methods for the propensity score which help to stabilize the treatment effect estimators by shrinking the propensity scores away from these boundaries.

The basic idea is to shrink the estimated propensity score, $\hat{p}_i = \hat{E}[D_i = 1|X_i = x]$, towards the estimated unconditional mean $\bar{D} = \frac{1}{n} \sum_i^n D_i$, i.e.

$$\hat{p}_i^s = (1 - \lambda_i(n))\hat{p}_i + \lambda_i(n)\bar{D}, \quad (3.15)$$

where $0 \leq \lambda_i(n) \leq 1$ is the tuning parameter, which may depend on the sample size. Equation (3.15) implies that our proposed shrunken propensity score is always closer to the share of treated and, therefore, the shrunken propensity scores have a lower variance than the conventional propensity scores. This enables us to estimate the treatment effects with a lower MSE.

Shrinking towards the unconditional mean avoids propensity scores close to one or zero. The usual way to deal with propensity scores that are too small or too large is to apply a trimming rule. A trimming rule basically determines an upper and a lower limit for the propensity score. Observations with propensity scores outside of the chosen limits are dropped from the estimation sample.² Obviously, dropping observations will cause both a loss of information as well as a loss of efficiency. Since shrinkage pushes the estimated propensity score away from the boundaries, we neither need to apply a trimming rule nor do we have to work with a reduced sample.

Shrinkage could also be applied directly to the parameter estimates of the propensity score by imposing a L_1 -norm (Lasso) or L_2 -norm (ridging) on the parameter estimates. Besides the higher computational burden, the interpretation of the shrinkage parameter in terms of the shrunken propensity score is, however, not straight forward. Moreover, our approach avoids shrinking a propensity score from one extreme to the other, i.e. we avoid shrinking a large propensity score $\hat{p}_i > \bar{D}$ towards zero, since $\hat{p}_i^s > \bar{D}$ always holds.

A crucial issue for any shrinkage estimator is the choice of the tuning parameter $\lambda_i(n)$. As we are interested in improving the small sample performance of weighting and double robust estimators, we propose to choose $\lambda_i(n)$ such that the penalty vanishes asymptotically. For $\lambda_i(n) = \mathcal{O}(n^{-\delta})$ with $\delta > 0$, the shrinkage estimator is consistent and converges to the true propensity score. For $\delta > 1/2$, the \hat{p}_i^s has the same asymptotic distribution as the conventional propensity score \hat{p}_i .

In the following, we consider three alternative methods of choosing $\lambda_i(n)$. The first method, the fixed tuning parameter method, is based on the functional form $\lambda_i(n) = \frac{c}{n^\delta}$. For a given values of c and δ , this method is easy to implement with no computational cost but is not optimized with respect to any MSE criterion.

In the second method, the MSE minimizing tuning parameter method, $\lambda_i(n)$ is determined by minimizing the MSE of the shrunken propensity score in (3.15).

²In the following section, we explain two trimming rules which are often used in the literature. For other trimming rules see Busso et al. (2009).

Thus, as shown in Appendix 3.C, the optimal $\lambda_i(n)$ is given by

$$\lambda_i^*(n) = \frac{V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D})}{V[\hat{p}_i] + \frac{E[D_i](1-E[D_i])}{n} + (E[D_i] - E[\hat{p}_i])^2 - 2\text{Cov}(\hat{p}_i, \bar{D})}, \quad (3.16)$$

where we assume $E[\hat{p}_i] \approx p_i$. Since $\lambda_i^*(n)$ depends on unknown parameters, we replace the squared bias, variance and covariances by their bootstrapped quantities. As for the tuning parameter in the first method, the MSE minimizing tuning parameter (3.16) also converges to zero as the sample size increases. Note that the MSE minimizing tuning parameter method yields different optimal λ 's for each observation in the sample. In order to stabilize the estimation resulting from the estimation noise, the estimated MSE minimizing tuning parameter can be replaced by the mean MSE minimizing tuning parameter, $\bar{\lambda}^*(n) = \frac{1}{n} \sum_{i=1}^n \lambda_i^*(n)$.³ Additionally, this guarantees that the ordering of the propensity scores does not change.

In the third method, the optimal λ is chosen by means of cross-validation. The idea is to minimize the mean squared prediction error of the estimated propensity score with respect to λ . The mean squared prediction error is calculated by leave-one-out cross validation for each λ in an equally spaced grid of $k+1$ λ 's, i.e. $[0, \lambda_{(1)}, \dots, \lambda_{(k-1)}, 1]$. The optimal λ is then the one which leads to the smallest cross validated mean squared prediction error. This method again yields a different λ for each setting, but is computationally less burdensome than the MSE minimizing choice of λ in the second method.

3.4 Monte Carlo Study

We demonstrate the efficiency gains due to propensity score shrinkage via a comprehensive Monte Carlo study. We adopt the same data generating processes as Busso et al. (2009) to make our results comparable with theirs. Since our approach shrinks the propensity score towards the treated-to-control ratio, it is especially valuable in situations where the overlap but not the strict overlap assumption is fulfilled. In the following, we, therefore, concentrate on those designs of Busso et al. (2009), which are not consistent with the strict overlap assumption. For the simulation study, D

³An alternative would be to choose λ such that the MSE of the vector of the shrunken propensity scores is minimized. The results are comparable to those obtained by using $\bar{\lambda}^*(n)$ and are available upon request.

and Y are generated as follows:

$$D_i = \mathbb{1}\{\eta + \kappa X_i - u_i > 0\} \quad (3.17)$$

$$Y_i = D_i + m(p(X_i)) + \gamma D_i m(p(X_i)) + \varepsilon_i, \quad (3.18)$$

where the error terms u_i and ε_i are independent of each other and the confounding variable X_i which is assumed to be a standard normally distributed random variable. $p(X_i)$ is the propensity score and $m(\cdot)$ is a function of the propensity score. We use two different functions in the Monte Carlo study given in Table 3.1.

Table 3.1: Functional form for $m(q)$

$m(q)$	Formula	Description
$m_1(q)$	$0.15 + 0.7q$	Linear
$m_2(q)$	$0.2 + \sqrt{1-q} - 0.6(0.9 - q)^2$	Nonlinear

The error term u_i is drawn from a standard normal distribution leading to the following propensity score function:

$$p(X_i) = \Phi(\eta + \kappa X_i). \quad (3.19)$$

We generate various treated-to-control ratios by choosing three different combinations for η and κ . Table 3.2 summarizes the parameter values and resulting ratios.

Table 3.2: Treated-to-control ratios

η	κ	Treated-to-control ratio
0	0.95	1:1
0.3	-0.8	3:2
-0.3	0.8	2:3

The error term in the outcome equation, ε_i , is specified as

$$\varepsilon_i = \psi(e_i p(X_i) + e_i D_i) + (1 - \psi)e_i, \quad (3.20)$$

where e_i is iid standard normal random variable and ψ is a parameter which controls heteroscedasticity, i.e. for $\psi = 0$, ε_i is a homoscedastic error term and if $\psi \neq 0$, ε_i is heteroscedastic. By choosing different values of γ , we specify whether the treatment effect is homogeneous or not. Treatment homogeneity implies that the treatment effect does not vary with different X 's. In this case, the causal effect of the treatment is the same for all individuals. As in Busso et al. (2009), we use the following

combinations of ψ and γ to create four different settings.

Table 3.3: Parameter combinations

γ	ψ	Description
0	0	homogeneous treatment, homoscedastic
1	0	heterogeneous treatment, homoscedastic
0	2	homogeneous treatment, heteroscedastic
1	2	heterogeneous treatment, heteroscedastic

Our simulations are based on 10,000, 5,000 and 2,000 Monte Carlo samples for sample sizes $n = 100$, $n = 200$ and $n = 500$, respectively. The choice to make the number of replications proportional to the sample size is motivated by the fact that simulation noise depends negatively on the number of replications and positively on the variance of the estimators (see Huber et al. (2013)), which again depends negatively on the chosen sample size. Hence, the simulation noise is constant if the Monte Carlo samples are chosen proportional to the sample size. Our Monte Carlo study consists of two parts. In the first part, we apply the methods without using any trimming rules. In the second part, we incorporate two different trimming rules to the conventional as well as shrunken propensity scores.

Propensity Score Shrinkage without Trimming

To provide a reference point for the optimal choices of λ in the different settings without applying any trimming rules, we perform a Monte Carlo study for a hypothetical case where the true ATE is known. Due to the computational burden of this procedure, we do this only for one specification which we believe is the most realistic one and only for sample size $n = 100$. This specification allows heteroscedasticity in the error term ($\psi = 2$) and heterogeneity in the treatment effect ($\gamma = 1$). Furthermore, we also consider the most challenging treated control ratio where we have more control units than treated units ($\eta = -0.3$, $\kappa = 0.8$). Lastly, the outcome equation is chosen to be a nonlinear function of the propensity score ($m_2(q)$). We apply the following procedure to get the optimal λ for known ATE:

1. We draw 10000 Monte Carlo samples for this specification.
2. For each Monte Carlo sample, we estimate the shrunken propensity scores for $\lambda = 0, 0.01, 0.02, \dots, 1$ and the ATE by the four methods with each of these shrunken propensity scores.
3. We calculate the MSE over 10000 Monte Carlo samples for each λ and choose the MSE minimizing λ .

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

4. Steps (1)-(3) are repeated 500 times.

The minimum, mean, maximum and standard error of the mean over 500 optimal λ 's are displayed in Table 3.4.

Table 3.4: Descriptive statistics for the optimal λ with known ATE

	IPW1	IPW2	IPW3	DR
Min	0.06	0.09	0.16	0.08
Mean	0.82	0.24	0.33	0.35
Max	1.00	0.45	0.50	0.66
Std. Err.	0.028	0.005	0.005	0.009

Note: The MSE minimizing λ^* 's are obtained from a Monte Carlo study for the specification with $n = 100$, $\gamma = 1$, $\psi = 2$, $\eta = -0.3$, $\kappa = 0.8$ and $m_2(q)$. We use 10000 Monte Carlo replications and replicate this procedure 500 times.

The results show that most shrinkage is required for IPW 1 and the least for IPW 2. In all of the 500 replications, λ is never chosen equal to zero, which implies that shrinkage is always optimal.

As in Busso et al. (2009), we estimate the ATE given in Equation (3.2) for each possible DGP by all three weighting methods and the doubly robust method reviewed in Section 3.2 using estimated (unshrunk) propensity score, \hat{p}_i . \hat{p}_i is obtained by maximum likelihood probit estimation as suggested by the distribution of the error term u_i . Additionally, we estimate the ATEs using the shrunk propensity score, \hat{p}_i^s . The optimal tuning parameter λ is chosen in three different ways as introduced in Section 3.2. The goal is to demonstrate the gains in terms of MSE reduction of the ATE due to propensity score shrinkage as well as to investigate the relative performance of the different shrinkage methods. For the fixed tuning parameter method, we set $c = 1$ and $\delta = 1/2$, i.e. $\lambda_i(n) = 1/\sqrt{n}$. As a summary statistic, we also report the averages over sample sizes in bold letters. The figures in brackets indicate percentage losses due to the bias introduced by shrinkage, $\left(\frac{\text{bias}^2(\text{ATE}(\hat{p})) - \text{bias}^2(\text{ATE}(\hat{p}^s))}{\text{bias}^2(\text{ATE}(\hat{p})) + \text{Var}(\text{ATE}(\hat{p}))} \right)$.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.5: Average percentage improvement in MSE for the ATE

	ATE based on \hat{p}_i vs. ATE based on \hat{p}_i^* using											
	$\lambda = 1/\sqrt{n}$				$\lambda = \text{argmin MSE}$				$\lambda = \text{cross-validated}$			
	100	200	500	avg.	100	200	500	avg.	100	200	500	avg.
IPW1	47.4	51.6	37.4	45.5	54.9	61.9	42.3	53.0	8.3	5.1	2.4	5.2
	(-4.0)	(-3.8)	(-6.1)	(-4.7)	(-3.1)	(-3.9)	(-6.2)	(-4.4)	(-0.8)	(-0.8)	(-0.4)	(-0.7)
IPW2	16.5	18.6	18.8	18.0	16.6	18.3	21.7	18.9	5.3	3.7	2.6	3.9
	(-2.1)	(-2.4)	(-3.1)	(-2.6)	(-2.4)	(-2.5)	(-2.6)	(-2.5)	(-0.7)	(-0.4)	(-0.3)	(-0.4)
IPW3	6.7	5.9	4.8	5.8	6.8	6.2	5.2	6.0	3.0	2.0	1.4	2.1
	(-0.9)	(-0.9)	(-1.1)	(-1.0)	(-1.1)	(-1.0)	(-0.9)	(-1.0)	(-0.4)	(-0.2)	(-0.2)	(-0.3)
DR	5.7	6.4	7.5	6.5	5.6	6.6	7.8	6.6	2.5	2.0	1.7	2.1
	(0.0)	(0.0)	(-0.1)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.0)	(0.0)	(0.0)	(0.0)	(0.0)

Note: avg. denotes the average over the sample sizes. Percentage change which is due to the bias is given in brackets.

Table 3.5 reveals that, independent of the choice of λ , the improvement turns out to be more pronounced for IPW1 and IPW2, i.e. for those methods which are most vulnerable to very small or very large propensity scores. This result is especially striking since most estimates in the empirical literature are probably based on IPW2 (Busso et al. (2009)). Using the fixed tuning parameter method, the MSE of this estimator could be improved by 18.0% on average. Basically, this improvement comes at no costs due to the simplicity of the linear combination.

The MSE minimizing λ is chosen as in Equation (3.16) for each individual i . Figures 3.B.1 - 3.B.3 of Appendix 3.B plot the individual specific $\hat{\lambda}_i^*(n)$. It can be seen that the estimated MSE-minimal tuning parameter exhibits a high variation across observations. In small samples the λ_i 's vary strongly over individuals. Therefore, we shrink the propensity score using the average over all observations. The computationally more burdensome MSE-minimizing λ leads to a 18.9% improvement for IPW2. For both choices of λ , the average improvement of IPW3 and DR is still 6.0 to 6.6 percent. We see that the improvement is due to a large reduction of the variance but comes at the expense of introducing a comparatively small bias. For DR, the increase in the squared bias is nearly zero.

If we compare the average results in Table 3.5 obtained by the fixed tuning parameter method to the ones obtained from MSE minimization, we find that MSE-minimization yields better results for $n = 500$. For $n = 200$ and $n = 100$ this is the case for IPW1. For the other estimators, both methods give about the same result. On average, the cross-validated λ also yields a reduction of the MSE in all cases but is always dominated by the other two choices of λ .

The detailed simulation results for the first part are given in Tables 3.A.1 - 3.A.3 of Appendix 3.A. Tables 3.A.1 and 3.A.2 show that, in all 288 cases, the use of shrunken propensity scores leads to an improvement of the MSE of the ATE if the

fixed valued λ or the MSE-minimizing λ is chosen. Table 3.A.3 shows that the use of the shrunken propensity score leads to an improvement in 99.3% of the MSE comparisons if the cross-validated λ is taken.

Propensity Score Shrinkage with Trimming

In the second part of the Monte Carlo study, we evaluate the performance of the proposed shrinkage methods in combination with two trimming rules. Trimming rules are methods for the propensity score are usually applied to avoid the problems occurring if the propensity scores are close to the boundaries. From the various trimming rules proposed in the literature, we consider the two trimming rules which are most commonly used in empirical work and revealed the best performance in the study by Busso et al. (2009). These trimming rules are applied as follows:

1. The first trimming rule goes back to a suggestion by Dehejia and Wahba (1999). Let $T_i^{ATE} = \mathbb{1}(\hat{a} < \hat{p}(X_i) < \hat{b})$ setting \hat{b} to be the k^{th} largest propensity score in the control group and \hat{a} to be the k^{th} smallest propensity score in the treatment group. Then the estimators are computed based on the subsample for which $T_i^{ATE} = 1$.
2. In the second trimming rule suggested by Crump et al. (2009), all units with an estimated propensity score outside the interval $[0.1; 0.9]$ for the ATE are discarded.

As in the first part, we estimate the propensity scores by probit and shrink the propensity scores with the optimal λ 's chosen by the three different methods we propose. Different from the first part, we apply the trimming rules 1 and 2 to the conventional and trimming rule 2 to the shrunken propensity scores before estimating the ATEs by weighting and doubly robust methods. Finally, we compare the results based on the shrunken propensity score combined with trimming rule 2 with the results based on: (i) conventional propensity score, (ii) conventional propensity score combined with trimming rule 1, and (iii) conventional propensity score combined with trimming rule 2. As mentioned before, applying the trimming rules to the shrunken propensity score leads to a smaller reduction in the sample size since less observations lie outside the limits of the two trimming rules. For example, if we apply trimming rule 2 to the shrunken propensity score with fixed tuning parameter $\lambda(n) = 1/\sqrt{100} = 0.1$ in the setting where the treated-to-control ratio is 1:1, we use all observations with an conventional propensity score in the interval $[0.055; 0.944]$ instead of only those in the interval $[0.1; 0.9]$. We, therefore, still throw less information away than in the case where the unshrunken propensity scores are trimmed. The estimators based on this procedure converge to the estimators based

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

on the conventional propensity scores which are then trimmed using trimming rule 2.

Again, we perform a Monte Carlo experiment for the most realistic scenario to see which λ would be chosen by cross validation if the true ATE would be known. The results are displayed in Table 3.6.

Table 3.6: Descriptive statistics for the optimal λ for known ATE with trimming rule 2

	IPW1	IPW2	IPW3	DR
Min	0.12	0.14	0.16	0.07
Mean	0.16	0.26	0.32	0.17
Max	0.19	0.46	0.50	0.42
St. Dev.	0.01	0.05	0.06	0.03

Note: The MSE minimizing λ^* 's are obtained from a Monte Carlo study for the specification with $n = 100$, $\gamma = 1$, $\psi = 2$, $\eta = -0.3$, $\kappa = 0.8$ and $m_2(q)$. We use 10000 Monte Carlo replications and replicate this procedure 500 times.

In this case, most shrinkage is required for IPW 3 and the least for IPW 1. Again in none of the 500 replications λ is chosen equal to zero, implying that, for this estimation procedure, it also is always optimal to have shrinkage. If we compare the maximum λ 's in Table 3.6 to Table 3.4, we see that, especially for IPW 1, the degree of shrinkage is a lot smaller if we use trimming rule 2 after shrinking the propensity score.

The results for the fixed tuning parameter method, $\lambda_i(n) = 1/\sqrt{n}$, are given in Tables 3.A.4 - 3.A.6 of Appendix 3.A and summarized in Table 3.7 below, which contains the average MSE improvements. If we compare our estimation procedure to the estimators based on the conventional propensity scores, the largest percentage improvement in MSE can be obtained for the simple IPW1 estimator. For this weighting estimator, we obtain an improvement of up to 75.7%. When averaging over all settings, the MSE of this estimator is improved by 46.2%. The second largest improvement is obtained for the popular estimator IPW2. The weights of this estimator, compared to IPW1, are forced to add up to one and the average improvement here is 18.7%. IPW3, the estimator that minimizes the asymptotic variance for a known propensity score, can, on average, still be improved by 8.8% in terms of MSE. The MSE of the DR estimator can be reduced by 11.0% if propensity score shrinkage with trimming rule 2 is applied instead of using the unshrunk propensity scores.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.7: Average percentage improvement in MSE, fixed valued λ

ATE based on the shrunken propensity scores \hat{p}_i^s + trimming rule 2 vs.												
	(a) ATE based on \hat{p}_i				(b) ATE based on \hat{p}_i + trimming rule 1				(c) ATE based on \hat{p}_i + trimming rule 2			
	100	200	500	avg.	100	200	500	avg.	100	200	500	avg.
IPW1	45.0	51.6	42.1	46.2	25.8	16.3	13.0	18.4	20.5	15.1	9.9	15.2
	(-0.6)	(-0.5)	(-0.6)	(-0.6)	(-1.0)	(-0.9)	(-0.6)	(-0.8)	(-0.9)	(-0.3)	(0.7)	(-0.2)
IPW2	14.2	18.6	23.2	18.7	13.1	8.2	7.7	9.7	9.0	5.9	3.9	6.3
	(-0.5)	(-0.7)	(-1.2)	(-0.8)	(-0.4)	(-0.5)	(-1.1)	(-0.7)	(-0.4)	(-0.2)	(0.1)	(-0.1)
IPW3	6.7	8.2	11.5	8.8	12.8	7.6	6.7	9.1	6.6	4.3	2.9	4.6
	(-0.2)	(-0.5)	(-1.2)	(-0.6)	(-0.1)	(-0.3)	(-0.9)	(-0.4)	(-0.1)	(0.1)	(0.3)	(0.1)
DR	8.9	10.2	14.0	11.0	10.7	5.7	5.5	7.3	3.8	2.7	1.8	2.8
	(-0.1)	(-0.4)	(-1.2)	(-0.6)	(0.1)	(-0.1)	(-0.8)	(-0.3)	(0.1)	(0.2)	(0.4)	(0.2)

Note: avg. denotes the average over the sample sizes. Percentage change which is due to the bias is given in brackets.

Table 3.7 part (b) shows that the improvements for IPW1 and IPW2 are smaller than in part (a). For IPW3 and DR, the improvements are smaller in part (b) for sample sizes 200 and 500. Nevertheless, the estimators based on the shrunken propensity scores combined with trimming rule 2 improve IPW1 by 18.4%, IPW2 by 9.7%, IPW3 by 9.0% and DR by 7.3% on average.

The results in Table 3.7 part (c) reveal that the improvements for all four estimators are smaller than the improvement in part (a) and (b). The average improvements for the four estimators is between 2.8% and 15.2%. Even though we obtain a smaller improvement of the MSE, the suggested procedure reduces the variance and bias for all four estimators for $n = 500$. This even holds for all sample sizes for the double robust estimator. Moreover, the improvement is smaller for larger sample sizes. This is expected since, for large n , the estimator based on the shrunken weights combined with trimming rule 2 converge to the estimators based on the conventional estimators and trimming rule 2.

All in all, the four estimators, based on the conventional propensity score, never have a lower MSE than estimators using shrunken propensity scores combined with trimming rule 2 (see Table 3.A.4) in 72 settings. The estimators based on the conventional propensity scores combined with trimming rule 1 yield a lower MSE than the estimators based on our procedure (see Table 3.A.5) only once. However, the increase in MSE was only 0.04%. Only in 2 out of 288 cases was the MSE of the estimator based on the conventional propensity scores combined with trimming rule 2 smaller than our suggested procedure. Furthermore, the detailed results in Table 3.A.6 show that the losses in MSE in those two cases are only 0.1% and 0.5% .

For the fixed tuning parameter method, the effects of propensity score shrinkage on the distribution of the estimated ATE's can be seen from the boxplots given in

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Figures 3.B.4 - 3.B.6 of Appendix 3.B. We compare our method to the estimators based on the conventional propensity score. The introduction of propensity score shrinkage does not significantly change the interquartile ranges of the four estimators compared to the estimates without shrinkage. However, the number of outliers is substantially reduced by shrinkage. This holds in particular for the IPW1 estimator, which suffers from a high number of very large outliers and explains why the MSE gains due to propensity score shrinkage are largest for this estimator. Moreover, note that even for small sample sizes, propensity score shrinkage hardly generates any additional bias compared to the estimators without shrinkage.

Thus far, we simply set $\lambda_i(n) = 1/\sqrt{n}$. This choice of $\lambda_i(n)$ yields $\lambda_i(100) = 0.100$, $\lambda_i(200) = 0.071$ and $\lambda_i(500) = 0.045$ for all four estimators. If we compare those with optimal λ 's for known ATE in Table 3.6, we see that the improvements are obtained with $\lambda_i(n)$'s which are considerably lower than the optimal λ 's.

In this part, we use the average over the $\lambda_i(n)$, which minimize the MSEs of the shrunken propensity scores, for each setting. The results based on these $\bar{\lambda}^*(n)$'s are given in Tables 3.A.7 - 3.A.9 of Appendix 3.A. The chosen $\bar{\lambda}^*(n)$ depend on the sample size but, as they minimize the MSE of the shrunken propensity score, and, therefore, do not consider the second stage of the estimation procedure, they are equal for all four estimators. Table 3.8 below summarizes the average percentage improvements obtained by estimating the ATE based on the shrunken propensity combined with trimming rule 2.

Table 3.8: Average percentage improvement in MSE, $\text{MSE}(\hat{p}_i^s)$ -minimizing λ

	ATE based on the shrunken propensity scores \hat{p}_i^s + trimming rule 2 vs.											
	(a) ATE based on \hat{p}_i				(b) ATE based on \hat{p}_i + trimming rule 1				(c) ATE based on \hat{p}_i + trimming rule 2			
	100	200	500	avg.	100	200	500	avg.	100	200	500	avg.
IPW1	52.1	62.1	46.1	53.4	26.3	15.2	12.7	18.1	20.9	13.4	10.8	15.0
	(-0.8)	(-0.5)	(-0.6)	(-0.6)	(-1.2)	(-1.2)	(-0.8)	(-1.0)	(-1.0)	(-0.7)	(0.5)	(-0.4)
IPW2	14.5	18.6	25.1	19.4	13.9	7.9	6.3	9.4	8.6	5.5	3.3	5.8
	(-0.7)	(-0.6)	(-1.3)	(-0.9)	(-0.6)	(-0.6)	(-1.4)	(-0.9)	(-0.6)	(-0.2)	(0.0)	(-0.3)
IPW3	6.8	8.9	11.5	9.1	13.9	7.1	5.3	8.8	6.5	3.8	2.3	4.2
	(-0.3)	(-0.4)	(-1.3)	(-0.7)	(-0.3)	(-0.3)	(-1.1)	(-0.6)	(-0.2)	(0.0)	(0.3)	(0.0)
DR	8.8	10.8	14.1	11.3	11.9	5.0	4.0	7.0	3.8	2.0	1.1	2.3
	(-0.2)	(-0.3)	(-1.2)	(-0.6)	(0.1)	(-0.1)	(-1.0)	(-0.3)	(0.2)	(0.2)	(0.4)	(0.2)

Note: avg. denotes the average over the sample sizes. Percentage change which is due to the bias is given in brackets.

Table 3.8 shows that $\text{MSE}(\hat{p}_i^s)$ -minimizing λ leads to a considerable reduction in the MSE of the treatment effect. Table 3.8 part (a) shows that the MSE's of the ATE estimators based on $\text{MSE}(\hat{p}_i^s)$ -minimizing λ are on average between 6.8% and 62.1% smaller than those based on conventional propensity score.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Comparing the average results from Table 3.8 to those obtained by the fixed value λ in Table 3.7, we see that both choices of λ give about the same result. For $n = 100$, the average value of the MSE-minimizing λ over the 10000 Monte Carlo samples was 0.113. In 0.05% of the cases, the restriction $0 \leq \bar{\lambda}^*(n) \leq 1$ is binding for $n = 100$. For $n = 200$ ($n = 500$), the average value is 0.076 (0.046), the minimum is 0 (0.025) and the maximum 0.429 (0.100). For $n = 200$, $\bar{\lambda}^*(n)$ is set to zero in 0.01% of the cases and never set to one. For $n = 500$, it is never set to zero or one. These numbers highlight that shrinkage can be a useful tool especially for small sample sizes.

Since the average $\lambda^*(n)$ over the 72 different settings is larger than the fixed value λ for each sample size (Table 3.9), this choice of λ implies on average more shrinkage. If we compare the two resulting λ s for $n = 100$ to the optimal λ for known ATE in Table 3.6, we see that the MSE minimizing λ is closer to these true λ 's which explains the slightly higher MSE gains.

The pattern of the MSE reductions by our procedure with respect to the conventional propensity scores combined with trimming rule 2 is analogous to the pattern with respect to the conventional propensity scores combined with trimming rule 1.

The detailed results in Tables 3.A.7 - 3.A.9 show that IPW1 is improved by up to 86.0%, IPW2 up to 38.6%, IPW3 up to 20.9%. For DR, the largest decrease in MSE is 24.1%. Out of the 864 cases (72 settings for four estimators compared to three alternatives), our procedure yields an improvement in the MSE 854 times. In the other 10 cases, the average increase in MSE is only 0.925%.

Next, we use the cross-validated alternative to choose the optimal λ . Table 3.9 summarizes the λ 's chosen by all three methods proposed:

Table 3.9: Average values for λ obtained through the different methods.

	100	200	500	avg.
$\lambda = 1/\sqrt{n}$	0.100	0.071	0.045	0.072
$\lambda = \text{argmin MSE}$	0.113	0.076	0.046	0.078
$\lambda = \text{cross-validated}$	0.061	0.029	0.013	0.034

Note: avg. denotes the average over the sample sizes.

For the fixed valued λ method, the values are independent of the different designs except for the sample size. For the MSE minimizing λ and the cross-validated λ , we obtain different values for each setting. In these cases, Table 3.9 reports the average values of λ over the different settings for each sample size. In 7.3% of the Monte Carlo samples with $n = 100$, cross-validated λ is equal to 0 and the largest optimal

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

λ is 0.89. For $n = 200$ ($n = 500$), the respective values are 15.3% (31.1%) and 0.13 (0.06). As for the other two methods, these numbers indicated that less shrinkage is optimal for larger sample sizes.

It turns out that the optimal choice of λ by cross-validation is lower than the λ 's chosen by the other methods. Thus, the cross-validated λ is not as close to the optimal λ for the known ATE given in Table 3.6 and lower MSE gains are expected.

The results based on these λ 's are given in Tables 3.A.10 - 3.A.12 of Appendix 3.A. Table 3.10 below summarizes the average percentage improvements. The results show that, on average, the cross-validation method also leads to a gain in the MSE of the ATEs for all four estimators. However, we see that the fixed valued λ 's (Table 3.7) and the MSE-minimizing method (Table 3.8) provide larger average improvements in the MSEs of the ATEs.

Table 3.10: Average percentage improvement in MSE, cross-validated λ

ATE based on the shrunken propensity scores \hat{p}_i^s + trimming rule 2 vs.												
	(a) ATE based on \hat{p}_i				(b) ATE based on \hat{p}_i + trimming rule 1				(c) ATE based on \hat{p}_i + trimming rule 2			
	100	200	500	avg.	100	200	500	avg.	100	200	500	avg.
IPW1	41.0	40.0	43.5	41.5	17.4	10.5	5.3	11.1	12.1	7.3	4.0	7.8
	(-0.1)	(0.0)	(-0.6)	(-0.2)	(-0.2)	(-0.1)	(-0.7)	(-0.4)	(0.1)	(0.2)	(0.7)	(0.3)
IPW2	12.1	15.5	20.5	16.0	9.5	5.6	4.4	6.5	4.4	2.8	1.7	3.0
	(-0.2)	(-0.3)	(-1.2)	(-0.6)	(-0.1)	(-0.2)	(-1.0)	(-0.4)	(0.0)	(0.2)	(0.2)	(0.1)
IPW3	5.0	5.9	8.0	6.3	10.3	5.9	3.7	6.7	3.0	1.9	1.1	2.0
	(-0.1)	(-0.3)	(-1.3)	(-0.6)	(0.0)	(-0.2)	(-1.0)	(-0.4)	(0.0)	(0.1)	(0.2)	(0.1)
DR	7.9	8.9	11.4	9.4	9.2	4.8	3.0	5.7	1.4	1.0	0.7	1.0
	(-0.2)	(-0.4)	(-1.4)	(-0.7)	(0.1)	(-0.2)	(-1.1)	(-0.4)	(0.1)	(0.1)	(0.1)	(0.1)

Note: avg. denotes the average over the sample sizes. Percentage change which is due to the bias is given in brackets.

Using the cross validation method to determine the optimal λ , the improvement in MSEs can be split into the cases where $\lambda = 0$, e.g. no shrinkage, is chosen and into the cases where $\lambda > 0$ is chosen. If we only look at the Monte Carlo samples where at least some shrinkage is chosen, we obtain larger improvements in MSEs. In those cases where λ is set to zero, the MSEs of the estimators are obviously equal.

As reported in Table 3.10 part (c), even though the gains are less pronounced, our procedure not only reduces the variance but, on average, also leads to a lower squared bias for all four estimators and all sample sizes.

The detailed results in Tables 3.A.10 - 3.A.12 show that IPW1 is improved by up to 81.3%, IPW2 up to 42.0%, IPW3 up to 18.8%. For DR, the largest MSE-reduction is 23.8%. In 18 out of the 288 cases, propensity score shrinkage fails to outper-

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

form the estimators based on conventional propensity score estimates. Traditional estimates based on the conventional propensity score combined with trimming rule 1 outperform our shrinkage approaches in only 9 out of the 288 cases. Moreover, failure of MSE reduction due to shrinkage are not only rare, but the losses are also small in magnitude. The average MSE increase over these 27 cases where our procedure is outperformed is only 1.3%. If we compare our procedure to the conventional propensity score combined with trimming rule 2, we see from Table 3.A.12 that in all of the 288 cases our procedure yields a lower MSE.

To summarize our findings, we explicitly look at the most realistic setting described before. We focus, thereby, on our suggested procedure and its asymptotic equivalent. The results are given in Table 3.11:

Table 3.11: Average percentage improvement in MSE for the most realistic setting and different λ s.

	ATE based on \hat{p}_i + trimming rule 2 vs. ATE based on \hat{p}_i^s + trimming rule 2 using											
	$\lambda = 1/\sqrt{n}$				$\lambda = \text{argmin MSE}$				$\lambda = \text{cross-validated}$			
	100	200	500	avg.	100	200	500	avg.	100	200	500	avg.
IPW 1	20.5	15.2	10.4	15.4	20.5	14.4	10.3	15.1	13.6	9.1	6.1	9.6
	(-1.2)	(-0.4)	(0.5)	(-0.4)	(-1.7)	(-1.2)	(0.1)	(-0.9)	(-0.2)	(-0.1)	(1.1)	(0.3)
IPW 2	6.3	4.2	3.2	4.6	5.6	4.0	3.1	4.2	2.7	2.3	1.1	2.0
	(0.1)	(0.4)	(0.6)	(0.4)	(0.1)	(0.3)	(0.7)	(0.4)	(0.2)	(0.1)	(0.2)	(0.2)
IPW 3	5.3	3.7	2.8	3.9	4.8	3.2	2.6	3.5	2.2	1.9	0.7	1.6
	(0.1)	(0.3)	(0.5)	(0.3)	(0.1)	(0.3)	(0.6)	(0.3)	(0.1)	(0.1)	(0.2)	(0.1)
DR	4.1	2.9	2.3	3.1	3.8	2.6	1.7	2.7	1.7	1.6	0.6	1.3
	(0.1)	(0.2)	(0.2)	(0.2)	(0.1)	(0.2)	(0.3)	(0.2)	(0.1)	(0.0)	(0.1)	(0.1)

Note: avg. denotes the average over the sample sizes. Simulation for the specification with $\delta = 1$, $\psi = 2$, $\eta = -0.3$, $\kappa = 0.8$ and $m_2(q)$. Average percentage change which is due to the bias is given in brackets.

Table 3.11 shows that, like for the average results, the cross validated λ leads to the smallest improvements. Comparing the results based on the fixed valued λ to those based on the MSE minimizing λ , we again find that the improvements are very similar. Moreover, for this setting, the MSE of the ATE is not only reduced due to a variance reduction but also due to a lower squared bias with the exception of IPW1 for $n = 100$ and $n = 200$.

Like for the average results, we also see that the gain of the MSE of the ATE is largest for small samples and that the sample size increases as the gains decrease. Asymptotically, the ATE based on the shrunken propensity score is equal to the ATE based on the conventional propensity score.

Table 3.12 below summarizes our results by means of a linear regression. We regress the average percentage MSE improvement on the features of our data generating process represented by a set of dummy variables. The results show that the improvement

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

is less pronounced if the error term is heteroscedastic (Set3 and Set4). The MSE improvement of the weighting estimators is larger when the outcome equation depends on the propensity score in a nonlinear way. Moreover, there is a tendency towards the improvement being higher when the treated-to-control ratio is balanced (Ratio1). If we look at the influence of the sample size on the MSE improvement, we see that compared to part (a) the improvement is higher for larger sample sizes. If the conventional propensity score is combined with a trimming rule, the improvement is significantly lower for larger sample sizes.

Table 3.12: Regression of the average percentage improvement in MSE for $N = 72$ settings on a set of dummy variables describing the setting.

Average percentage MSE improvement for ATE based on \hat{p}_i^s + trimming rule 2 vs.												
Variables	(a) ATE based on \hat{p}_i				(b) ATE based on \hat{p}_i + trimming rule 1				(c) ATE based on \hat{p}_i + trimming rule 2			
	IPW1	IPW2	IPW3	DR	IPW1	IPW2	IPW3	DR	IPW1	IPW2	IPW3	DR
Cons	54.2 (2.3)	23.6 (0.6)	12.2 (0.6)	14.9 (0.6)	28.3 (0.5)	15.7 (0.5)	15.0 (0.5)	12.1 (0.5)	21.8 (0.7)	9.8 (0.6)	6.5 (0.4)	2.9 (0.3)
Set2	2.8 (2.2)	-1.0 (0.5)	-1.6 (0.6)	-1.2 (0.6)	3.8 (0.5)	-0.2 (0.5)	-0.2 (0.5)	-0.5 (0.5)	3.5 (0.7)	0.7 (0.5)	0.7 (0.4)	0.5 (0.3)
Set3	-13.7 (2.2)	-13.1 (0.5)	-9.6 (0.6)	-9.4 (0.6)	-5.6 (0.5)	-4.6 (0.5)	-3.7 (0.5)	-1.2 (0.5)	-3.9 (0.7)	-2.9 (0.5)	-1.1 (0.4)	1.5 (0.3)
Set4	-10.0 (2.2)	-13.7 (0.5)	-10.6 (0.6)	-10.2 (0.6)	-2.7 (0.5)	-4.7 (0.5)	-3.8 (0.5)	-1.5 (0.5)	-1.3 (0.7)	-2.4 (0.5)	-0.6 (0.4)	1.8 (0.3)
Curve2	4.1 (1.5)	1.0 (0.4)	0.7 (0.4)	0.2 (0.4)	0.9 (0.3)	1.3 (0.4)	0.8 (0.3)	0.0 (0.3)	2.8 (0.5)	1.6 (0.4)	1.1 (0.3)	0.1 (0.2)
Ratio2	-14.7 (1.9)	-3.3 (0.5)	-0.5 (0.5)	-2.4 (0.5)	-4.1 (0.4)	-0.7 (0.4)	-0.8 (0.4)	-1.6 (0.4)	-4.4 (0.6)	0.1 (0.5)	0.2 (0.3)	-0.3 (0.2)
Ratio3	-3.3 (1.9)	-5.5 (0.5)	-0.8 (0.5)	-0.3 (0.5)	-1.5 (0.4)	-1.9 (0.4)	-1.2 (0.4)	-0.2 (0.4)	-2.3 (0.6)	-1.6 (0.5)	-0.8 (0.3)	-0.1 (0.2)
N2	6.6 (1.9)	4.4 (0.5)	1.5 (0.5)	1.4 (0.5)	-9.5 (0.4)	-4.9 (0.4)	-5.2 (0.4)	-5.0 (0.4)	-5.4 (0.6)	-3.0 (0.5)	-2.2 (0.3)	-1.1 (0.2)
N3	-2.9 (1.9)	9.0 (0.5)	4.8 (0.5)	5.1 (0.5)	-12.8 (0.4)	-5.4 (0.4)	-6.2 (0.4)	-5.2 (0.4)	-10.6 (0.6)	-5.1 (0.5)	-3.7 (0.3)	-2.0 (0.2)
R^2	0.74	0.96	0.91	0.9	0.96	0.86	0.87	0.79	0.9	0.79	0.73	0.66

Note: OLS standard errors in brackets. Set1 = homogeneous treatment and homoscedasticity, Set2 = heterogeneous treatment and homoscedasticity, Set3 = homogeneous treatment and heteroscedasticity, Set4 = heterogeneous treatment and heteroscedasticity, Curve1 = Outcome equation depends linearly on the propensity score, Curve2 = Outcome equation depends on the propensity score in a nonlinear way, Ratio1 = Treated-to-Control Ratio 1:1, Ratio2 = Treated-to-Control Ratio 3:2, Ratio3 = Treated-to-Control Ratio 2:3, N1 = sample size 100, N2 = sample size 200, N3 = sample size 500.

3.5 Conclusion

Estimators that rely on propensity score weighting are among the most popular methods used in the literature on estimation of causal treatment effects. In this paper, we propose a simple and easy-to-implement method to improve those estimators in terms of MSE. The considerable gains in terms of MSE are demonstrated by a comprehensive Monte Carlo simulation study.

The methods we consider here require the first step estimation of the propensity

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

scores. We show that the MSE improvements of the first step estimation lead to MSE reduction of the treatment effect estimator in finite samples. To improve the first step propensity score estimation, we propose a simple shrinkage towards the unconditional mean of the treatment variable. Since the shrinkage parameter is a choice parameter, we suggest three different methods for choosing a λ which satisfies certain optimality conditions.

In the Monte Carlo study, we evaluate the finite sample properties with and without applying trimming rules. All three suggested choices of λ lead to an average improvement in the MSE of the average treatment effects for all four estimators.

In the first part of the Monte Carlo study we compare the estimators based on the shrunken propensity scores to the estimators based on the conventional propensity scores. We obtain a lower MSE in all of the settings if we use the fixed valued or the MSE minimizing tuning parameter. For the cross validated tuning parameter the MSE is reduced in 99.3% of the cases, respectively.

In the second part we base the estimators on the shrunken propensity scores combined with trimming rule proposed by Crump et al. (2009). If we compare this procedure to the estimators based on the conventional propensity score, conventional propensity score combined with trimming rule 1 as well as the conventional propensity score combined with trimming rule we find the following: With the fixed tuning parameter method our procedure leads in 99.7% of the cases to a lower MSE. With the MSE minimizing λ and the cross-validated λ our procedure outperforms in 98.8% and 96.9% cases, respectively.

Given this insight and the fact that the MSE minimizing choice of λ has a much higher computational cost than the fixed valued λ , the latter provides the best trade off between MSE gain and computational burden.

The main advantage of our approach is that it is very simple and can be implemented at basically no cost. Since the shrunken propensity scores are a simple linear combination of the conventional propensity scores and the share of treated, every improvement can be obtained without computational burdens.

Bibliography

- ABADIE, A. AND G. W. IMBENS (2006): “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74, 235–267.
- BUSSO, M., J. DINARDO, AND J. MCCRARY (2009): “Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects,” Unpublished manuscript, http://emlab.berkeley.edu/~jmccrary/BDM_JBES.pdf.
- CRUMP, R. K., V. J. HOTZ, G. W. IMBENS, AND O. A. MITNIK (2009): “Dealing with Limited Overlap in Estimation of Average Treatment Effects,” *Biometrika*, 96, 187–199.
- DEHEJIA, R. H. AND S. WAHBA (1999): “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053–1062.
- FRÖLICH, M. (2004): “Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators,” *Review of Economics and Statistics*, 86, 77–90.
- HAHN, J. (1998): “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66, 315–331.
- HECKMAN, J., H. ICHIMURA, AND P. TODD (1997): “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *The Review of Economic Studies*, 64, 605–654.
- HIRANO, K. AND G. IMBENS (2001): “Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization,” *Health Services and Outcomes Research Methodology*, 2, 259–278.
- HUBER, M., M. LECHNER, AND C. WUNSCH (2013): “The performance of estimators based on the propensity score,” *Journal of Econometrics*, 175, 1–21.
- IMBENS, G. (2004): “Nonparametric Estimation of Average Treatment Effects Under Exogeneity,” *Review of Economics and Statistics*, 86, 4–29.
- IMBENS, G. W. AND J. M. WOOLDRIDGE (2009): “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47, 5–86.
- KHAN, S. AND E. TAMER (2010): “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” *Econometrica*, 78, 2021–2042.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

LUNCEFORD, J. AND M. DAVIDIAN (2004): “Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study,” *Statistics in Medicine*, 23, 2937–2960.

ROBINS, J. M. AND A. ROTNITZKY (1995): “Semiparametric Efficiency in Multivariate Regression Models with Missing Data,” *Journal of the American Statistical Association*, 90, 122–129.

ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (1994): “Estimation of Regression Coefficients when Some Regressors are not Always Observed,” *Journal of the American Statistical Association*, 89, 846–866.

ROSENBAUM, P. R. AND D. B. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.

RUBIN, D. B. (1974): “Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies,” *Journal of Educational Psychology*, 66, 688–701.

SEIFERT, B. AND T. GASSER (1996): “Finite-Sample Variance of Local Polynomials: Analysis and Solutions,” *Journal of the American Statistical Association*, 91, 267–275.

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

——— (2007): “Inverse Probability Weighted Estimation for General Missing Data Problems,” *Journal of Econometrics*, 141, 1281–1301.

Appendix 3.A Tables

Table 3.A.1: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used instead of the conventional propensity scores, fixed valued λ

	N	linear			nonlinear			linear			nonlinear		
		100	200	500	100	200	500	100	200	500	100	200	500
Ratio 1:1		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	59.8	60.9	57.0	56.4	58.9	52.1	60.7	61.6	57.9	58.4	58.2	49.5
		(-0.2)	(-0.3)	(-0.1)	(-6.3)	(-7.0)	(-8.5)	(-0.5)	(-0.7)	(-0.5)	(-11.5)	(-13.1)	(-17.1)
	IPW 2	24.4	26.7	25.3	26.8	30.5	33.8	20.5	22.6	19.4	26.0	29.7	33.3
	(-3.9)	(-4.6)	(-6.8)	(-1.0)	(-1.0)	(-0.5)	(-8.5)	(-10.0)	(-13.7)	(-1.4)	(-1.4)	(-0.8)	
	IPW 3	10.3	9.0	6.4	11.4	10.2	8.6	8.2	6.9	3.8	11.1	10.0	8.5
		(-1.6)	(-1.6)	(-2.4)	(-0.4)	(-0.4)	(-0.1)	(-3.5)	(-3.7)	(-4.8)	(-0.6)	(-0.5)	(0.0)
	DR	8.5	9.9	11.0	8.5	10.0	11.3	8.5	9.9	11.1	8.5	10.0	11.0
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(-0.2)	(-0.3)	(-0.6)
Ratio 1:1		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	45.7	48.1	46.0	43.9	48.2	42.9	46.9	49.2	47.2	48.3	49.9	42.7
		(-0.2)	(-0.3)	(-0.1)	(-5.0)	(-5.6)	(-6.8)	(-0.4)	(-0.5)	(-0.4)	(-9.7)	(-11.1)	(-14.2)
	IPW 2	12.7	14.5	15.6	14.2	17.4	21.9	10.3	11.9	11.5	13.7	16.9	21.6
	(-2.5)	(-3.2)	(-4.9)	(-0.6)	(-0.7)	(-0.4)	(-5.7)	(-7.0)	(-10.0)	(-0.9)	(-0.9)	(-0.5)	
	IPW 3	4.8	4.4	2.8	5.5	5.2	4.3	3.5	3.1	1.1	5.3	5.1	4.3
		(-1.0)	(-1.1)	(-1.6)	(-0.3)	(-0.2)	(0.0)	(-2.2)	(-2.4)	(-3.3)	(-0.4)	(-0.3)	(0.0)
	DR	5.5	6.5	7.2	5.6	6.6	7.4	5.5	6.5	7.3	5.7	6.6	7.2
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.2)	(-0.4)
Ratio 3:2		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	55.9	47.9	38.1	52.5	43.2	35.2	56.1	48.3	38.7	52.8	43.1	35.2
		(-0.4)	(-0.5)	(-1.0)	(-1.1)	(-1.6)	(-1.4)	(-0.3)	(-0.3)	(-0.7)	(-3.4)	(-4.6)	(-4.8)
	IPW 2	20.6	22.6	22.0	22.2	23.8	23.0	18.8	21.2	20.6	21.5	22.9	22.3
	(-2.0)	(-2.0)	(-2.6)	(-1.1)	(-1.5)	(-1.4)	(-3.9)	(-4.0)	(-4.9)	(-1.6)	(-2.1)	(-1.9)	
	IPW 3	8.6	7.6	6.7	9.0	7.6	7.0	7.6	6.6	5.7	8.7	7.3	6.8
		(-0.9)	(-0.8)	(-1.0)	(-0.6)	(-0.7)	(-0.5)	(-1.8)	(-1.7)	(-1.9)	(-0.8)	(-1.0)	(-0.7)
	DR	5.9	6.6	8.3	5.9	6.7	8.4	5.9	6.7	8.2	6.0	6.9	8.6
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Ratio 3:2		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	40.6	33.9	27.4	37.1	29.3	24.7	41.0	34.3	27.9	38.4	30.1	25.4
		(-0.3)	(-0.4)	(-0.7)	(-0.8)	(-1.1)	(-1.0)	(-0.2)	(-0.2)	(-0.5)	(-2.6)	(-3.4)	(-3.5)
	IPW 2	10.7	12.9	13.8	11.7	13.7	14.2	9.6	12.2	12.9	11.2	13.2	13.8
	(-1.3)	(-1.3)	(-1.8)	(-0.7)	(-1.0)	(-1.0)	(-2.5)	(-2.6)	(-3.4)	(-1.1)	(-1.4)	(-1.3)	
	IPW 3	3.6	3.5	3.5	3.8	3.6	3.7	3.0	2.9	2.8	3.6	3.4	3.6
		(-0.6)	(-0.5)	(-0.6)	(-0.4)	(-0.4)	(-0.3)	(-1.2)	(-1.0)	(-1.3)	(-0.5)	(-0.6)	(-0.4)
	DR	3.9	4.5	5.8	3.9	4.6	5.9	3.9	4.5	5.8	4.0	4.7	6.0
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Ratio 2:3		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	47.7	65.5	36.3	46.8	66.1	33.0	48.7	66.5	36.8	48.1	67.4	31.6
		(-2.5)	(-2.0)	(-4.0)	(-9.2)	(-6.8)	(-14.6)	(-3.2)	(-2.6)	(-5.4)	(-15.4)	(-10.9)	(-24.6)
	IPW 2	19.7	21.9	18.8	21.6	23.2	21.7	17.0	19.5	15.8	21.5	23.2	21.8
	(-2.0)	(-2.3)	(-3.2)	(-0.1)	(-0.1)	(0.0)	(-5.2)	(-5.9)	(-7.6)	(-0.2)	(-0.1)	(0.0)	
	IPW 3	8.9	7.5	5.7	9.7	8.4	7.0	7.4	6.0	4.0	9.6	8.3	7.0
		(-0.9)	(-0.9)	(-1.2)	(-0.1)	(0.0)	(0.0)	(-2.3)	(-2.4)	(-2.8)	(-0.1)	(0.0)	(0.1)
	DR	6.2	6.3	7.3	6.2	6.4	7.4	6.2	6.3	7.3	6.1	6.3	7.1
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.1)	(-0.3)	(-0.4)
Ratio 2:3		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	36.5	54.8	28.4	37.1	57.2	27.0	37.6	56.3	29.2	40.2	60.6	27.3
		(-1.8)	(-1.7)	(-2.9)	(-7.1)	(-5.8)	(-10.9)	(-2.4)	(-2.2)	(-3.9)	(-12.5)	(-9.7)	(-19.6)
	IPW 2	9.9	11.8	11.2	11.0	12.4	13.2	8.3	10.4	9.3	11.0	12.4	13.3
	(-1.3)	(-1.5)	(-2.1)	(-0.1)	(-0.1)	(0.0)	(-3.3)	(-3.9)	(-5.2)	(-0.1)	(-0.1)	(0.0)	
	IPW 3	4.4	3.7	2.8	4.9	4.2	3.7	3.5	2.8	1.8	4.8	4.2	3.6
		(-0.5)	(-0.6)	(-0.8)	(0.0)	(0.0)	(0.0)	(-1.4)	(-1.5)	(-1.8)	(0.0)	(0.0)	(0.0)
	DR	4.0	4.1	5.0	4.0	4.2	5.0	4.0	4.1	5.0	4.0	4.1	4.8
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.2)	(-0.3)	

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{p}^*))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{p}^*))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.2: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used instead of the conventional propensity scores, $MSE(\hat{p}_i^s)$ -minimizing λ

N	linear			nonlinear			linear			nonlinear			
	100	200	500	100	200	500	100	200	500	100	200	500	
Ratio 1:1	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	60.9	66.3	66.3	57.5	63.3	58.4	61.6	67.3	66.3	58.1	64.6	56.9
		(-0.2)	(-0.3)	(-0.2)	(-6.2)	(-6.2)	(-7.3)	(-0.4)	(-0.6)	(-0.5)	(-11.8)	(-11.0)	(-13.9)
	IPW 2	24.6	26.5	31.3	27.7	29.0	34.3	20.5	22.5	28.0	27.0	28.3	33.8
	(-4.3)	(-3.9)	(-5.3)	(-0.9)	(-1.3)	(-0.9)	(-9.2)	(-8.9)	(-10.7)	(-1.3)	(-1.7)	(-1.3)	
	IPW 3	9.9	9.3	7.3	11.4	10.0	8.7	7.6	7.3	5.3	11.2	9.8	8.6
		(-1.8)	(-1.4)	(-1.7)	(-0.4)	(-0.5)	(-0.3)	(-4.0)	(-3.2)	(-3.6)	(-0.5)	(-0.6)	(-0.3)
	DR	8.5	9.6	11.2	8.6	9.8	11.7	8.4	9.6	11.1	8.6	9.9	11.7
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.2)	(-0.2)	(-0.4)
Ratio 1:1	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	46.4	54.1	56.5	44.8	52.6	49.7	47.5	55.5	56.8	47.9	56.1	50.2
		(-0.2)	(-0.3)	(-0.1)	(-4.8)	(-5.1)	(-6.1)	(-0.3)	(-0.5)	(-0.4)	(-9.7)	(-9.5)	(-12.2)
	IPW 2	11.9	15.8	21.4	13.7	17.2	23.2	9.5	13.4	19.7	13.3	16.8	22.8
	(-2.7)	(-2.6)	(-3.9)	(-0.6)	(-0.9)	(-0.7)	(-5.9)	(-5.9)	(-7.9)	(-0.9)	(-1.2)	(-0.9)	
	IPW 3	4.1	4.5	3.9	5.0	4.9	5.0	2.7	3.3	2.5	4.9	4.8	4.9
		(-1.1)	(-0.8)	(-1.2)	(-0.3)	(-0.3)	(-0.2)	(-2.4)	(-2.0)	(-2.5)	(-0.3)	(-0.4)	(-0.2)
	DR	4.8	6.6	7.7	4.9	6.7	8.3	4.9	6.6	7.7	5.0	6.8	8.4
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.1)	(-0.3)
Ratio 3:2	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	51.9	86.8	42.0	47.7	71.8	41.0	52.4	86.7	42.9	48.8	70.6	42.8
		(-0.5)	(-0.2)	(-0.7)	(-1.4)	(-0.7)	(-1.8)	(-0.4)	(-0.1)	(-0.5)	(-4.3)	(-2.4)	(-5.0)
	IPW 2	21.0	21.3	22.2	22.0	22.6	24.3	19.1	19.3	20.6	21.1	21.8	23.5
	(-2.4)	(-2.5)	(-2.0)	(-1.2)	(-1.4)	(-2.0)	(-4.7)	(-4.8)	(-4.1)	(-1.8)	(-2.0)	(-2.6)	
	IPW 3	9.5	8.1	6.9	9.9	8.6	6.7	8.3	7.0	6.0	9.6	8.3	6.4
		(-1.2)	(-1.0)	(-0.6)	(-0.6)	(-0.6)	(-0.8)	(-2.3)	(-2.1)	(-1.4)	(-0.9)	(-0.9)	(-1.0)
	DR	6.1	6.7	8.3	6.2	6.7	8.5	6.1	6.6	8.3	6.3	6.9	8.8
		(0.0)	(0.0)	(-0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Ratio 3:2	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	36.1	79.4	30.5	32.0	59.6	29.5	36.6	79.3	31.3	34.2	59.1	31.9
		(-0.4)	(-0.2)	(-0.5)	(-1.0)	(-0.6)	(-1.3)	(-0.3)	(-0.1)	(-0.3)	(-3.1)	(-2.1)	(-3.8)
	IPW 2	10.1	11.8	13.9	10.7	12.9	15.7	9.0	10.5	12.8	10.1	12.4	15.1
	(-1.5)	(-1.7)	(-1.4)	(-0.8)	(-0.9)	(-1.4)	(-3.0)	(-3.3)	(-2.9)	(-1.2)	(-1.3)	(-1.8)	
	IPW 3	3.8	3.9	3.7	4.1	4.2	3.6	3.1	3.1	3.1	3.8	4.0	3.4
		(-0.7)	(-0.7)	(-0.4)	(-0.4)	(-0.4)	(-0.5)	(-1.4)	(-1.4)	(-1.0)	(-0.6)	(-0.6)	(-0.7)
	DR	3.8	4.5	5.9	3.9	4.6	6.0	3.8	4.5	5.9	4.0	4.7	6.2
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Ratio 2:3	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	72.5	56.0	38.6	74.2	56.0	33.9	73.5	57.1	38.8	75.6	56.9	34.5
		(-1.3)	(-2.6)	(-4.7)	(-4.9)	(-9.2)	(-16.0)	(-1.7)	(-3.4)	(-6.2)	(-7.9)	(-15.0)	(-25.2)
	IPW 2	21.3	21.7	24.5	23.6	24.0	24.8	18.3	18.7	22.9	23.5	23.9	25.0
	(-2.7)	(-2.6)	(-2.0)	(-0.1)	(-0.1)	(-0.1)	(-6.4)	(-6.5)	(-5.7)	(-0.1)	(-0.1)	(-0.1)	
	IPW 3	9.2	8.0	6.5	10.5	9.1	7.1	7.3	6.3	5.1	10.5	9.0	7.1
		(-1.3)	(-1.1)	(-0.7)	(0.0)	(0.0)	(-0.1)	(-3.1)	(-2.7)	(-2.0)	(0.0)	(0.0)	(-0.1)
	DR	6.4	7.1	7.6	6.4	7.1	7.8	6.3	7.1	7.5	6.3	7.0	7.8
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.1)	(-0.2)	(-0.2)	(-0.2)
Ratio 2:3	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	61.1	44.9	29.9	64.9	46.6	27.1	62.7	46.3	30.3	68.7	49.7	29.1
		(-1.1)	(-2.0)	(-3.6)	(-4.2)	(-7.3)	(-12.6)	(-1.4)	(-2.6)	(-4.8)	(-7.1)	(-12.7)	(-21.2)
	IPW 2	9.7	11.8	15.6	11.1	13.2	15.5	8.1	10.1	15.1	11.1	13.2	15.7
	(-1.6)	(-1.6)	(-1.4)	(-0.1)	(-0.1)	(-0.1)	(-4.0)	(-4.1)	(-4.0)	(-0.1)	(-0.1)	(-0.1)	
	IPW 3	3.8	3.7	3.2	4.6	4.3	3.6	2.7	2.6	2.4	4.6	4.3	3.6
		(-0.7)	(-0.7)	(-0.5)	(0.0)	(0.0)	(-0.1)	(-1.8)	(-1.7)	(-1.3)	(0.0)	(0.0)	(-0.1)
	DR	3.5	4.6	4.9	3.6	4.6	5.1	3.5	4.6	4.8	3.6	4.6	5.2
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.2)	(-0.1)

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{p}^s))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.3: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used instead of the conventional propensity scores, data-driven λ

	N	linear			nonlinear			linear			nonlinear		
		100	200	500	100	200	500	100	200	500	100	200	500
Ratio 1:1		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	9.1 (0.0)	6.4 (-0.1)	1.9 (0.0)	7.7 (-1.1)	4.9 (-1.2)	1.7 (-0.5)	9.1 (-0.1)	6.4 (-0.2)	2.0 (-0.1)	6.0 (-2.2)	3.3 (-2.1)	1.3 (-0.9)
	IPW 2	7.0 (-1.1)	4.8 (-0.4)	3.5 (-0.4)	8.1 (-0.2)	5.8 (-0.3)	4.5 (-0.1)	5.6 (-2.3)	3.5 (-1.1)	2.8 (-0.9)	8.1 (-0.3)	5.6 (-0.4)	4.5 (-0.1)
	IPW 3	4.0 (-0.6)	2.6 (-0.2)	1.7 (-0.3)	4.4 (-0.1)	2.9 (-0.2)	2.2 (-0.1)	3.3 (-1.3)	2.0 (-0.6)	1.2 (-0.6)	4.6 (-0.1)	2.9 (-0.2)	2.2 (0.0)
DR	3.4 (0.0)	2.7 (0.0)	2.5 (0.0)	3.4 (0.0)	2.8 (0.0)	2.7 (0.0)	3.6 (0.0)	2.7 (0.0)	2.4 (0.0)	3.6 (-0.1)	2.7 (0.0)	2.7 (-0.1)	
Ratio 1:1		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	7.4 (0.0)	5.2 (-0.1)	1.8 (0.0)	6.4 (-0.9)	4.0 (-1.0)	1.5 (-0.4)	7.5 (-0.1)	5.2 (-0.1)	1.8 (-0.1)	5.3 (-1.9)	2.9 (-1.8)	1.2 (-0.8)
	IPW 2	3.1 (-0.7)	2.3 (-0.3)	2.2 (-0.3)	3.9 (-0.1)	2.9 (-0.2)	2.8 (-0.1)	2.2 (-1.5)	1.5 (-0.7)	1.8 (-0.7)	4.0 (-0.2)	2.8 (-0.3)	2.7 (-0.1)
	IPW 3	1.6 (-0.4)	1.2 (-0.1)	1.0 (-0.2)	1.9 (-0.1)	1.3 (-0.1)	1.3 (0.0)	1.2 (-0.8)	0.8 (-0.4)	0.7 (-0.4)	2.1 (-0.1)	1.3 (-0.1)	1.2 (0.0)
DR	2.1 (0.0)	1.6 (0.0)	1.7 (0.0)	2.1 (0.0)	1.7 (0.0)	1.8 (0.0)	2.2 (0.0)	1.6 (0.0)	1.7 (0.0)	2.2 (0.0)	1.7 (0.0)	1.8 (0.0)	
Ratio 3:2		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	13.2 (-0.1)	8.3 (0.0)	5.4 (-0.1)	13.1 (-0.5)	8.3 (-0.3)	5.5 (-0.2)	13.4 (-0.1)	8.5 (0.0)	5.6 (0.0)	12.8 (-1.5)	8.1 (-0.8)	5.4 (-0.5)
	IPW 2	7.6 (-0.9)	4.9 (-0.4)	3.0 (-0.3)	8.0 (-0.4)	5.1 (-0.3)	3.1 (-0.1)	6.7 (-1.7)	4.3 (-0.8)	2.5 (-0.6)	7.7 (-0.6)	4.9 (-0.4)	2.9 (-0.2)
	IPW 3	4.5 (-0.5)	2.8 (-0.3)	1.7 (-0.2)	4.7 (-0.3)	2.8 (-0.2)	1.7 (-0.1)	4.0 (-1.0)	2.5 (-0.5)	1.5 (-0.4)	4.6 (-0.4)	2.7 (-0.3)	1.7 (-0.1)
DR	3.0 (0.0)	2.3 (0.0)	1.7 (0.0)	3.0 (0.0)	2.4 (0.0)	1.8 (0.0)	3.0 (0.0)	2.3 (0.0)	1.7 (0.0)	3.0 (0.0)	2.4 (0.0)	1.9 (0.0)	
Ratio 3:2		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	8.7 (-0.1)	5.7 (0.0)	3.8 (0.0)	8.4 (-0.4)	5.6 (-0.2)	3.5 (-0.1)	8.9 (0.0)	5.9 (0.0)	3.9 (0.0)	8.6 (-1.1)	5.7 (-0.6)	3.5 (-0.4)
	IPW 2	3.2 (-0.5)	2.4 (-0.3)	1.8 (-0.2)	3.4 (-0.2)	2.5 (-0.2)	1.6 (-0.1)	2.7 (-1.1)	2.0 (-0.5)	1.5 (-0.4)	3.2 (-0.4)	2.3 (-0.3)	1.4 (-0.1)
	IPW 3	1.6 (-0.3)	1.2 (-0.2)	1.0 (-0.1)	1.7 (-0.2)	1.2 (-0.1)	0.8 (-0.1)	1.3 (-0.6)	1.0 (-0.3)	0.9 (-0.3)	1.6 (-0.2)	1.1 (-0.2)	0.7 (-0.1)
DR	1.8 (0.0)	1.5 (0.0)	1.1 (0.0)	1.8 (0.0)	1.5 (0.0)	1.1 (0.0)	1.8 (0.0)	1.5 (0.0)	1.1 (0.0)	1.8 (0.0)	1.6 (0.0)	1.2 (0.0)	
Ratio 2:3		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	9.7 (-0.3)	6.4 (-0.5)	3.8 (-0.3)	6.2 (-1.7)	2.9 (-1.8)	0.1 (-1.1)	9.2 (-0.4)	6.0 (-0.6)	3.4 (-0.4)	3.6 (-2.8)	0.2 (-2.9)	-2.9 (-1.8)
	IPW 2	7.3 (-0.8)	5.2 (-0.2)	3.2 (-0.4)	8.3 (0.0)	5.8 (-0.1)	4.0 (0.0)	5.7 (-2.0)	4.1 (-0.8)	1.9 (-0.8)	8.3 (0.0)	5.8 (-0.1)	4.0 (0.0)
	IPW 3	4.3 (-0.5)	3.0 (-0.1)	1.9 (-0.2)	4.9 (0.0)	3.2 (0.0)	2.1 (0.0)	3.4 (-1.2)	2.5 (-0.5)	1.3 (-0.5)	4.9 (0.0)	3.2 (0.0)	2.2 (0.0)
DR	2.9 (0.0)	2.3 (0.0)	2.0 (0.0)	2.9 (0.0)	2.4 (0.0)	2.0 (0.0)	2.9 (0.0)	2.3 (0.0)	1.9 (0.0)	2.8 (-0.1)	2.3 (0.0)	1.9 (-0.1)	
Ratio 2:3		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	7.7 (-0.2)	5.0 (-0.4)	2.9 (-0.2)	5.3 (-1.3)	2.6 (-1.4)	0.0 (-0.8)	7.4 (-0.3)	4.8 (-0.5)	2.5 (-0.3)	3.3 (-2.3)	0.5 (-2.4)	-2.5 (-1.5)
	IPW 2	2.9 (-0.5)	2.5 (-0.1)	1.8 (-0.2)	3.6 (0.0)	2.9 (-0.1)	2.3 (0.0)	2.0 (-1.3)	1.8 (-0.5)	1.0 (-0.5)	3.6 (0.0)	2.9 (-0.1)	2.3 (0.0)
	IPW 3	1.7 (-0.3)	1.4 (-0.1)	1.0 (-0.1)	2.0 (0.0)	1.6 (0.0)	1.2 (0.0)	1.2 (-0.7)	1.1 (-0.3)	0.7 (-0.3)	2.0 (0.0)	1.6 (0.0)	1.2 (0.0)
DR	1.7 (0.0)	1.4 (0.0)	1.3 (0.0)	1.7 (0.0)	1.4 (0.0)	1.3 (0.0)	1.7 (0.0)	1.4 (0.0)	1.3 (0.0)	1.6 (-0.1)	1.4 (0.0)	1.3 (-0.1)	

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^s))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.4: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores, fixed valued λ

N	linear			nonlinear			linear			nonlinear			
	100	200	500	100	200	500	100	200	500	100	200	500	
Ratio 1:1	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	53.9	59.0	55.8	55.4	62.2	56.7	54.9	59.8	56.1	64.2	68.3	63.1
		(0.0)	(0.0)	(0.0)	(-1.3)	(-1.0)	(-0.7)	(-0.1)	(0.0)	(0.0)	(-1.2)	(-0.5)	(0.1)
	IPW 2	22.1	29.2	33.6	22.7	31.0	37.3	22.0	29.3	33.6	22.6	30.2	34.0
	(-0.7)	(-0.9)	(-0.7)	(-0.1)	(0.0)	(0.1)	(-1.6)	(-1.9)	(-1.5)	(0.1)	(-0.5)	(-3.1)	
	IPW 3	11.2	15.7	17.8	11.8	16.7	18.8	10.5	14.8	16.6	11.8	15.6	14.2
		(-0.2)	(-0.4)	(-0.1)	(0.0)	(0.1)	(0.0)	(-0.5)	(-0.8)	(-0.3)	(0.0)	(-0.9)	(-4.9)
	DR	12.9	18.2	20.7	13.4	18.7	22.6	12.9	18.5	19.9	13.7	17.8	18.6
		(0.1)	(0.1)	(-0.1)	(0.0)	(0.0)	(0.2)	(0.2)	(0.3)	(-0.1)	(-0.6)	(-1.9)	(-5.6)
Ratio 1:1	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	38.2	43.8	42.2	40.5	48.5	43.9	39.6	45.0	43.0	51.2	56.7	51.9
		(0.0)	(0.0)	(0.0)	(-1.0)	(-0.8)	(-0.5)	(0.0)	(0.0)	(0.0)	(-1.0)	(-0.4)	(0.1)
	IPW 2	8.9	13.5	18.5	9.6	15.1	21.1	9.2	13.6	18.6	9.6	14.8	18.7
	(-0.5)	(-0.7)	(-0.6)	(-0.1)	(0.0)	(0.1)	(-1.1)	(-1.4)	(-1.2)	(0.1)	(-0.3)	(-2.4)	
	IPW 3	2.1	4.6	5.3	2.7	5.3	5.7	1.8	3.8	4.5	2.7	4.7	2.5
		(-0.1)	(-0.3)	(-0.1)	(0.0)	(0.1)	(0.0)	(-0.4)	(-0.6)	(-0.3)	(0.0)	(-0.6)	(-3.4)
	DR	4.8	7.3	7.6	5.1	7.5	8.9	5.0	7.3	7.0	5.4	7.0	6.1
		(0.1)	(0.1)	(-0.1)	(0.0)	(0.0)	(0.1)	(0.1)	(0.2)	(-0.1)	(-0.4)	(-1.3)	(-4.1)
Ratio 3:2	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	51.8	45.1	42.9	46.9	38.5	37.2	50.4	42.8	39.0	50.5	43.4	42.6
		(-0.1)	(-0.2)	(-0.1)	(-2.0)	(-2.3)	(-2.1)	(-1.1)	(-2.0)	(-3.4)	(-1.8)	(-1.0)	(0.0)
	IPW 2	19.7	22.8	29.7	21.4	24.7	30.4	20.0	23.6	30.2	20.9	23.9	27.0
	(-0.6)	(-0.5)	(-0.6)	(-0.1)	(-0.1)	(-0.2)	(-0.2)	(0.0)	(0.0)	(0.1)	(-0.4)	(-3.3)	
	IPW 3	10.9	10.9	19.0	11.7	11.9	19.3	11.1	10.8	17.8	11.1	10.7	14.8
		(-0.3)	(-0.2)	(-0.2)	(0.1)	(0.1)	(0.0)	(0.2)	(0.1)	(-0.4)	(-0.1)	(-0.7)	(-4.4)
	DR	12.5	12.6	21.2	12.6	12.9	21.4	11.8	11.1	18.0	11.4	10.7	15.4
		(0.1)	(0.1)	(0.0)	(0.0)	(0.1)	(0.0)	(-0.5)	(-1.1)	(-2.6)	(-1.2)	(-2.3)	(-6.5)
Ratio 3:2	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	36.7	31.1	30.9	31.9	25.0	25.4	35.8	29.7	28.4	35.7	29.4	29.8
		(-0.1)	(-0.1)	(-0.1)	(-1.5)	(-1.6)	(-1.5)	(-0.8)	(-1.4)	(-2.5)	(-1.4)	(-0.7)	(0.0)
	IPW 2	7.1	10.9	17.9	7.9	12.0	17.0	7.5	11.8	18.7	7.6	11.4	14.4
	(-0.4)	(-0.4)	(-0.4)	(0.0)	(0.0)	(-0.1)	(-0.1)	(0.0)	(0.0)	(0.0)	(-0.3)	(-2.3)	
	IPW 3	1.0	2.7	9.4	1.4	3.4	8.8	1.2	2.8	9.1	1.0	2.5	5.5
		(-0.2)	(-0.2)	(-0.1)	(0.0)	(0.1)	(0.0)	(0.1)	(0.1)	(-0.2)	(0.0)	(-0.5)	(-3.0)
	DR	2.1	4.0	10.3	2.2	4.2	10.5	1.7	3.1	8.5	1.4	2.7	6.2
		(0.0)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.3)	(-0.7)	(-1.7)	(-0.8)	(-1.6)	(-4.5)
Ratio 2:3	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	45.2	66.2	41.9	50.3	70.7	46.5	46.6	67.4	40.9	58.0	75.7	53.7
		(0.2)	(0.1)	(0.0)	(-0.3)	(-0.2)	(-0.4)	(0.1)	(-0.2)	(-1.6)	(-0.5)	(-0.3)	(-0.1)
	IPW 2	18.2	23.2	26.2	18.7	23.3	27.2	16.3	20.6	20.5	18.8	23.2	27.2
	(-0.5)	(-0.7)	(-0.5)	(-0.1)	(0.0)	(0.0)	(-3.1)	(-4.5)	(-6.9)	(0.0)	(0.0)	(-0.2)	
	IPW 3	10.6	12.3	16.6	10.9	13.0	17.6	8.6	9.0	9.8	11.0	12.8	17.3
		(-0.2)	(-0.3)	(-0.1)	(-0.1)	(0.0)	(0.0)	(-2.2)	(-3.6)	(-5.9)	(0.0)	(-0.1)	(-0.3)
	DR	12.5	14.3	18.9	12.6	14.5	19.7	12.7	13.9	15.8	13.1	14.8	19.8
		(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(-0.7)	(-2.6)	(0.0)	(-0.2)	(-0.4)
Ratio 2:3	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	30.1	52.1	29.8	35.9	58.4	35.3	31.8	53.8	29.8	45.3	65.9	44.0
		(0.1)	(0.1)	(0.0)	(-0.2)	(-0.2)	(-0.3)	(0.1)	(-0.1)	(-1.1)	(-0.4)	(-0.2)	(-0.1)
	IPW 2	7.7	10.1	14.3	8.1	10.1	15.1	6.8	8.4	10.4	8.3	10.2	15.1
	(-0.3)	(-0.4)	(-0.3)	(-0.1)	(0.0)	(0.0)	(-2.0)	(-3.0)	(-4.5)	(0.0)	(0.0)	(-0.1)	
	IPW 3	4.0	3.4	7.4	4.2	4.0	8.0	2.8	1.2	3.0	4.3	4.0	7.9
		(-0.1)	(-0.2)	(0.0)	(0.0)	(0.0)	(0.0)	(-1.4)	(-2.3)	(-3.6)	(0.0)	(0.0)	(-0.2)
	DR	8.1	6.1	10.0	8.0	6.3	10.2	8.3	5.6	8.2	8.4	6.7	10.4
		(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(-0.4)	(-1.6)	(0.0)	(-0.1)	(-0.2)

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^*))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.5: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores with trimming rule 1, fixed valued λ

	N	linear			nonlinear			linear			nonlinear		
		100	200	500	100	200	500	100	200	500	100	200	500
Ratio 1:1		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	26.7	19.3	17.0	29.4	19.4	13.7	34.0	24.3	19.4	34.2	24.7	17.6
		(0.0)	(-0.1)	(0.1)	(-2.2)	(-2.4)	(-1.4)	(-0.1)	(-0.1)	(0.1)	(-3.0)	(-0.8)	(2.1)
	IPW 2	15.0	11.4	10.5	15.4	12.3	11.6	15.1	10.8	9.6	16.7	14.2	11.3
	(-1.0)	(-1.3)	(-1.7)	(-0.2)	(-0.1)	(0.2)	(-2.3)	(-2.8)	(-3.1)	(1.1)	(1.4)	(-0.3)	
	IPW 3	13.9	10.3	8.8	14.1	10.9	9.6	14.5	10.2	8.6	15.2	12.4	8.6
		(-0.5)	(-0.6)	(-0.8)	(-0.1)	(0.0)	(0.2)	(-1.1)	(-1.4)	(-1.5)	(1.1)	(1.2)	(-1.0)
	DR	10.2	7.0	7.2	9.9	7.1	7.7	11.1	7.5	7.7	10.3	7.8	5.3
		(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.1)	(0.9)	(0.3)	(-2.7)
Ratio 1:1		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	21.8	13.5	10.6	23.8	14.1	8.6	27.0	16.7	12.0	27.8	18.6	12.0
		(0.0)	(0.0)	(0.1)	(-1.4)	(-1.5)	(-0.8)	(-0.1)	(-0.1)	(0.1)	(-2.0)	(-0.5)	(1.5)
	IPW 2	11.9	6.3	5.4	12.1	7.2	6.1	12.1	5.4	4.4	12.9	8.7	6.3
	(-0.6)	(-0.8)	(-1.1)	(-0.1)	(-0.1)	(0.1)	(-1.4)	(-1.7)	(-2.0)	(0.7)	(0.9)	(-0.2)	
	IPW 3	12.1	6.4	4.7	12.2	6.9	5.2	12.6	5.8	4.1	12.8	8.2	4.9
		(-0.3)	(-0.4)	(-0.5)	(-0.1)	(0.0)	(0.1)	(-0.7)	(-0.8)	(-1.0)	(0.7)	(0.8)	(-0.7)
	DR	11.6	6.0	4.8	11.4	6.1	5.1	12.3	5.9	4.5	11.6	6.7	4.0
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.0)	(0.6)	(0.2)	(-1.7)
Ratio 3:2		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	23.7	15.9	15.0	24.2	14.9	12.3	28.1	17.5	11.9	28.6	20.1	18.1
		(-0.2)	(-0.3)	(-0.2)	(-2.8)	(-3.2)	(-2.9)	(-1.5)	(-2.6)	(-4.8)	(-1.9)	(-0.7)	(1.3)
	IPW 2	15.0	10.0	10.6	14.8	10.7	12.2	16.7	11.3	10.9	16.2	12.4	11.2
	(-0.7)	(-0.6)	(-0.8)	(-0.2)	(-0.3)	(-0.2)	(-0.3)	(0.1)	(0.0)	(1.2)	(0.6)	(-2.3)	
	IPW 3	14.3	8.9	9.5	14.1	9.3	10.4	15.8	9.8	8.8	15.1	10.5	8.4
		(-0.4)	(-0.3)	(-0.3)	(-0.1)	(-0.1)	(-0.1)	(0.0)	(0.2)	(-0.5)	(1.0)	(0.3)	(-3.1)
	DR	10.3	6.0	8.0	10.1	6.0	8.2	10.5	5.0	4.7	10.1	5.5	3.9
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.4)	(-1.0)	(-2.8)	(0.4)	(-0.9)	(-5.2)
Ratio 3:2		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	19.0	10.1	8.0	19.3	9.5	5.9	22.1	11.2	6.2	22.5	12.9	10.0
		(-0.1)	(-0.2)	(-0.1)	(-1.8)	(-1.9)	(-1.8)	(-1.0)	(-1.6)	(-3.1)	(-1.2)	(-0.3)	(0.9)
	IPW 2	11.0	5.0	3.7	10.7	5.5	4.2	12.0	5.8	4.0	11.7	6.7	3.8
	(-0.4)	(-0.4)	(-0.5)	(-0.1)	(-0.1)	(-0.1)	(-0.2)	(0.0)	(0.0)	(0.7)	(0.4)	(-1.5)	
	IPW 3	11.2	4.8	3.1	11.0	5.1	3.3	12.1	5.4	2.8	11.7	5.9	2.3
		(-0.2)	(-0.2)	(-0.2)	(-0.1)	(-0.1)	(0.0)	(0.0)	(0.1)	(-0.3)	(0.7)	(0.2)	(-2.0)
	DR	10.2	4.2	2.4	10.1	4.3	2.5	10.3	3.5	0.3	10.3	4.1	0.0
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.2)	(-0.5)	(-1.7)	(0.3)	(-0.6)	(-3.4)
Ratio 2:3		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	23.9	16.3	16.7	28.1	18.5	15.8	30.3	20.7	16.9	33.1	22.2	17.5
		(0.2)	(0.0)	(0.0)	(-0.5)	(-0.9)	(-0.7)	(-0.1)	(-0.6)	(-2.1)	(-1.8)	(-1.4)	(-0.2)
	IPW 2	13.4	9.1	11.1	13.1	9.6	11.2	11.7	5.1	3.4	13.5	10.5	12.1
	(-0.6)	(-0.8)	(-0.7)	(-0.1)	(0.0)	(0.0)	(-3.2)	(-5.1)	(-8.2)	(0.4)	(0.8)	(1.0)	
	IPW 3	12.7	8.5	10.3	12.5	8.9	10.3	11.8	5.6	4.3	12.9	9.7	11.0
		(-0.3)	(-0.4)	(-0.3)	(-0.1)	(0.0)	(0.0)	(-2.2)	(-3.7)	(-6.2)	(0.5)	(0.7)	(0.9)
	DR	9.4	6.0	8.8	9.1	5.9	8.6	10.1	5.7	6.5	9.4	6.4	9.0
		(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.1)	(-0.2)	(-0.8)	(-2.5)	(0.6)	(0.6)	(0.7)
Ratio 2:3		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	19.1	10.1	11.2	22.3	12.1	11.1	23.6	13.1	11.6	26.6	15.5	13.1
		(0.1)	(0.0)	(0.0)	(-0.3)	(-0.6)	(-0.5)	(-0.1)	(-0.4)	(-1.3)	(-1.2)	(-1.0)	(-0.2)
	IPW 2	11.2	4.6	6.3	11.0	5.3	6.3	10.3	2.2	1.6	11.3	5.9	6.9
	(-0.4)	(-0.5)	(-0.4)	(-0.1)	(0.0)	(0.0)	(-1.8)	(-3.0)	(-4.8)	(0.3)	(0.5)	(0.6)	
	IPW 3	11.5	5.0	6.1	11.4	5.5	6.0	11.1	3.1	2.4	11.6	6.1	6.5
		(-0.2)	(-0.3)	(-0.2)	(0.0)	(0.0)	(0.0)	(-1.2)	(-2.2)	(-3.6)	(0.3)	(0.4)	(0.6)
	DR	11.8	5.1	5.9	11.6	5.1	5.7	12.3	4.7	4.5	11.7	5.5	6.1
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(-0.1)	(-0.5)	(-1.4)	(0.4)	(0.4)	(0.5)

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^*))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.6: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores with trimming rule 2, fixed valued λ

	N	linear			nonlinear			linear			nonlinear		
		100	200	500	100	200	500	100	200	500	100	200	500
Ratio 1:1		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	24.0	18.0	10.2	24.7	18.5	9.9	25.4	19.0	11.0	30.1	26.1	19.6
		(-0.1)	(0.0)	(0.1)	(-2.8)	(-2.4)	(-1.3)	(-0.2)	(-0.1)	(0.1)	(-1.9)	(1.7)	(6.9)
	IPW 2	12.0	6.6	2.7	12.5	7.4	4.1	11.2	5.0	1.4	13.7	10.0	8.9
	(-1.1)	(-1.5)	(-2.0)	(-0.3)	(-0.1)	(0.2)	(-2.5)	(-3.2)	(-3.7)	(1.4)	(2.8)	(5.4)	
	IPW 3	7.5	4.0	1.2	7.8	4.3	1.9	7.0	3.1	0.7	8.8	6.4	5.9
		(-0.5)	(-0.7)	(-1.0)	(-0.2)	(0.0)	(0.2)	(-1.3)	(-1.6)	(-1.8)	(1.2)	(2.3)	(4.4)
	DR	3.0	1.4	0.1	2.8	1.1	-0.1	3.2	1.4	0.3	3.1	2.0	2.1
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.7)	(1.2)	(2.3)
Ratio 1:1		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	19.2	13.6	6.9	19.9	14.6	7.0	20.2	14.4	7.3	24.2	20.5	13.9
		(-0.1)	(0.0)	(0.1)	(-1.7)	(-1.4)	(-0.7)	(-0.1)	(0.0)	(0.1)	(-1.2)	(1.2)	(4.9)
	IPW 2	8.5	4.0	1.9	8.8	5.2	3.2	8.0	3.0	1.1	9.6	6.8	6.1
	(-0.7)	(-0.9)	(-1.2)	(-0.2)	(-0.1)	(0.2)	(-1.5)	(-1.9)	(-2.3)	(0.9)	(1.7)	(3.3)	
	IPW 3	7.2	3.6	1.7	7.4	4.2	2.4	6.9	3.0	1.3	8.0	5.5	4.8
		(-0.3)	(-0.4)	(-0.6)	(-0.1)	(0.0)	(0.1)	(-0.8)	(-0.9)	(-1.1)	(0.7)	(1.4)	(2.7)
	DR	6.1	3.7	2.1	5.9	3.6	2.2	6.2	3.8	2.1	5.9	4.1	3.5
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.4)	(0.7)	(1.4)
Ratio 3:2		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	19.3	13.9	8.0	18.0	12.4	6.3	18.7	13.0	6.5	23.6	20.1	18.2
		(-0.2)	(-0.3)	(-0.2)	(-3.1)	(-3.4)	(-3.2)	(-1.0)	(-1.5)	(-1.6)	(-1.1)	(1.7)	(7.2)
	IPW 2	10.2	7.2	4.0	10.0	7.6	4.6	11.8	9.8	8.5	11.6	10.5	8.9
	(-0.7)	(-0.7)	(-0.8)	(-0.3)	(-0.3)	(-0.2)	(0.4)	(1.6)	(3.5)	(1.7)	(2.9)	(4.5)	
	IPW 3	7.2	4.8	2.6	7.0	4.9	2.9	8.3	6.7	5.8	8.3	7.3	6.3
		(-0.4)	(-0.3)	(-0.4)	(-0.2)	(-0.2)	(-0.1)	(0.5)	(1.4)	(2.9)	(1.5)	(2.4)	(3.6)
	DR	2.6	1.8	0.8	2.6	1.8	0.7	2.8	2.3	1.6	3.1	2.9	2.1
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.3)	(0.8)	(0.7)	(1.2)	(1.4)
Ratio 3:2		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	15.0	10.1	5.9	14.2	9.2	5.0	14.7	9.4	5.2	18.1	14.2	13.0
		(-0.1)	(-0.2)	(-0.1)	(-1.9)	(-2.0)	(-1.9)	(-0.7)	(-0.9)	(-1.0)	(-0.7)	(1.3)	(4.7)
	IPW 2	6.1	4.6	1.4	5.9	5.0	1.8	7.0	6.2	4.2	6.9	6.8	4.5
	(-0.4)	(-0.4)	(-0.5)	(-0.2)	(-0.2)	(-0.1)	(0.2)	(0.8)	(2.2)	(1.0)	(1.8)	(2.7)	
	IPW 3	5.1	3.9	1.1	4.9	4.1	1.2	5.7	5.0	3.0	5.7	5.5	3.4
		(-0.2)	(-0.2)	(-0.2)	(-0.1)	(-0.1)	(-0.1)	(0.3)	(0.8)	(1.7)	(0.9)	(1.5)	(2.2)
	DR	4.2	3.5	0.9	4.1	3.5	0.8	4.2	3.6	1.4	4.5	4.2	1.5
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.1)	(0.5)	(0.5)	(0.8)	(0.9)
Ratio 2:3		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	21.0	14.6	11.1	23.0	16.0	11.7	22.6	16.3	12.6	26.9	19.6	15.0
		(0.0)	(0.0)	(0.0)	(-1.2)	(-0.9)	(-0.7)	(0.4)	(1.1)	(1.4)	(-1.8)	(-0.6)	(0.8)
	IPW 2	10.1	6.4	5.1	10.0	6.9	6.1	7.7	3.2	0.2	10.2	7.5	7.0
	(-0.6)	(-0.9)	(-0.9)	(-0.1)	(0.0)	(0.0)	(-3.2)	(-4.2)	(-5.6)	(0.1)	(0.6)	(1.0)	
	IPW 3	6.7	4.3	4.1	6.6	4.6	4.7	5.2	2.3	1.1	6.7	5.0	5.5
		(-0.3)	(-0.5)	(-0.5)	(-0.1)	(0.0)	(0.0)	(-1.9)	(-2.4)	(-3.3)	(0.1)	(0.5)	(0.8)
	DR	2.4	1.7	2.9	2.2	1.6	3.0	2.7	2.2	3.5	2.1	1.7	3.2
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.5)	(0.6)	(0.1)	(0.3)	(0.3)
Ratio 2:3		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	15.1	10.6	7.2	17.0	12.1	7.7	16.4	12.0	8.4	20.5	15.2	10.4
		(0.0)	(0.0)	(0.0)	(-0.7)	(-0.6)	(-0.5)	(0.3)	(0.7)	(0.9)	(-1.2)	(-0.4)	(0.5)
	IPW 2	6.0	3.2	2.2	6.2	3.9	2.5	4.5	1.3	-0.5	6.3	4.2	3.2
	(-0.4)	(-0.5)	(-0.5)	(-0.1)	(0.0)	(0.0)	(-1.9)	(-2.4)	(-3.2)	(0.1)	(0.4)	(0.6)	
	IPW 3	5.1	2.9	2.1	5.2	3.4	2.3	4.2	1.8	0.5	5.3	3.7	2.8
		(-0.2)	(-0.3)	(-0.2)	(0.0)	(0.0)	(0.0)	(-1.1)	(-1.3)	(-1.8)	(0.1)	(0.3)	(0.5)
	DR	4.3	3.0	2.2	4.2	2.9	2.2	4.5	3.4	2.5	4.1	2.9	2.3
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.4)	(0.4)	(0.1)	(0.2)	(0.2)

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^*))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.7: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores, $MSE(\hat{p}_i^s)$ -minimizing λ

	N	linear			nonlinear			linear			nonlinear				
		100	200	500	100	200	500	100	200	500	100	200	500		
Ratio 1:1	homogeneous, homoscedastic	IPW 1	55.2	64.1	66.9	55.8	64.9	64.5	55.7	64.9	66.9	62.8	71.7	70.2	
			(-0.1)	(-0.1)	(0.0)	(-1.6)	(-1.3)	(-0.6)	(-0.1)	(-0.2)	(0.0)	(-1.7)	(-0.7)	(0.1)	
			23.3	28.6	37.3	23.9	29.2	38.6	23.0	28.7	38.4	23.8	29.0	35.0	
		IPW 2	(-0.8)	(-0.5)	(-1.2)	(-0.1)	(-0.2)	(0.0)	(-1.9)	(-1.3)	(-2.1)	(0.1)	(-0.1)	(-3.1)	
			IPW 3	11.9	15.3	19.0	12.5	15.8	20.9	11.0	14.8	18.2	12.5	15.4	15.9
				(-0.3)	(-0.1)	(-0.5)	(0.0)	(0.0)	(0.0)	(-0.7)	(-0.3)	(-0.9)	(0.0)	(-0.3)	(-4.6)
	DR	13.8		17.6	22.6	14.5	18.4	24.1	13.5	17.6	22.7	14.9	18.3	19.5	
		(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(-0.1)	(0.1)	(0.3)	(0.1)	(-0.7)	(-1.2)	(-5.8)		
		heterogeneous, homoscedastic	IPW 1	38.1	50.3	53.0	39.9	52.2	50.4	39.1	51.6	53.3	48.9	61.0	58.0
	(-0.1)			(-0.1)	(0.0)	(-1.3)	(-1.1)	(-0.4)	(-0.1)	(-0.1)	(0.0)	(-1.4)	(-0.6)	(0.1)	
	7.8			15.9	22.8	8.3	16.2	23.0	7.9	16.1	25.0	8.4	16.1	20.3	
	IPW 2		(-0.5)	(-0.3)	(-1.0)	(-0.1)	(-0.1)	(0.0)	(-1.2)	(-0.9)	(-1.7)	(0.1)	(0.0)	(-2.4)	
IPW 3			1.1	6.0	7.0	1.6	6.3	8.4	0.6	5.6	6.9	1.7	6.1	5.1	
			(-0.2)	(-0.1)	(-0.5)	(0.0)	(0.0)	(0.0)	(-0.4)	(-0.2)	(-0.7)	(0.0)	(-0.2)	(-3.4)	
	DR	4.2	8.3	9.4	4.8	8.9	11.3	4.0	8.3	9.4	5.3	8.9	8.7		
(0.0)		(0.1)	(-0.1)	(0.0)	(0.0)	(-0.1)	(0.1)	(0.2)	(0.0)	(-0.4)	(-0.8)	(-4.3)			
Ratio 3:2		homogeneous, homoscedastic	IPW 1	47.9	86.0	43.7	41.7	69.1	40.2	46.9	85.3	40.0	46.1	70.0	47.1
	(-0.1)			(0.0)	(-0.2)	(-2.4)	(-1.1)	(-2.4)	(-0.9)	(-0.4)	(-3.7)	(-2.6)	(-0.6)	(-0.1)	
	20.8			22.0	28.3	21.5	23.4	31.1	21.5	22.7	28.8	20.9	22.5	28.2	
	IPW 2		(-0.9)	(-0.7)	(-0.4)	(-0.2)	(-0.2)	(-0.2)	(-0.6)	(-0.1)	(-0.1)	(0.1)	(-0.5)	(-2.4)	
			IPW 3	12.3	11.7	17.1	12.9	12.9	17.9	12.6	11.9	16.5	12.3	11.6	14.0
				(-0.4)	(-0.3)	(-0.2)	(0.0)	(0.0)	(0.2)	(0.0)	(0.2)	(-0.7)	(0.0)	(-0.8)	(-3.3)
	DR	12.8		12.8	19.5	13.1	13.3	20.1	12.3	11.7	16.6	11.8	11.1	15.1	
		(0.0)	(0.1)	(0.2)	(0.0)	(0.0)	(0.1)	(-0.4)	(-1.0)	(-2.6)	(-1.1)	(-2.3)	(-5.3)		
		heterogeneous, homoscedastic	IPW 1	31.9	78.0	31.0	26.5	55.8	28.4	31.4	77.3	28.7	31.1	57.1	35.0
	(0.0)			(0.0)	(-0.1)	(-1.6)	(-1.0)	(-1.7)	(-0.6)	(-0.4)	(-2.6)	(-1.8)	(-0.6)	(-0.1)	
	6.4			8.6	15.9	7.0	9.8	18.5	7.0	9.3	16.3	6.5	9.2	16.3	
	IPW 2		(-0.6)	(-0.5)	(-0.4)	(-0.1)	(-0.1)	(-0.1)	(-0.4)	(-0.1)	(-0.1)	(0.1)	(-0.3)	(-1.8)	
IPW 3			1.2	1.7	7.8	1.6	2.5	8.6	1.4	1.9	7.5	1.2	1.6	5.8	
			(-0.3)	(-0.2)	(-0.2)	(0.0)	(0.0)	(0.1)	(0.0)	(0.1)	(-0.4)	(0.0)	(-0.5)	(-2.4)	
	DR	1.9	2.7	10.1	2.1	3.1	10.4	1.6	2.0	8.2	1.3	1.7	6.9		
(0.0)		(0.0)	(0.1)	(0.0)	(0.0)	(0.1)	(-0.2)	(-0.7)	(-1.7)	(-0.8)	(-1.5)	(-3.9)			
Ratio 2:3		homogeneous, homoscedastic	IPW 1	71.1	55.9	44.6	75.9	61.0	48.2	72.2	57.0	43.7	80.3	67.1	56.7
	(0.0)			(0.0)	(0.4)	(-0.4)	(-0.6)	(-0.2)	(0.0)	(-0.2)	(-1.0)	(-0.7)	(-0.8)	(-0.2)	
	20.3			23.0	29.7	21.3	24.1	29.1	17.8	19.7	24.9	21.3	24.2	28.8	
	IPW 2		(-1.1)	(-0.8)	(-0.9)	(-0.1)	(-0.1)	(0.2)	(-4.5)	(-4.9)	(-7.7)	(0.0)	(0.0)	(-0.2)	
			IPW 3	11.4	12.8	15.7	12.1	13.6	16.4	8.6	9.2	8.9	12.2	13.6	15.7
				(-0.6)	(-0.4)	(-0.6)	(0.0)	(0.0)	(0.1)	(-3.2)	(-3.8)	(-7.1)	(0.0)	(0.0)	(-0.5)
	DR	12.7		14.9	17.9	13.1	15.2	18.5	12.4	14.3	15.2	13.8	15.8	18.3	
		(0.0)	(0.0)	(0.2)	(0.0)	(0.0)	(0.0)	(-0.2)	(-0.6)	(-2.3)	(0.0)	(-0.1)	(-1.0)		
		heterogeneous, homoscedastic	IPW 1	57.0	42.1	29.6	64.1	48.3	33.7	59.0	43.7	29.2	71.2	56.5	43.8
	(0.0)			(0.0)	(0.3)	(-0.3)	(-0.5)	(-0.1)	(0.0)	(-0.1)	(-0.9)	(-0.6)	(-0.7)	(-0.1)	
	7.4			12.3	17.8	7.9	12.8	16.9	6.0	10.5	15.4	8.0	12.9	16.8	
	IPW 2		(-0.6)	(-0.5)	(-0.7)	(0.0)	(-0.1)	(0.1)	(-2.8)	(-3.1)	(-5.5)	(0.0)	(0.0)	(-0.2)	
IPW 3			2.9	5.9	6.8	3.3	6.2	7.2	1.3	3.8	2.6	3.4	6.3	6.9	
			(-0.3)	(-0.2)	(-0.5)	(0.0)	(0.0)	(0.1)	(-1.9)	(-2.3)	(-4.8)	(0.0)	(0.0)	(-0.4)	
	DR	6.9	8.7	8.4	7.1	8.9	9.5	6.7	8.5	6.4	7.5	9.3	9.8		
(0.0)		(0.0)	(0.1)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.3)	(-1.7)	(0.0)	(0.0)	(-0.7)			

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{p})) - \text{bias}^2(\text{ATE}(\hat{p}^s))}{\text{bias}^2(\text{ATE}(\hat{p})) + \text{Var}(\text{ATE}(\hat{p}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.8: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores with trimming rule 1, $MSE(\hat{\rho}_i^s)$ -minimizing λ

	N	linear		nonlinear			linear		nonlinear				
		100	200	500	100	200	500	100	200	500			
Ratio 1:1		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	28.6	18.1	18.6	30.4	17.0	18.1	35.4	23.2	22.2	34.7	20.9	24.1
		(-0.1)	(-0.2)	(0.1)	(-2.6)	(-3.2)	(-1.5)	(-0.1)	(-0.3)	(0.0)	(-3.4)	(-2.4)	(1.7)
	IPW 2	17.2	10.7	9.5	17.8	10.8	11.7	16.7	10.6	9.5	19.4	12.5	10.0
	(-1.3)	(-0.8)	(-1.7)	(-0.2)	(-0.3)	(0.0)	(-2.8)	(-2.0)	(-3.1)	(1.4)	(1.0)	(-1.6)	
	IPW 3	16.5	8.8	8.3	16.8	8.8	9.5	16.6	9.3	8.6	18.1	10.2	7.1
		(-0.6)	(-0.3)	(-0.9)	(-0.1)	(-0.2)	(0.0)	(-1.4)	(-0.8)	(-1.6)	(1.3)	(0.9)	(-2.2)
	DR	12.5	4.8	7.4	12.1	4.9	7.3	13.2	5.7	8.7	12.8	5.4	3.4
		(0.0)	(0.0)	(-0.1)	(0.0)	(0.0)	(-0.1)	(0.0)	(0.0)	(-0.1)	(1.0)	(0.2)	(-3.9)
Ratio 1:1		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	21.6	11.7	9.5	23.2	11.2	9.3	26.4	15.2	12.7	26.7	14.3	14.1
		(-0.1)	(-0.1)	(0.0)	(-1.7)	(-2.0)	(-0.9)	(-0.1)	(-0.2)	(0.0)	(-2.3)	(-1.6)	(1.2)
	IPW 2	12.4	5.3	2.8	12.8	5.5	4.0	12.2	5.3	3.5	13.8	6.6	2.8
	(-0.7)	(-0.5)	(-1.2)	(-0.2)	(-0.2)	(0.0)	(-1.6)	(-1.2)	(-2.2)	(0.7)	(0.6)	(-1.3)	
	IPW 3	13.3	4.9	2.6	13.4	5.0	3.4	13.3	5.2	3.4	14.2	5.9	1.7
		(-0.3)	(-0.2)	(-0.7)	(-0.1)	(-0.1)	(0.0)	(-0.8)	(-0.5)	(-1.1)	(0.7)	(0.5)	(-1.7)
	DR	13.4	3.9	2.3	13.2	4.1	2.4	13.8	4.3	3.5	13.5	4.5	-0.2
		(0.0)	(0.0)	(-0.1)	(0.0)	(0.0)	(-0.1)	(0.0)	(0.0)	(-0.1)	(0.5)	(0.1)	(-2.8)
Ratio 3:2		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	24.1	16.0	13.8	24.4	13.8	10.1	28.5	17.7	11.4	28.4	17.8	15.9
		(-0.1)	(-0.2)	(-0.4)	(-3.1)	(-3.2)	(-3.6)	(-1.1)	(-2.2)	(-5.3)	(-2.7)	(-0.7)	(1.0)
	IPW 2	15.2	9.8	8.7	15.7	9.8	9.3	16.3	11.4	10.1	17.1	11.1	7.9
	(-1.1)	(-0.9)	(-0.5)	(-0.3)	(-0.3)	(-0.2)	(-0.9)	(-0.2)	(0.1)	(1.2)	(0.8)	(-2.0)	
	IPW 3	14.7	8.8	7.5	15.0	8.6	7.7	15.8	10.0	7.9	16.2	9.3	5.5
		(-0.6)	(-0.5)	(-0.2)	(-0.2)	(-0.1)	(-0.1)	(-0.3)	(0.1)	(-0.5)	(1.1)	(0.5)	(-2.6)
	DR	11.2	5.3	5.8	11.0	5.4	5.7	11.4	4.8	3.7	11.0	4.6	1.1
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.2)	(-0.9)	(-2.8)	(0.4)	(-0.8)	(-4.7)
Ratio 3:2		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	18.9	11.1	8.0	19.1	9.5	5.7	22.1	12.5	6.6	21.8	12.0	9.5
		(-0.1)	(-0.1)	(-0.2)	(-1.8)	(-2.1)	(-2.2)	(-0.7)	(-1.4)	(-3.3)	(-1.7)	(-0.6)	(0.6)
	IPW 2	10.4	5.3	3.4	10.5	5.0	3.9	11.2	6.6	4.4	11.3	5.6	3.0
	(-0.6)	(-0.5)	(-0.4)	(-0.1)	(-0.2)	(-0.1)	(-0.5)	(-0.1)	(0.1)	(0.7)	(0.5)	(-1.5)	
	IPW 3	10.6	5.1	2.9	10.6	4.8	3.2	11.3	6.1	3.2	11.2	5.1	1.7
		(-0.4)	(-0.3)	(-0.2)	(-0.1)	(-0.1)	(0.0)	(-0.2)	(0.1)	(-0.2)	(0.6)	(0.3)	(-1.9)
	DR	9.6	4.0	2.6	9.5	4.1	2.6	9.8	3.9	1.2	9.4	3.5	-0.4
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.5)	(-1.6)	(0.2)	(-0.5)	(-3.2)
Ratio 2:3		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	26.8	15.9	13.7	30.4	16.9	14.0	33.0	20.1	14.3	34.7	19.4	16.9
		(0.1)	(0.1)	(0.0)	(-1.1)	(-1.3)	(-0.9)	(-0.2)	(-0.5)	(-2.1)	(-2.8)	(-2.5)	(-0.9)
	IPW 2	15.1	9.8	9.6	15.6	10.1	10.9	12.6	5.8	1.6	16.0	10.7	11.4
	(-1.2)	(-1.0)	(-1.3)	(-0.1)	(-0.1)	(0.0)	(-4.6)	(-5.7)	(-9.8)	(0.6)	(0.6)	(0.3)	
	IPW 3	14.9	8.8	8.5	15.1	9.0	9.3	13.4	6.0	1.8	15.5	9.6	9.6
		(-0.7)	(-0.5)	(-0.7)	(0.0)	(0.0)	(0.0)	(-3.1)	(-4.0)	(-7.5)	(0.6)	(0.5)	(0.2)
	DR	11.7	5.6	7.0	11.4	5.7	7.4	12.2	5.4	4.1	11.5	6.2	7.4
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.2)	(-0.7)	(-3.0)	(0.7)	(0.5)	(-0.2)
Ratio 2:3		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	19.7	11.1	8.2	22.5	12.0	8.6	24.0	14.1	9.1	26.3	14.3	11.2
		(0.1)	(0.0)	(0.0)	(-0.7)	(-0.9)	(-0.5)	(-0.1)	(-0.3)	(-1.5)	(-1.9)	(-1.7)	(-0.6)
	IPW 2	11.1	5.7	4.4	11.3	5.7	4.9	9.6	3.5	-0.2	11.6	6.1	5.1
	(-0.7)	(-0.6)	(-0.8)	(0.0)	(-0.1)	(0.0)	(-2.6)	(-3.3)	(-6.3)	(0.3)	(0.3)	(0.2)	
	IPW 3	11.8	5.8	4.0	11.9	5.8	4.4	10.9	4.3	0.3	12.1	6.2	4.5
		(-0.4)	(-0.3)	(-0.5)	(0.0)	(0.0)	(0.0)	(-1.8)	(-2.3)	(-4.8)	(0.3)	(0.3)	(0.1)
	DR	12.6	5.7	3.6	12.4	5.8	3.9	12.9	5.6	2.2	12.5	6.1	3.8
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.4)	(-2.0)	(0.4)	(0.3)	(-0.2)

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^s))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.9: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores with trimming rule 2, $MSE(\hat{\rho}_i^s)$ -minimizing λ

	N	linear			nonlinear			linear			nonlinear			
		100	200	500	100	200	500	100	200	500	100	200	500	
Ratio 1:1	homogeneous, homoscedastic													
		heterogeneous, homoscedastic												
		IPW 1	24.6	14.4	12.4	24.6	13.3	13.0	25.7	15.0	13.4	29.4	19.6	23.1
			(-0.1)	(-0.3)	(0.1)	(-3.1)	(-3.6)	(-1.5)	(-0.2)	(-0.5)	(0.1)	(-2.3)	(-0.6)	(6.5)
		IPW 2	12.4	7.5	3.6	13.2	7.6	5.5	11.1	6.7	2.0	14.5	9.8	9.5
			(-1.4)	(-0.7)	(-1.8)	(-0.3)	(-0.4)	(0.1)	(-3.2)	(-2.0)	(-3.4)	(1.6)	(2.1)	(4.6)
	IPW 3	8.7	4.3	2.3	9.0	4.3	3.3	7.9	3.9	1.5	10.2	6.1	6.7	
		(-0.7)	(-0.3)	(-0.9)	(-0.2)	(-0.2)	(0.1)	(-1.7)	(-0.9)	(-1.6)	(1.4)	(1.7)	(3.8)	
	DR	4.0	0.9	1.6	3.7	1.0	1.2	4.1	0.9	1.8	4.0	1.8	2.8	
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.8)	(0.8)	(1.9)	
	homogeneous, heteroscedastic													
	heterogeneous, heteroscedastic													
Ratio 1:1	homogeneous, homoscedastic													
		heterogeneous, homoscedastic												
		IPW 1	18.3	11.0	9.2	18.7	10.7	9.7	19.2	11.4	9.9	22.5	15.2	17.0
			(-0.1)	(-0.2)	(0.1)	(-1.9)	(-2.2)	(-0.8)	(-0.1)	(-0.3)	(0.1)	(-1.6)	(-0.4)	(4.5)
		IPW 2	7.3	4.6	1.9	7.8	5.1	2.9	6.7	4.2	1.0	8.6	6.5	5.5
			(-0.7)	(-0.4)	(-1.2)	(-0.2)	(-0.2)	(0.1)	(-1.7)	(-1.2)	(-2.2)	(0.9)	(1.3)	(2.8)
	IPW 3	6.5	3.7	1.6	6.7	4.0	2.1	6.1	3.5	1.1	7.4	5.1	4.2	
		(-0.4)	(-0.2)	(-0.6)	(-0.1)	(-0.1)	(0.1)	(-0.9)	(-0.5)	(-1.1)	(0.8)	(1.0)	(2.2)	
	DR	6.1	3.4	1.6	5.9	3.5	1.3	6.2	3.4	1.6	6.1	4.0	2.2	
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.4)	(0.5)	(1.0)	
	homogeneous, heteroscedastic													
	heterogeneous, heteroscedastic													
Ratio 3:2	homogeneous, homoscedastic													
		heterogeneous, homoscedastic												
		IPW 1	21.5	13.4	9.4	20.0	11.7	7.3	21.3	12.7	7.5	24.8	19.1	20.3
			(-0.1)	(-0.2)	(-0.4)	(-3.3)	(-3.3)	(-3.8)	(-0.7)	(-1.3)	(-2.6)	(-1.8)	(1.7)	(7.6)
		IPW 2	10.4	6.1	2.8	10.8	6.7	3.0	11.7	8.6	6.8	12.5	9.4	8.4
			(-1.1)	(-1.0)	(-0.6)	(-0.3)	(-0.3)	(-0.4)	(-0.3)	(1.3)	(3.3)	(2.0)	(3.0)	(5.5)
	IPW 3	7.1	3.8	1.0	7.3	4.2	1.0	8.1	5.7	3.8	8.8	6.4	5.6	
		(-0.7)	(-0.5)	(-0.2)	(-0.2)	(-0.2)	(-0.3)	(0.1)	(1.3)	(2.5)	(1.8)	(2.5)	(4.7)	
	DR	2.5	1.0	-1.0	2.4	1.1	-1.3	2.9	1.6	-0.6	2.9	1.9	1.0	
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.4)	(0.1)	(0.9)	(1.1)	(2.5)	
	homogeneous, heteroscedastic													
	heterogeneous, heteroscedastic													
Ratio 3:2	homogeneous, homoscedastic													
		heterogeneous, homoscedastic												
		IPW 1	16.0	9.8	6.5	15.2	8.7	5.3	16.1	9.4	5.4	18.5	13.6	14.1
			(-0.1)	(-0.2)	(-0.2)	(-1.9)	(-2.1)	(-2.2)	(-0.4)	(-0.8)	(-1.6)	(-1.1)	(1.0)	(5.3)
		IPW 2	6.2	3.0	1.2	6.4	3.3	1.5	7.0	4.5	3.7	7.4	5.0	4.9
			(-0.7)	(-0.5)	(-0.4)	(-0.2)	(-0.2)	(-0.2)	(-0.2)	(0.8)	(2.0)	(1.2)	(1.7)	(3.5)
	IPW 3	5.2	2.3	0.7	5.3	2.5	0.8	5.8	3.5	2.4	6.1	3.8	3.7	
		(-0.4)	(-0.3)	(-0.2)	(-0.1)	(-0.1)	(-0.1)	(0.1)	(0.8)	(1.5)	(1.0)	(1.5)	(3.0)	
	DR	4.6	2.1	0.6	4.5	2.1	0.5	4.9	2.4	0.8	4.8	2.6	1.8	
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.3)	(0.1)	(0.5)	(0.6)	(1.5)	
	homogeneous, heteroscedastic													
	heterogeneous, heteroscedastic													
Ratio 2:3	homogeneous, homoscedastic													
		heterogeneous, homoscedastic												
		IPW 1	21.2	13.9	9.1	23.3	15.4	10.0	23.2	15.8	10.4	27.2	18.9	13.3
			(0.0)	(0.0)	(0.0)	(-1.5)	(-1.5)	(-1.0)	(0.9)	(1.1)	(1.3)	(-2.5)	(-1.8)	(0.0)
		IPW 2	8.1	5.7	2.9	9.0	6.9	4.5	5.2	2.2	-2.3	9.2	7.4	5.6
			(-1.4)	(-1.1)	(-1.3)	(-0.1)	(-0.1)	(0.0)	(-4.8)	(-4.8)	(-6.4)	(0.3)	(0.5)	(1.2)
	IPW 3	5.4	3.5	2.0	5.9	4.1	2.8	3.6	1.4	-1.2	6.0	4.5	3.7	
		(-0.8)	(-0.6)	(-0.7)	(0.0)	(-0.1)	(0.0)	(-2.9)	(-2.8)	(-3.6)	(0.3)	(0.5)	(1.0)	
	DR	1.1	0.7	0.9	0.9	0.7	0.8	1.7	1.3	1.6	0.7	0.8	1.2	
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.4)	(0.5)	(0.8)	(0.2)	(0.2)	(0.5)	
	homogeneous, heteroscedastic													
	heterogeneous, heteroscedastic													
Ratio 2:3	homogeneous, homoscedastic													
		heterogeneous, homoscedastic												
		IPW 1	15.6	10.3	6.7	17.2	11.5	7.6	17.1	11.7	7.7	20.5	14.4	10.3
			(0.0)	(0.0)	(0.0)	(-1.0)	(-1.0)	(-0.6)	(0.5)	(0.7)	(0.8)	(-1.7)	(-1.2)	(0.1)
		IPW 2	5.1	3.0	1.5	5.5	3.6	2.5	3.5	1.1	-1.8	5.6	4.0	3.1
			(-0.8)	(-0.6)	(-0.8)	(0.0)	(-0.1)	(0.0)	(-2.7)	(-2.7)	(-4.0)	(0.1)	(0.3)	(0.7)
	IPW 3	4.6	2.6	1.6	4.8	2.8	2.0	3.6	1.5	-0.4	4.8	3.2	2.6	
		(-0.4)	(-0.3)	(-0.4)	(0.0)	(0.0)	(0.0)	(-1.6)	(-1.6)	(-2.3)	(0.1)	(0.3)	(0.6)	
	DR	4.1	2.4	1.6	3.9	2.4	1.5	4.3	2.9	1.9	3.8	2.6	1.7	
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.3)	(0.4)	(0.1)	(0.2)	(0.3)	

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^s))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.10: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores, data-driven λ

N	linear			nonlinear			linear			nonlinear			
	100	200	500	100	200	500	100	200	500	100	200	500	
Ratio 1:1	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	55.4	56.9	81.3	53.5	53.4	73.0	55.4	56.8	80.8	58.9	59.1	73.5
		(0.0)	(0.1)	(0.1)	(0.0)	(0.2)	(0.1)	(0.0)	(0.2)	(0.1)	(0.4)	(0.6)	(-0.6)
	IPW 2	18.2	24.1	41.0	19.4	22.7	36.9	18.9	25.5	42.0	18.5	21.4	32.9
	(0.1)	(0.0)	(0.1)	(0.1)	(0.2)	(0.0)	(0.1)	(0.1)	(0.3)	(-0.6)	(-0.9)	(-3.8)	
	IPW 3	9.0	10.5	18.0	9.8	10.8	18.8	9.1	10.5	17.6	8.9	9.4	13.8
		(0.1)	(0.1)	(0.4)	(0.1)	(0.3)	(0.0)	(0.3)	(0.3)	(0.7)	(-0.8)	(-1.2)	(-5.0)
	DR	13.0	14.4	22.4	13.8	15.1	23.8	13.2	14.4	22.0	13.6	14.6	20.4
		(0.0)	(0.2)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.4)	(0.1)	(-1.2)	(-1.9)	(-4.8)
Ratio 1:1	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	39.2	41.9	71.7	37.6	38.8	61.4	39.6	42.2	71.5	45.0	46.3	63.4
		(0.0)	(0.1)	(0.1)	(-0.1)	(0.2)	(0.1)	(0.0)	(0.1)	(0.1)	(0.3)	(0.5)	(-0.6)
	IPW 2	6.0	10.5	24.9	6.3	9.0	21.9	6.8	11.7	26.6	5.9	8.2	18.7
	(0.1)	(0.0)	(0.1)	(0.0)	(0.1)	(0.0)	(0.2)	(0.0)	(0.2)	(-0.3)	(-0.6)	(-2.9)	
	IPW 3	-0.8	0.0	6.8	-0.6	0.0	7.8	-0.6	-0.1	6.8	-0.9	-0.8	4.2
		(0.1)	(0.0)	(0.2)	(0.1)	(0.2)	(0.0)	(0.3)	(0.2)	(0.4)	(-0.4)	(-0.8)	(-3.5)
	DR	1.8	2.7	10.0	2.4	3.2	11.6	2.0	2.7	9.9	2.6	3.1	9.4
		(0.0)	(0.2)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.3)	(0.1)	(-0.7)	(-1.3)	(-3.5)
Ratio 3:2	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	35.8	39.6	35.3	32.4	36.2	28.3	34.1	37.9	32.6	36.1	41.2	27.8
		(0.0)	(0.0)	(0.0)	(-1.0)	(-0.5)	(-0.1)	(-0.8)	(-1.3)	(-1.9)	(-0.2)	(-0.1)	(-3.5)
	IPW 2	18.5	23.3	20.9	20.0	24.1	21.1	19.1	23.5	20.5	18.8	21.7	12.8
	(-0.1)	(-0.1)	(-0.1)	(0.0)	(0.1)	(0.0)	(0.2)	(-0.1)	(-1.1)	(-0.3)	(-1.7)	(-7.5)	
	IPW 3	10.8	12.9	10.7	11.3	13.8	11.5	11.0	12.1	9.4	10.0	11.1	2.1
		(0.0)	(0.0)	(0.0)	(0.1)	(0.2)	(0.1)	(0.3)	(-0.2)	(-1.2)	(-0.4)	(-2.1)	(-8.5)
	DR	13.0	16.2	13.7	13.4	16.6	14.0	12.4	14.6	11.7	11.8	13.7	4.7
		(0.0)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.6)	(-1.1)	(-2.3)	(-1.5)	(-3.1)	(-9.1)
Ratio 3:2	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	20.2	23.9	19.7	17.6	21.2	13.8	19.2	23.0	17.8	21.3	26.5	14.6
		(0.0)	(0.0)	(0.0)	(-0.7)	(-0.3)	(0.0)	(-0.7)	(-1.0)	(-1.4)	(-0.2)	(-0.1)	(-2.7)
	IPW 2	4.6	9.0	6.7	5.9	9.8	7.6	5.2	9.3	6.5	5.2	8.2	2.0
	(-0.1)	(-0.1)	(-0.1)	(0.0)	(0.1)	(0.0)	(0.2)	(-0.1)	(-0.7)	(-0.2)	(-1.1)	(-5.2)	
	IPW 3	-0.6	1.5	-1.1	-0.2	2.2	0.1	-0.5	1.1	-2.1	-0.9	0.4	-6.0
		(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(0.2)	(-0.1)	(-0.8)	(-0.2)	(-1.3)	(-5.8)
	DR	0.3	3.3	1.0	0.6	3.7	1.2	-0.1	2.4	-0.3	-0.2	1.8	-5.0
		(0.0)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.4)	(-0.8)	(-1.5)	(-0.9)	(-1.9)	(-6.3)
Ratio 2:3	homogeneous, homoscedastic						heterogeneous, homoscedastic						
	IPW 1	48.8	41.4	37.6	56.1	46.3	42.1	50.4	42.1	35.7	64.1	53.5	50.1
		(0.1)	(0.2)	(0.0)	(-0.1)	(0.1)	(0.1)	(0.0)	(0.1)	(-2.2)	(0.0)	(0.4)	(0.0)
	IPW 2	16.9	19.1	24.5	17.8	20.0	24.8	15.6	18.2	22.6	17.7	20.2	24.1
	(-0.2)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-2.1)	(-1.7)	(-3.8)	(0.0)	(0.1)	(-0.8)	
	IPW 3	9.7	10.2	13.8	10.2	10.8	14.3	8.1	8.4	9.1	10.1	11.1	13.4
		(-0.1)	(0.0)	(0.1)	(0.0)	(0.0)	(0.0)	(-1.8)	(-1.5)	(-4.1)	(-0.1)	(0.1)	(-1.0)
	DR	12.6	13.0	17.4	12.8	13.5	18.1	12.7	12.7	14.3	13.2	14.3	18.0
		(0.0)	(0.2)	(0.0)	(0.0)	(0.0)	(0.1)	(-0.2)	(0.0)	(-3.1)	(-0.1)	(-0.1)	(-0.9)
Ratio 2:3	homogeneous, heteroscedastic						heterogeneous, heteroscedastic						
	IPW 1	33.4	27.5	23.4	41.7	33.0	28.3	35.6	28.6	22.4	52.1	41.6	37.7
		(0.0)	(0.1)	(0.0)	(-0.1)	(0.0)	(0.1)	(0.0)	(0.0)	(-1.7)	(0.0)	(0.3)	(0.0)
	IPW 2	6.3	7.8	13.0	6.8	8.2	13.7	5.9	7.4	12.4	6.8	8.4	13.4
	(-0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-1.2)	(-1.1)	(-2.6)	(0.0)	(0.0)	(-0.5)	
	IPW 3	2.0	1.4	5.9	2.1	1.8	6.6	1.2	0.4	3.2	2.1	2.1	6.1
		(0.0)	(0.0)	(0.1)	(0.0)	(0.0)	(0.0)	(-1.0)	(-0.9)	(-2.6)	(0.0)	(0.0)	(-0.6)
	DR	5.8	4.3	8.8	5.9	4.6	9.7	6.0	4.2	6.9	6.2	5.3	9.9
		(0.0)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(0.0)	(-2.0)	(0.0)	(-0.1)	(-0.5)

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^*))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Table 3.A.11: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores with trimming rule 1, data-driven λ

	N	linear			nonlinear			linear			nonlinear			
		100	200	500	100	200	500	100	200	500	100	200	500	
Ratio 1:1	homogeneous, homoscedastic	IPW 1	14.5	11.2	8.8	17.6	13.4	7.5	22.1	16.5	11.9	21.6	16.6	7.1
			(0.0)	(0.0)	(0.0)	(-0.3)	(-0.3)	(0.0)	(0.0)	(-0.1)	(0.0)	(0.2)	(0.2)	(-1.9)
			IPW 2	8.8	6.7	7.7	8.3	7.6	7.4	9.9	7.2	9.6	9.2	8.6
		(-0.3)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.1)	(-0.6)	(-0.1)	(-0.1)	(0.0)	(1.0)	(0.1)	(-3.1)
		IPW 3	9.9	7.6	7.2	9.7	8.2	7.2	11.1	8.2	8.6	10.4	8.9	4.4
		(-0.2)	(0.0)	(0.0)	(0.0)	(-0.1)	(0.0)	(-0.4)	(0.0)	(0.0)	(0.0)	(0.9)	(0.0)	(-3.1)
	DR	8.3	6.7	6.2	8.0	6.4	6.6	9.4	7.7	7.4	8.2	6.6	3.9	
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.7)	(-0.1)	(-3.1)	
		heterogeneous, homoscedastic												
	Ratio 1:1	IPW 1	11.4	7.6	3.9	13.5	9.0	3.1	16.3	11.0	6.6	16.5	11.5	2.6
			(0.0)	(0.0)	(0.0)	(-0.3)	(-0.2)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.1)	(-1.5)
			IPW 2	7.5	4.9	3.7	7.4	5.3	3.8	8.2	5.1	5.7	8.1	6.0
(-0.1)		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.3)	(-0.1)	(0.0)	(0.6)	(0.0)	(-2.1)		
IPW 3		8.5	5.4	3.8	8.5	5.6	4.0	9.2	5.6	5.3	9.1	6.1	1.8	
(-0.1)		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.2)	(0.0)	(0.0)	(0.5)	(0.0)	(-2.1)		
DR	8.9	5.0	3.5	8.8	4.9	3.9	9.5	5.4	4.8	9.1	5.1	1.8		
	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.4)	(-0.1)	(-2.1)		
	heterogeneous, heteroscedastic													
Ratio 3:2	homogeneous, homoscedastic	IPW 1	16.6	9.4	7.9	17.8	9.9	6.7	21.4	12.2	7.7	21.3	12.2	3.2
			(0.0)	(-0.1)	(0.1)	(-1.2)	(-0.7)	(0.0)	(-1.0)	(-1.8)	(-2.7)	(0.5)	(1.3)	(-2.1)
			IPW 2	12.7	7.1	7.1	12.1	7.4	7.7	14.1	8.1	7.2	13.2	7.4
		(-0.3)	(-0.1)	(-0.1)	(-0.1)	(0.0)	(0.1)	(0.0)	(-0.2)	(-1.4)	(1.0)	(-0.5)	(-5.3)	
		IPW 3	13.2	7.5	5.8	12.8	7.6	6.2	14.4	8.1	5.5	13.7	7.2	0.2
		(-0.2)	(-0.1)	(0.0)	(-0.1)	(0.0)	(0.1)	(0.0)	(-0.3)	(-1.6)	(1.0)	(-0.6)	(-5.6)	
	DR	10.6	5.8	4.9	10.5	5.8	4.9	10.6	5.1	3.5	10.6	4.5	-2.0	
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.5)	(-1.2)	(-2.5)	(0.5)	(-1.2)	(-6.4)	
		heterogeneous, heteroscedastic												
	Ratio 3:2	IPW 1	12.2	4.8	1.0	12.7	5.3	0.4	15.3	6.8	0.9	14.7	7.0	-1.7
			(-0.1)	(0.0)	(0.1)	(-0.8)	(-0.4)	(0.0)	(-0.7)	(-1.1)	(-1.7)	(0.2)	(0.8)	(-1.4)
			IPW 2	9.1	2.6	0.8	8.4	3.0	1.8	10.0	3.2	0.9	9.0	3.1
(-0.1)		(-0.1)	(0.0)	(-0.1)	(0.0)	(0.1)	(0.0)	(-0.1)	(-0.8)	(0.5)	(-0.3)	(-3.4)		
IPW 3		9.5	3.0	0.0	9.1	3.3	0.7	10.3	3.4	-0.3	9.6	3.1	-2.9	
(-0.1)		(0.0)	(0.0)	(-0.1)	(0.0)	(0.0)	(0.0)	(-0.2)	(-1.0)	(0.5)	(-0.3)	(-3.5)		
DR	8.7	2.3	-0.2	8.7	2.3	-0.1	8.7	2.0	-1.2	8.7	1.6	-4.2		
	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.3)	(-0.7)	(-1.6)	(0.2)	(-0.7)	(-4.0)		
	heterogeneous, heteroscedastic													
Ratio 2:3	IPW 1	16.4	10.1	7.1	20.6	12.5	8.0	22.9	14.6	6.9	25.4	15.8	10.4	
		(0.1)	(0.1)	(0.0)	(-0.3)	(-0.1)	(0.0)	(-0.2)	(-0.3)	(-3.3)	(-0.7)	(-0.3)	(-0.2)	
		IPW 2	9.7	6.1	6.7	9.4	6.0	6.6	8.7	4.9	3.5	9.9	6.7	6.4
	(-0.4)	(0.0)	(-0.1)	(0.0)	(-0.1)	(0.0)	(-2.2)	(-1.8)	(-4.7)	(0.6)	(0.4)	(-0.1)		
	IPW 3	10.5	6.3	5.9	10.3	6.3	6.0	9.9	5.6	2.3	10.7	6.9	5.7	
	(-0.3)	(0.0)	(-0.1)	(0.0)	(-0.1)	(0.0)	(-1.8)	(-1.5)	(-4.5)	(0.6)	(0.4)	(-0.2)		
DR	8.1	5.0	4.9	7.9	4.8	5.2	8.5	5.5	2.4	8.2	5.2	5.1		
	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(0.0)	(-0.3)	(-0.5)	(-3.5)	(0.6)	(0.5)	(-0.1)		
	heterogeneous, heteroscedastic													
Ratio 2:3	IPW 1	13.4	6.7	3.1	16.4	8.3	3.9	17.8	9.7	3.4	20.3	11.0	5.8	
		(0.0)	(0.0)	(0.0)	(-0.2)	(-0.1)	(0.0)	(-0.1)	(-0.2)	(-2.1)	(-0.5)	(-0.2)	(-0.1)	
		IPW 2	8.5	4.4	3.4	8.5	4.2	3.5	8.0	3.8	2.0	8.8	4.6	3.5
	(-0.2)	(0.0)	(-0.1)	(0.0)	(-0.1)	(0.0)	(-1.2)	(-1.0)	(-2.8)	(0.3)	(0.3)	(0.0)		
	IPW 3	9.4	4.6	3.2	9.4	4.5	3.5	9.2	4.2	1.5	9.7	4.9	3.4	
	(-0.1)	(0.0)	(0.0)	(0.0)	(-0.1)	(0.0)	(-0.9)	(-0.8)	(-2.7)	(0.3)	(0.3)	(0.0)		
DR	9.9	4.3	3.0	9.8	4.1	3.4	10.2	4.5	1.9	10.1	4.4	3.4		
	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.3)	(-2.0)	(0.4)	(0.3)	(0.0)		

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^*))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

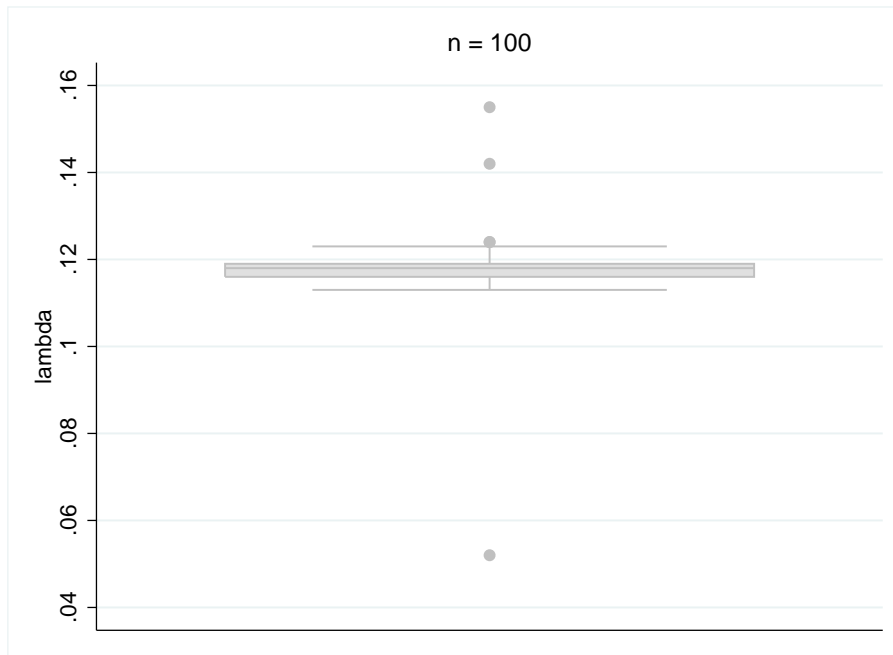
Table 3.A.12: Percentage improvement in MSE for the ATE if the shrunken propensity scores are used with trimming rule 2 instead of the conventional propensity scores with trimming rule 2, data-driven λ

	N	linear			nonlinear			linear			nonlinear		
		100	200	500	100	200	500	100	200	500	100	200	500
Ratio 1:1		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	11.5	7.2	2.2	13.0	8.5	3.6	12.2	7.6	2.4	17.9	13.3	8.4
		(0.0)	(-0.1)	(0.0)	(-0.4)	(-0.6)	(-0.1)	(0.0)	(-0.1)	(0.0)	(1.8)	(2.1)	(3.2)
	IPW 2	4.9	3.3	1.0	5.0	3.2	1.4	4.9	3.2	1.1	5.9	4.3	2.9
	(-0.4)	(0.0)	(0.0)	(0.0)	(-0.1)	(-0.1)	(-0.9)	(-0.1)	(0.0)	(1.1)	(1.0)	(1.5)	
	IPW 3	2.6	1.8	0.6	2.7	1.8	0.8	2.5	1.6	0.5	3.4	2.6	2.0
		(-0.3)	(0.0)	(0.0)	(0.0)	(-0.1)	(0.0)	(-0.6)	(-0.1)	(0.0)	(0.9)	(0.8)	(1.1)
	DR	0.5	0.8	0.4	0.3	0.6	0.3	0.6	0.7	0.4	0.5	0.8	0.9
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.4)	(0.4)	(0.6)
Ratio 1:1		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	8.2	5.3	2.6	9.4	6.2	3.2	8.7	5.5	2.7	13.1	9.7	6.4
		(0.0)	(0.0)	(0.0)	(-0.4)	(-0.3)	(0.0)	(0.0)	(-0.1)	(0.0)	(1.1)	(1.4)	(2.2)
	IPW 2	2.6	2.1	1.4	2.7	2.2	1.3	2.5	2.0	1.5	3.3	2.8	2.0
	(-0.2)	(0.0)	(0.0)	(0.0)	(-0.1)	(0.0)	(-0.4)	(-0.1)	(0.0)	(0.6)	(0.6)	(0.8)	
	IPW 3	2.0	1.6	1.1	2.1	1.7	1.0	1.9	1.5	1.1	2.5	2.2	1.5
		(-0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.3)	(0.0)	(0.0)	(0.5)	(0.4)	(0.6)
	DR	1.8	1.7	1.1	1.7	1.5	1.0	1.8	1.6	1.1	1.8	1.6	1.2
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.2)	(0.3)
Ratio 3:2		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	12.6	6.0	2.7	12.3	5.7	2.3	12.3	5.4	1.9	16.8	10.6	7.0
		(-0.1)	(-0.1)	(0.0)	(-1.3)	(-0.7)	(0.0)	(-0.4)	(-0.6)	(-0.5)	(1.3)	(3.3)	(4.7)
	IPW 2	6.5	3.1	2.3	6.2	2.9	1.9	7.8	4.8	4.2	7.4	4.3	3.3
	(-0.3)	(-0.2)	(-0.2)	(-0.2)	(-0.1)	(0.0)	(0.8)	(1.3)	(1.3)	(1.4)	(1.6)	(1.8)	
	IPW 3	4.2	1.6	1.4	4.1	1.6	1.2	5.1	2.6	2.6	5.0	2.7	2.4
		(-0.2)	(-0.1)	(-0.1)	(-0.1)	(0.0)	(0.0)	(0.6)	(0.9)	(0.8)	(1.2)	(1.3)	(1.4)
	DR	1.2	0.1	0.6	1.0	0.1	0.6	1.4	0.1	0.7	1.3	0.5	1.2
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.1)	(0.1)	(0.5)	(0.5)	(0.6)
Ratio 3:2		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	8.8	4.4	2.2	8.5	4.3	1.8	8.7	4.1	1.8	11.2	7.2	4.6
		(0.0)	(0.0)	(0.0)	(-0.8)	(-0.4)	(0.0)	(-0.2)	(-0.3)	(-0.3)	(0.7)	(1.9)	(3.0)
	IPW 2	3.6	1.9	1.5	3.3	1.9	0.9	4.5	3.0	2.7	3.9	2.6	1.7
	(-0.1)	(-0.1)	(-0.1)	(-0.1)	(0.0)	(0.0)	(0.5)	(0.8)	(0.8)	(0.8)	(0.8)	(1.0)	
	IPW 3	2.9	1.4	1.0	2.7	1.4	0.6	3.4	2.0	1.8	3.2	1.9	1.2
		(-0.1)	(-0.1)	(-0.1)	(-0.1)	(0.0)	(0.0)	(0.4)	(0.5)	(0.5)	(0.6)	(0.7)	(0.8)
	DR	2.4	1.2	0.5	2.3	1.2	0.6	2.5	1.2	0.7	2.4	1.4	0.7
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.3)
Ratio 2:3		homogeneous, homoscedastic						heterogeneous, homoscedastic					
	IPW 1	12.9	7.0	4.1	14.9	8.3	5.6	14.3	7.9	5.1	18.8	11.4	8.9
		(0.0)	(-0.1)	(0.1)	(-0.5)	(-0.6)	(0.2)	(0.7)	(0.4)	(0.9)	(-0.1)	(-0.1)	(1.6)
	IPW 2	5.4	3.3	2.1	5.4	2.8	1.7	4.1	2.2	0.9	5.6	3.1	2.2
	(-0.5)	(0.1)	(-0.1)	(0.0)	(-0.1)	(0.0)	(-2.1)	(-1.3)	(-1.5)	(0.3)	(0.2)	(0.5)	
	IPW 3	3.3	2.0	1.1	3.3	1.8	0.9	2.5	1.3	0.4	3.5	2.0	1.3
		(-0.3)	(0.0)	(-0.1)	(0.0)	(-0.1)	(0.0)	(-1.3)	(-0.8)	(-0.9)	(0.3)	(0.1)	(0.4)
	DR	0.6	0.5	0.4	0.4	0.4	0.3	0.9	0.8	0.6	0.3	0.4	0.5
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.3)	(0.2)	(0.4)	(0.1)	(0.1)	(0.2)
Ratio 2:3		homogeneous, heteroscedastic						heterogeneous, heteroscedastic					
	IPW 1	9.0	6.0	2.7	10.5	6.9	3.7	10.0	6.7	3.3	13.6	9.1	6.1
		(0.0)	(-0.1)	(0.0)	(-0.4)	(-0.4)	(0.1)	(0.4)	(0.2)	(0.5)	(-0.2)	(-0.1)	(1.1)
	IPW 2	2.5	2.5	1.1	2.5	2.1	0.9	1.9	1.9	0.4	2.7	2.3	1.1
	(-0.2)	(0.0)	(-0.1)	(0.0)	(-0.1)	(0.0)	(-1.2)	(-0.7)	(-1.0)	(0.2)	(0.1)	(0.2)	
	IPW 3	2.1	2.1	0.6	2.1	1.8	0.5	1.6	1.7	0.1	2.2	1.9	0.7
		(-0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(-0.7)	(-0.4)	(-0.6)	(0.1)	(0.1)	(0.2)
	DR	1.8	1.7	0.5	1.7	1.6	0.5	2.0	1.9	0.6	1.7	1.6	0.6
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.1)

Note: Percentage change which is due to the bias $\left(\frac{\text{bias}^2(\text{ATE}(\hat{\rho})) - \text{bias}^2(\text{ATE}(\hat{\rho}^*))}{\text{bias}^2(\text{ATE}(\hat{\rho})) + \text{Var}(\text{ATE}(\hat{\rho}))} \right)$ is given in brackets.

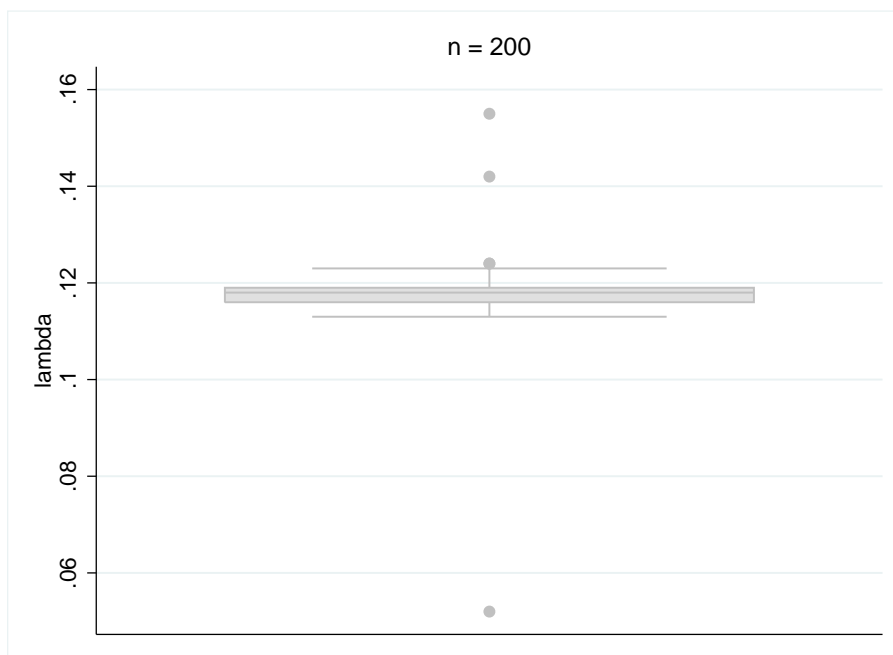
Appendix 3.B Figures

Figure 3.B.1: Individual MSE minimizing λ s for $n = 100$, $\delta = 1$, $\psi = 2$, $\nu = -0.3$, $\kappa = 0.8$ and $m_2(q)$.



Note: For each individual the average is taken over 10000 Monte Carlo samples.

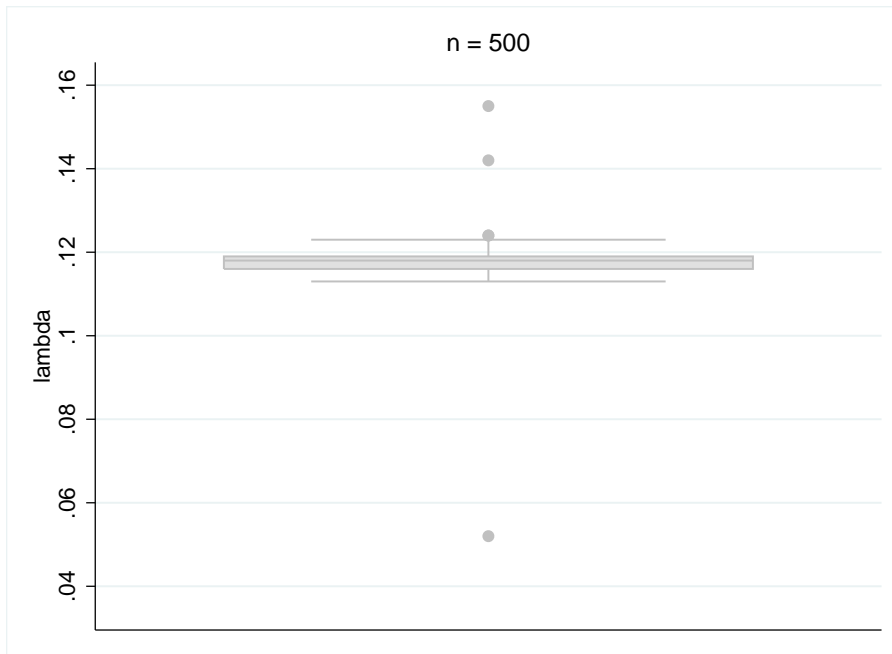
Figure 3.B.2: Individual MSE minimizing λ s for $n = 200$, $\delta = 1$, $\psi = 2$, $\nu = -0.3$, $\kappa = 0.8$ and $m_2(q)$.



Note: For each individual the average is taken over 5000 Monte Carlo samples.

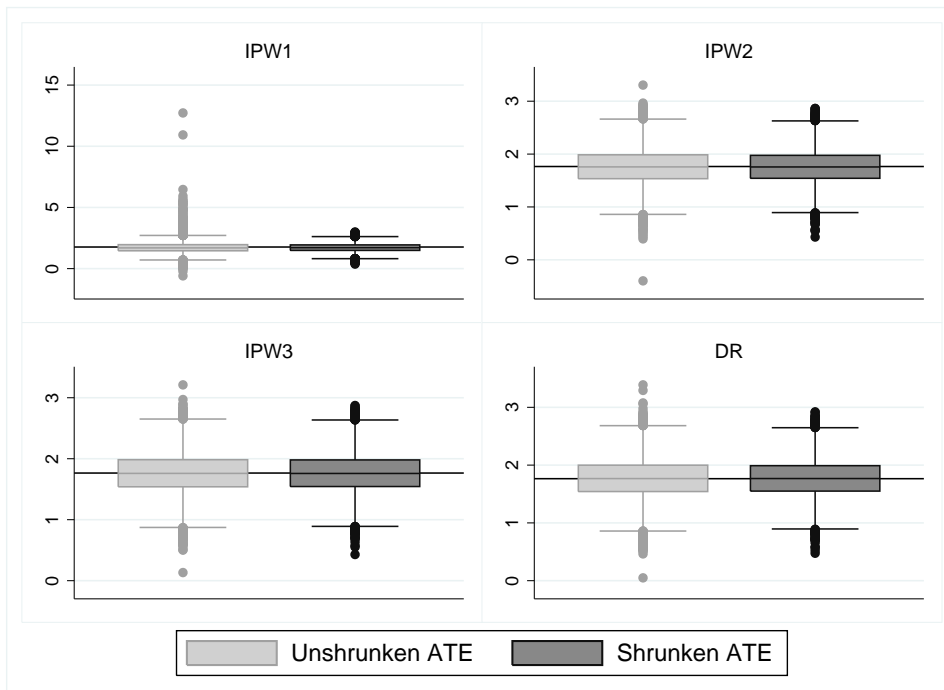
3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Figure 3.B.3: Individual MSE minimizing λ s for $n = 500$, $\delta = 1$, $\psi = 2$, $\nu = -0.3$, $\kappa = 0.8$ and $m_2(q)$.



Note: For each individual the average is taken over 2000 Monte Carlo samples.

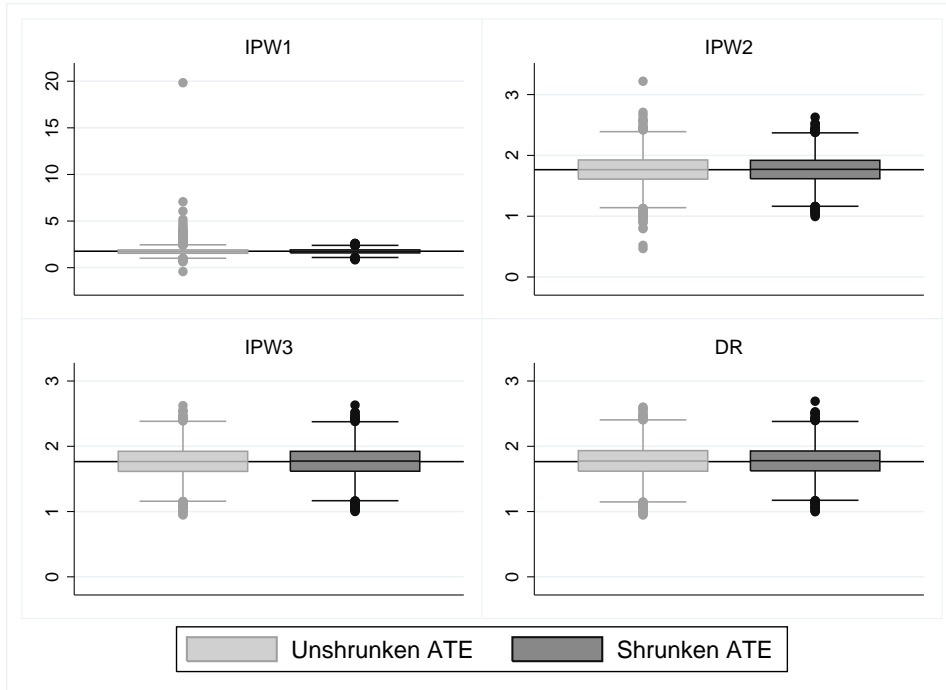
Figure 3.B.4: ATE's for $n = 100$, $\delta = 1$, $\psi = 2$, $\nu = -0.3$, $\kappa = 0.8$ and $m_2(q)$.



Note: For shrunk ATE's fixed valued λ 's are used.

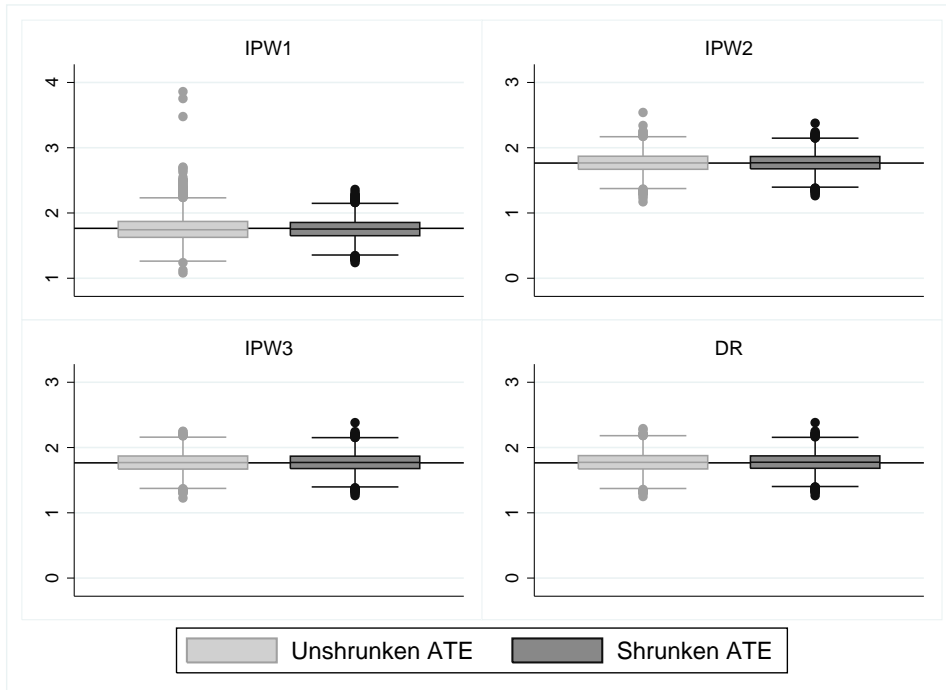
3. A SIMPLE AND SUCCESSFUL METHOD TO SHRINK THE WEIGHT

Figure 3.B.5: ATE's for $n = 200$, $\delta = 1$, $\psi = 2$, $\nu = -0.3$, $\kappa = 0.8$ and $m_2(q)$.



Note: For shrunk ATE's fixed valued λ 's are used.

Figure 3.B.6: ATE's for $n = 500$, $\delta = 1$, $\psi = 2$, $\nu = -0.3$, $\kappa = 0.8$ and $m_2(q)$.



Note: For shrunk ATE's fixed valued λ 's are used.

Appendix 3.C Supplementary Proofs

Theorem 3.C.1. *Under the assumption that $E[\hat{p}_i] \approx p_i$ the MSE minimizing $\lambda_i^*(n)$ is given by*

$$\lambda_i^*(n) = \frac{V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D})}{V[\hat{p}_i] + \frac{E[D_i](1-E[D_i])}{n} + (E[D_i] - E[\hat{p}_i])^2 - 2\text{Cov}(\hat{p}_i, \bar{D})}.$$

Proof 3.C.1.

$$\begin{aligned} \hat{p}_i^s &= (1 - \lambda_i(n))\hat{p}_i + \lambda_i(n)\bar{D} \\ E[\hat{p}_i^s] &= (1 - \lambda_i(n))E[\hat{p}_i] + \lambda_i(n)E[\bar{D}] \\ \text{bias}(\hat{p}_i^s) &= E[\hat{p}_i] - p_i + \lambda_i(n)[E[\bar{D}] - E[\hat{p}_i]] \\ V[\hat{p}_i^s] &= (1 - \lambda_i(n))^2 V[\hat{p}_i] + \lambda_i(n)^2 V[\bar{D}] + 2\lambda_i(n)(1 - \lambda_i(n))\text{Cov}(\hat{p}_i, \bar{D}) \\ \text{MSE}(\hat{p}_i^s) &= [E[\hat{p}_i] - p_i]^2 + 2\lambda_i(n)[E[\hat{p}_i] - p_i][E[\bar{D}] - E[\hat{p}_i]] + \lambda_i(n)^2 [E[\bar{D}] - E[\hat{p}_i]]^2 \\ &\quad + (1 - \lambda_i(n))^2 V[\hat{p}_i] + \lambda_i(n)^2 V[\bar{D}] + 2\lambda_i(n)(1 - \lambda_i(n))\text{Cov}(\hat{p}_i, \bar{D}) \end{aligned}$$

Therefore, we get:

$$\begin{aligned} \frac{\partial \text{MSE}(\hat{p}_i^s)}{\partial \lambda_i(n)} &= 2[E[\hat{p}_i] - p_i][E[\bar{D}] - E[\hat{p}_i]] + 2\lambda_i(n)[E[\bar{D}] - E[\hat{p}_i]]^2 \\ &\quad - 2(1 - \lambda_i(n))V[\hat{p}_i] + 2\lambda_i(n)V[\bar{D}] + 2(1 - 2\lambda_i(n))\text{Cov}(\hat{p}_i, \bar{D}) \\ &\stackrel{!}{=} 0 \\ \Leftrightarrow \lambda_i(n) &= \frac{V[\hat{p}_i] + V[\bar{D}] - 2\text{Cov}(\hat{p}_i, \bar{D}) + [E[\bar{D}] - E[\hat{p}_i]]^2}{V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D}) - [E[\hat{p}_i] - p_i][E[\bar{D}] - E[\hat{p}_i]]} \\ \lambda_i^*(n) &= \frac{V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D}) - [E[\hat{p}_i] - p_i][E[\bar{D}] - E[\hat{p}_i]]}{V[\hat{p}_i] + V[\bar{D}] + (E[D_i] - E[\hat{p}_i])^2 - 2\text{Cov}(\hat{p}_i, \bar{D})} \end{aligned}$$

using $E[\hat{p}_i] \approx p_i$ and $V[\bar{D}] = \frac{E[D_i](1-E[D_i])}{n}$ yields:

$$\lambda_i^*(n) = \frac{V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D})}{V[\hat{p}_i] + \frac{E[D_i](1-E[D_i])}{n} + (E[D_i] - E[\hat{p}_i])^2 - 2\text{Cov}(\hat{p}_i, \bar{D})} \quad \square$$

Theorem 3.C.2. *Under the assumption that $E[\hat{p}_i] \approx p_i$ the $\lambda^*(n)$, which minimizes the sum of individual MSEs is given by*

$$\lambda^*(n) = \frac{\sum_i [V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D})]}{\sum_i \left[V[\hat{p}_i] + \frac{E[D_i](1-E[D_i])}{n} + (E[D_i] - E[\hat{p}_i])^2 - 2\text{Cov}(\hat{p}_i, \bar{D}) \right]}.$$

Proof 3.C.2.

$$\begin{aligned}
 \hat{p}_i^s &= (1 - \lambda(n))\hat{p}_i + \lambda(n)\bar{D} \\
 E[\hat{p}_i^s] &= (1 - \lambda(n))E[\hat{p}_i] + \lambda(n)E[\bar{D}] \\
 \text{bias}(\hat{p}_i^s) &= E[\hat{p}_i] - p_i + \lambda(n)[E[\bar{D}] - E[\hat{p}_i]] \\
 V[\hat{p}_i^s] &= (1 - \lambda(n))^2 V[\hat{p}_i] + \lambda(n)^2 V[\bar{D}] + 2\lambda(n)(1 - \lambda(n)) \text{Cov}(\hat{p}_i, \bar{D}) \\
 \sum_i \text{MSE}(\hat{p}_i^s) &= \sum_i \left[[E[\hat{p}_i] - p_i]^2 + 2\lambda(n)[E[\hat{p}_i] - p_i][E[\bar{D}] - E[\hat{p}_i]] + \lambda(n)^2 [E[\bar{D}] - E[\hat{p}_i]]^2 \right. \\
 &\quad \left. + (1 - \lambda(n))^2 V[\hat{p}_i] + \lambda(n)^2 V[\bar{D}] + 2\lambda(n)(1 - \lambda(n)) \text{Cov}(\hat{p}_i, \bar{D}) \right]
 \end{aligned}$$

Therefore, we get:

$$\begin{aligned}
 \frac{\partial \sum_i \text{MSE}(\hat{p}_i^s)}{\partial \lambda(n)} &= \sum_i \left[2[E[\hat{p}_i] - p_i][E[\bar{D}] - E[\hat{p}_i]] + 2\lambda(n)[E[\bar{D}] - E[\hat{p}_i]]^2 \right. \\
 &\quad \left. - 2(1 - \lambda(n))V[\hat{p}_i] + 2\lambda(n)V[\bar{D}] + 2(1 - 2\lambda(n))\text{Cov}(\hat{p}_i, \bar{D}) \right] \\
 &\stackrel{!}{=} 0 \\
 \Leftrightarrow \lambda(n) &\left[\sum_i \left(V[\hat{p}_i] + V[\bar{D}] - 2\text{Cov}(\hat{p}_i, \bar{D}) + [E[\bar{D}] - E[\hat{p}_i]]^2 \right) \right] \\
 &= \sum_i \left[V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D}) - [E[\hat{p}_i] - p_i][E[\bar{D}] - E[\hat{p}_i]] \right] \\
 \lambda^*(n) &= \frac{\sum_i \left[V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D}) - [E[\hat{p}_i] - p_i][E[\bar{D}] - E[\hat{p}_i]] \right]}{\sum_i \left[V[\hat{p}_i] + V[\bar{D}] + (E[D_i] - E[\hat{p}_i])^2 - 2\text{Cov}(\hat{p}_i, \bar{D}) \right]}
 \end{aligned}$$

using $E[\hat{p}_i] \approx p_i$ and $V[\bar{D}] = \frac{E[D_i](1-E[D_i])}{n}$ yields:

$$\lambda^*(n) = \frac{\sum_i \left[V[\hat{p}_i] - \text{Cov}(\hat{p}_i, \bar{D}) \right]}{\sum_i \left[V[\hat{p}_i] + \frac{E[D_i](1-E[D_i])}{n} + (E[D_i] - E[\hat{p}_i])^2 - 2\text{Cov}(\hat{p}_i, \bar{D}) \right]} \quad \square$$

Complete Bibliography

- ABADIE, A. AND G. W. IMBENS (2006): “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74, 235–267.
- AKKOYUNLU-WIGLEY, A. AND S. WIGLEY (2008): “Basic Education and Capability Development in Turkey,” in *Education in Turkey, Vol. 26*, ed. by A.-M. Nohl, A. Akkoyunlu-Wigley, and S. Wigley, Waxmann Publishing, New York/Münster.
- AMMERMÜLLER, A. (2007): “PISA: What makes the Difference? Explaining the Gap in Test Scores between Finland and Germany,” *Empirical Economics*, 33, 263–287.
- AMMERMÜLLER, A., H. HEIJKE, AND L. WÖSSMANN (2005): “Schooling Quality in Eastern Europe: Educational Production during Transition,” *Economics of Education Review*, 24, 579–599.
- AYPAY, A., M. ERDOĞAN, AND M. SÖZER (2007): “Variation among Schools on Classroom Practices in Science based on TIMSS-1999 in Turkey,” *Journal of Research in Science Teaching*, 44, 1417–1435.
- BARSKY, R., J. BOUND, K. CHARLES, AND J. LUPTON (2002): “Accounting for the Black-White Wealth Gap: A Nonparametric Approach,” *Journal of the American Statistical Association*, 97, 663–673.
- BEIRNE, J. AND N. F. CAMPOS (2007): “Educational Inputs and Outcomes before the Transition from Communism,” *The Economics of Transition*, 15, 57–76.
- BJØRNSKOV, C. AND N. POTRAFKE (2011): “Politics and Privatization in Central and Eastern Europe: A Panel Data Analysis,” *The Economics of Transition*, 19, 201–230.
- BLANCY, N. K. AND A. SASMAZ (2011): “PISA 2009: Where does Turkey Stand?” *Turkish Policy Quarterly*, 10, 125–134.
- BLINDER, A. (1973): “Wage Discrimination: Reduced Form and Structural Estimates,” *Journal of Human Resources*, 8, 436–455.
- BOTEZAT, A. AND R. R. SEIBERLICH (2013): “Educational Performance Gaps in Eastern Europe,” *The Economics of Transition*, forthcoming.
- BROSNAN, M. (2006): “Digit Ratio and Faculty Membership: Implications for the

COMPLETE BIBLIOGRAPHY

- Relationship between Prenatal Testosterone and Academia,” *British Journal of Psychology*, 97, 455–466.
- BUSSO, M., J. DINARDO, AND J. MCCRARY (2009): “Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects,” Unpublished manuscript, http://emlab.berkeley.edu/~jmccrary/BDM_JBES.pdf.
- CAHILL, L. (2012): “His Brain, Her Brain,” *Special Editions, Scientific American*, 1, 46–53.
- CARD, D. AND J. ROTHSTEIN (2007): “Racial Segregation and the Black-White Test Score Gap,” *Journal of Public Economics*, 91, 2158–2184.
- CARNEIRO, P. AND J. HECKMAN (2003): “Human Capital Policy,” *Working Paper 9495*, National Bureau of Economic Research.
- CARRELL, S., M. PAGE, AND J. WEST (2009): “Sex and Science: How Professor Gender Perpetuates the Gender Gap,” *The Quarterly Journal of Economics*, 125, 1101–1144.
- CAYGILL, R. (2003): “PISA 2006: Student Attitudes to and Engagement with Science - How Ready are our 15-year-olds for Tomorrow’s World?” Technical Report, Ministry of Education, Wellington, New Zealand.
- CECI, S., W. WILLIAMS, AND S. BARNETT (2009): “Women’s Underrepresentation in Science: Sociocultural and Biological Considerations,” *Psychological Bulletin*, 135, 218–261.
- CERYCH, L. (1997): “Educational Reforms in Central and Eastern Europe: Processes and Outcomes,” *European Journal of Education*, 32, 75–96.
- CRUMP, R. K., V. J. HOTZ, G. W. IMBENS, AND O. A. MITNIK (2009): “Dealing with Limited Overlap in Estimation of Average Treatment Effects,” *Biometrika*, 96, 187–199.
- DAVISON, K. AND E. SUSMAN (2001): “Are Hormone Levels and Cognitive Ability Related during early Adolescence?” *International Journal of Behavioral Development*, 25, 416–428.
- DAYIOĞLU, M., M. KIRDAR, AND A. TANSEL (2009): “Impact of Sibship Size, Birth Order and Sex Composition on School Enrolment in Urban Turkey,” *Oxford Bulletin of Economics and Statistics*, 71, 399–426.

COMPLETE BIBLIOGRAPHY

- DAYIOĞLU, M. AND S. TÜRÜT-AŞIK (2007): “Gender Differences in Academic Performance in a Large Public University in Turkey,” *Higher Education*, 53, 255–277.
- DEHEJIA, R. H. AND S. WAHBA (1999): “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053–1062.
- DINCER, M. AND G. UYSAL (2010): “The Determinants of Student Achievement in Turkey,” *International Journal of Educational Development*, 30, 592–598.
- DUNCAN, K. C. AND J. SANDY (2007): “Explaining the Performance Gap Between Public and Private School Students,” *Eastern Economic Journal*, 33, 177–191.
- ERBERBER, E. (2010): “Analyzing Turkey’s Data from TIMSS 2007 to Investigate Regional Disparities in Eighth Grade Science Achievement,” in *The Impact of International Achievement Studies on National Education Policymaking (International Perspectives on Education and Society, Vol. 13)*, ed. by A. W. Wiseman, Amsterdam: North-Holland.
- EURYDICE (2010): “Organisation of the Education System in Turkey,” Technical Report, European Commission, Brussels, Belgium.
- FORTIN, N., T. LEMIEUX, AND S. FIRPO (2010): “Decomposition Methods in Economics,” in *Handbook of Labour Economics*, ed. by O. Ashenfelter and D. Card, Amsterdam: North-Holland.
- FRÖLICH, M. (2004): “Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators,” *Review of Economics and Statistics*, 86, 77–90.
- (2007): “Propensity Score Matching without Conditional Independence Assumption—With an Application to the Gender Wage Gap in the United Kingdom,” *Econometrics Journal*, 10, 359–407.
- FRYER, R. G. AND S. D. LEVITT (2010): “An Empirical Analysis of the Gender Gap in Mathematics,” *American Economic Journal: Applied Economics*, 2, 210–240.
- FUCHS, T. AND L. WÖSSMANN (2007): “What accounts for International Differences in Student Performance? A Re-Examination using PISA Data,” *Empirical Economics*, 32, 433–464.

COMPLETE BIBLIOGRAPHY

- GUIO, L., F. MONTE, P. SAPIENZA, AND L. ZINGALES (2008): “Culture, Gender, and Math,” *Science*, 320, 1164–1165.
- HAHN, J. (1998): “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66, 315–331.
- HANUSHEK, E., J. F. KAIN, AND S. G. RIVKIN (2009): “New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement,” *Journal of Labor Economics*, 27, 349–383.
- HANUSHEK, E. A. (2003): “The Failure of Input-Based Schooling Policies,” *The Economic Journal*, 113, F64–F98.
- HANUSHEK, E. A. AND D. D. KIMKO (2000): “Schooling, Labor-Force Quality, and the Growth of Nations,” *The American Economic Review*, 90, 1184–1208.
- HANUSHEK, E.A., V. L. AND K. HITOMI (2008): “Do Students Care about School Quality? Determinants of Dropout Behavior in Developing Countries,” *Journal of Human Capital*, 2, 69–105.
- HECKMAN, J., H. ICHIMURA, AND P. TODD (1997): “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *The Review of Economic Studies*, 64, 605–654.
- HECKMAN, J., J. LALONDE, AND J. SMITH (1999): “The Economics and Econometrics of Active Labor Market Programs,” in *Handbook of Labor Economics*, Vol. 3A, ed. by O. Ashenfelter and D. Card, Elsevier Science, 1865–2097.
- HILLE, A. (2011): “The Gender Gap in Mathematics in French Primary School,” Master’s thesis.
- HIRANO, K. AND G. IMBENS (2001): “Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization,” *Health Services and Outcomes Research Methodology*, 2, 259–278.
- HISARCIKLILAR, M., A. MCKAY, AND P. WRIGHT (2010): “Gender Based Differences in Educational Achievement in Turkey: What Has Changed Over Time?” *Working Paper*, presented at the 30th Annual Conference of the MEEA.
- HUBER, M., M. LECHNER, AND C. WUNSCH (2013): “The performance of estimators based on the propensity score,” *Journal of Econometrics*, 175, 1–21.

COMPLETE BIBLIOGRAPHY

- IMBENS, G. (2004): “Nonparametric Estimation of Average Treatment Effects Under Exogeneity,” *Review of Economics and Statistics*, 86, 4–29.
- IMBENS, G. W. AND J. M. WOOLDRIDGE (2009): “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47, 5–86.
- JACOBS, J. AND J. ECCLES (1992): “The Impact of Mothers’ Gender-Role Stereotypic Beliefs on Mothers’ and Children’s Ability Perceptions,” *Journal of Personality and Social Psychology*, 63, 932–944.
- JAMISON, E., D. JAMISON, AND E. HANUSHEK (2007): “The Effects of Education Quality on Mortality Decline and Income Growth,” *Economics of Education Review*, 26, 772–789.
- KHAN, S. AND E. TAMER (2010): “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” *Econometrica*, 78, 2021–2042.
- KITSING, M. (2008): “PISA 2006–Estonian Results,” Technical Report, Ministry of Education and Research, External Evaluation Department, Tartu, Estonia.
- KRIEG, J. M. AND P. STORER (2006): “How Much Do Students Matter? Applying the Oaxaca Decomposition to Explain Determinants of Adequate Yearly Progress,” *Contemporary Economic Policy*, 24, 563–581.
- KUCIAN, K., T. LOENNEKER, T. DIETRICH, E. MARTIN, AND M. VON ASTER (2005): “Gender Differences in Brain Activation Patterns During Mental Rotation and Number Related Cognitive Tasks,” *Psychology Science*, 47, 112–131.
- LARA-CINISOMO, S., A. PEBLEY, M. VAIANA, E. MAGGIO, M. BERENDS, AND S. LUCAS (2004): “A Matter of Class,” *Rand Review*, 28, 10–5.
- LAVY, V. AND A. SCHLOSSER (2011): “Mechanisms and Impacts of Gender Peer Effects at School,” *American Economic Journal: Applied Economics*, 3, 1–33.
- LECHNER, M. (2010): “A Note on the Common Support Problem in Applied Evaluation Studies,” *Annals of Economics and Statistics*, 91–92, 217–234.
- LUNCEFORD, J. AND M. DAVIDIAN (2004): “Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study,” *Statistics in Medicine*, 23, 2937–2960.
- MARSHALL, N., R. CAYGILL, AND S. MAY (2008): “PISA2006: Reading Liter-

COMPLETE BIBLIOGRAPHY

- acy: How Ready are Our 15-year-olds for Tomorrow's World?" Technical Report, Ministry of Education, Wellington, New Zealand.
- MARTINS, L. AND P. VEIGA (2010): "Do Inequalities in Parents' Education Play an Important Role in PISA Students' Mathematics Achievement Test Score Disparities?" *Economics of Education Review*, 29, 1016–1033.
- MCEWAN, P. AND J. MARSHALL (2004): "Why does academic achievement vary across countries? Evidence from Cuba and Mexico," *Education Economics*, 12, 205–217.
- MCEWAN, P. J. (2004): "The Indigenous Test Score Gap in Bolivia and Chile," *Economic Development and Cultural Change*, 53, 157–190.
- MORA, R. (2008): "A Nonparametric Decomposition of the Mexican American Average Wage Gap," *Journal of Applied Econometrics*, 23, 463–485.
- MULLIGAN, C. (1999): "Galton versus the Human Capital Approach to Inheritance," *Journal of Political Economy*, 107, S184–S224.
- MURNANE, R., J. WILLETT, Y. DUHALDEBORDE, AND J. TYLER (2000): "How Important are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?" *Journal of Policy Analysis and Management*, 19, 547–568.
- NIEDERLE, M. AND L. VESTERLUND (2010): "Explaining the Gender Gap in Math Test Scores: The Role of Competition," *The Journal of Economic Perspectives*, 24, 129–144.
- NONOYAMA-TARUMI, Y. AND J. WILLMS (2010): "The Relative and Absolute Risks of Disadvantaged Family Background and Low Levels of School Resources on Student Literacy," *Economics of Education Review*, 29, 214–224.
- ÑOPO, H. (2008): "Matching as a Tool to Decompose Wage Gaps," *Review of Economics and Statistics*, 90, 290–299.
- OAXACA, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693–709.
- OECD (2007): "PISA 2006 Science Competencies for Tomorrow's World," Technical Report, OECD, Paris, France.

COMPLETE BIBLIOGRAPHY

- (2009): “Education at a Glance 2009,” Technical Report, OECD, Paris, France.
- (2010): “PISA 2009 Results: What Students Know and Can Do. Student Performance in Reading, Mathematics and Science,” Technical Report, OECD, Paris, France.
- PATACCHINI, E. AND Y. ZENOU (2009): “On the Sources of the Black-White Test Score Gap in Europe,” *Economics Letters*, 102, 49–52.
- PLUG, E. AND W. VIJVERBERG (2003): “Schooling, Family Background, and Adoption: Is it Nature or is it Nurture?” *Journal of Political Economy*, 111, 611–641.
- POPE, D. AND J. SYDNOR (2010): “Geographic Variation in the Gender Differences in Test Scores,” *The Journal of Economic Perspectives*, 24, 95–108.
- RADÓ, P. (2001): *Transition in Education*, The Open Society Institute, Institute for Educational Policy, Budapest, Hungary.
- ROBINS, J. M. AND A. ROTNITZKY (1995): “Semiparametric Efficiency in Multivariate Regression Models with Missing Data,” *Journal of the American Statistical Association*, 90, 122–129.
- ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (1994): “Estimation of Regression Coefficients when Some Regressors are not Always Observed,” *Journal of the American Statistical Association*, 89, 846–866.
- ROSENBAUM, P. R. AND D. B. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.
- RUBIN, D. B. (1974): “Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies,” *Journal of Educational Psychology*, 66, 688–701.
- SAKELLARIOU, C. (2008): “Peer Effects and the Indigenous/Non-Indigenous early Test-Score Gap in Peru,” *Education Economics*, 16, 371–390.
- SCHNEEWEIS, N. (2011): “Educational Institutions and the Integration of Migrants,” *Journal of Population Economics*, 24, 1281–1308.
- SCHÜTZ, G., H. W. URSPRUNG, AND L. WÖSSMANN (2008): “Education Policy and Equality of Opportunity,” *Kyklos*, 61, 279–308.
- SEIFERT, B. AND T. GASSER (1996): “Finite-Sample Variance of Local Polynomi-

COMPLETE BIBLIOGRAPHY

- als: Analysis and Solutions,” *Journal of the American Statistical Association*, 91, 267–275.
- (2000): “Data Adaptive Ridging in Local Polynomial Regression,” *Journal of Computational and Graphical Statistics*, 9, 338–360.
- SMITS, J. AND A. HOSGOR (2006): “Effects of Family Background Characteristics on Educational Participation in Turkey,” *International Journal of Educational Development*, 26, 545–560.
- SOHN, K. (2012): “A new Insight into the Gender Gap in Math,” *Bulletin of Economic Research*, 64, 135–155.
- TANSEL, A. (2002): “Determinants of School Attainment of Boys and Girls in Turkey: Individual, Household and Community Factors,” *Economics of Education Review*, 21, 455–470.
- TIEDEMANN, J. (2000): “Parents’ Gender Stereotypes and Teachers’ Beliefs as Predictors of Children’s Concept of their Mathematical Ability in Elementary School,” *Journal of Educational Psychology*, 92, 144–151.
- WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.
- (2007): “Inverse Probability Weighted Estimation for General Missing Data Problems,” *Journal of Econometrics*, 141, 1281–1301.
- WÖSSMANN, L. (2003): “Schooling Resources, Educational Institutions and Student Performance: The International Evidence,” *Oxford Bulletin of Economics and Statistics*, 65, 117–170.
- (2008): “How Equal are Educational Opportunities? Family Background and Student Achievement in Europe and the United States,” *Zeitschrift für Betriebswirtschaft*, 78, 45–70.
- WÖSSMANN, L., E. LÜDEMANN, G. SCHÜTZ, AND M. R. WEST (2009): *School Accountability, Autonomy and Choice around the World*, Elgar, Cheltenham.

Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Thema

**Three Essays on
Semiparametric Econometric Evaluation:
Methods and Applications**

ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Weitere Personen, insbesondere Promotionsberater, waren an der inhaltlich materiellen Erstellung dieser Arbeit nicht beteiligt.¹ Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Konstanz, den 18. April 2013



(Ruben Seiberlich)

¹Siehe hierzu die Eigenabgrenzung zu den einzelnen Kapiteln auf der folgenden Seite.

Eigenabgrenzung

Kapitel 1 entstammt einer gemeinsamen Arbeit mit Frau Alina Botezat (University of Iași). Meine individuelle Leistung bei der Erstellung dieser Arbeit beträgt 50%.

Kapitel 2 entstammt einer gemeinsamen Arbeit mit Jun.-Prof. Zahide Eylem Gevrek-Demiray, PhD (University of Konstanz). Meine individuelle Leistung bei der Erstellung dieser Arbeit beträgt 50%.

Kapitel 3 entstammt einer gemeinsamen Arbeit mit Prof. Dr. Winfried Pohlmeier (University of Konstanz) und Dr. Selver Derya Uysal (IHS Vienna). Meine individuelle Leistung bei der Erstellung dieser Arbeit beträgt 50%.