Essays on Randomized Information Campaigns and Nonparametric Mediation Analysis

THESIS

presented to the Faculty of Economics and Social Sciences
at the University of Fribourg (Switzerland),
in fulfillment of the requirements for the degree of
Doctor of Economics and Social Sciences

by

Anna Solovyeva

from Russia

Accepted by the Faculty of Economics and Social Sciences
on 23.09.2019 at the proposal of

Prof. Dr. Martin Huber (First Advisor) and
Prof. Dr. Michael Gerfin (Second Advisor)

Fribourg, Switzerland 2019

Contents

List of Figures

List of Tables

Introduction

Program and policy evaluation constitute a crucial component of evidence-based policy, now a standard approach to policy-making in many developed and developing countries. Evidence-based policy makes decisions informed by findings from credible research and uses systematic monitoring and evaluation to track implementation and measure outcomes, thus ensuring the continuous improvement of program performance. Therefore, it becomes critically important that research provides rigorous evidence for causal effects of policy interventions. The program evaluation literature is a very dynamic field that has flourished over the past two decades, adding new developments and modifications to the pre-existing set of econometric tools. One influential strand of relatively new theoretical and applied literature is focused on flexible non-parametric methods for estimating treatment effects that are based on the less restrictive functional form and distributional assumptions. The most recent trend revolutionizing the econometric field is machine learning, a new approach that has emerged as a tool to manipulate and analyze massive amounts of data collected by modern computers, that record vast amounts of information about human transactions. Analytic methods based on machine learning are rapidly gaining popularity in the field of applied econometrics.

This PhD thesis, organized as a collection of four independent essays, combines traditional and more innovative methods of program evaluation to identify and estimate causal effects in settings not previously considered in literature. The first two chapters examine randomized information campaigns, combining straightforward impact evaluation, built upon the notion of randomization with flexible adjustments for differences in covariates (in Chapter 1), and machine-learning algorithms, primarily to investigate heterogeneity of treatment effects (Chapter 2). The two subsequent chapters present a novel model of mediation analysis that allows estimating direct and indirect treatment effects when outcomes are only observed for some units (Chapter 3) and compare several mediation/decomposition methods in the estimation of gender-wage gap in the United States (US). In what follows, a non-technical summary of each chapter is laid out.

Chapter 1 written in collaboration with Martin Huber, Ana Kotevska, and Aleksandra Martinovska Stojcheska explores the impact of an information campaign about a rural

development program (RDP) targeting farmers in the former Yugoslav Republic (FYR) of Macedonia. Rural areas of Macedonia are stricken by poverty, inefficient use of agricultural land, undiversified economic activities, and limited access to markets and finance. In an effort to strengthen rural economic growth and increase agricultural competitiveness, the Government of FYR Macedonia introduced a support program comprised of various measures, including financing and training. Despite the availability of governmental means for rural development, the RDP uptake by farmers is low. The present study investigates whether in-person provision of information about RDP increases farmers' awareness and interest in program participation. The information campaign was planned to be randomized within selected villages, such that every other farmer household would receive an information brochure about RDP measures. However, the actual implementation of the campaign deviated from the initial plan, as data collectors did not fully follow the protocol due to low levels of trust from farmers. Instead of delivering the brochure to every second house, field personnel handed them out in public places, collecting contact information of recipients to survey them 1-2 weeks later. As reported by the local staff and reflected in the data, younger smaller scale farmers, without previous experience with RDP, were more likely to receive the brochure. These violations of the experimental design necessitated restriction of the evaluation sample to a specific subset of observations, for which observed background characteristics are well-balanced, and the application of estimation methods that account for the potential remaining differences between the treatment and control group. Towards this end, we invoke the conditional independence assumption and utilize propensity score matching and entropy balancing, in addition to standard OLS, to recover the causal effect of the information brochure on outcomes of interest. Our results suggest that while the intervention succeeded in informing farmers about RDP measures, it had a negative, albeit only marginally significant, effect on the reported possibility of using RDP support in the future. The latter impact is likely driven by an increased awareness of administrative burden associated with RDP participation, which is also reflected in our findings. As revealed by an additional heterogeneity analysis, the negative effect on the possibility of participation appears to be driven largely by a group of unprofitable farmers who are particularly sensitive to additional administrative burden related to RDP, and for whom the requirement of upfront co-financing of RDP projects may be untenable. Our recommendation to Macedonian policy makers is to consider ways of easing administrative hurdles associated with RDP participation.

Similar to Chapter 1, Chapter 2, a collaborative work with Elena Denisova-Schmidt, Martin Huber, and Elvira Leontyeva, considers a randomized information campaign, but in a different setting and applies different econometric tools for additional analysis. In this study, the impact of various information materials on university students' attitudes towards dishonest academic practices and corruption is investigated. Corruption in Russia is a rather understudied but hot topic, especially considering recent anti-corruption protest rallies have been attracting a growing number of young supporters. In our experiment, about 2,000 university student survey participants were randomly assigned to one of four different information treatments (brochures or videos) about the negative consequences of corruption or to a control group. Randomization of the treatment assignment was successful, such that students were on average comparable across the treatment groups. As a methodological advancement over previous research, we use several supervised machine-learning techniques for robustness checks, revealing effect heterogeneities, and for multiple hypothesis testing. The common task of supervised machine-learning methods is to find functions that produce good out-of-sample predictions. This is attained by randomly partitioning data into subsamples. One part of the data ("training data") is first used to find a function that best predicts in-sample; next, another piece of data ("validation data") is used to refine the coefficients obtained in the first step to obtain best prediction in the validation subsample; finally, the remaining data ("test data") are used to obtain out-of-sample predictions using the model built and fine-tuned in the previous two steps. Our analysis suggests that dishonest academic practices are quite common among surveyed students, while corruption is perceived negatively as a "crime" and "evil", yet students are not particularly interested to take part in corruption-awareness activities. No pronounced treatment effects are detected in the total sample. However, when inspecting a subsample of students who frequently plagiarize, we find them to develop stronger negative attitudes towards corruption in the aftermath of our intervention. Unexpectedly, some information materials lead to more tolerant views on corruption among those who plagiarize less frequently and in male students, while female students appear nearly non-responsive to provided information. Based on these findings, we recommend policy makers to scrutinize the possibility of (undesired) heterogeneous effects when designing an anti-corruption educational intervention.

Chapter 3 co-authored with Martin Huber presents a novel model of flexible mediation analysis that identifies and estimates average natural direct and indirect treatment effects in situations when outcome is observed only for some units in the population. The aim of

the mediation analysis is to investigate the "black box" of total treatment effect and to separate the direct effect of a treatment from the indirect component operating through intermediate variables called mediators. This study extends existing nonparametric mediation models, resting on a sequential conditional independence assumption on the assignment of the treatment and the mediator, by allowing the outcome variable to be missing for some observations. Two sets of assumptions about the patterns of missing outcome values are considered: first, that outcome values (after conditioning on observed variables) are missing at random (MAR); and by the second less restrictive set of assumptions, outcomes are not missing at random, i.e. they can be related to unobservable characteristics. In the first case, direct and indirect effects in the total population are obtained through reweighting observations by the inverse of the selection propensity given observed characteristics. For the latter case, identification relies on the use of a control function, a nonparametric analog of the inverse Mill's ratio in Heckman-type selection models; observations are then reweighted by the control function, in addition to the inverse of the selection propensity given the observed characteristics, to identify effects in the selected and total populations. We conduct a brief simulation study investigating finite-sample properties of the presented mediation models based on semiparametric IPW estimation with probit-based propensity scores. Furthermore, we provide an empirical illustration using data from the Program STAR, an educational experiment that randomly assigned kindergarten and primary school pupils to small classes in the United States. We evaluate the average natural direct and indirect effects of the program on standardized math test scores in the first grade of primary school mediated by absenteeism in kindergarten. Due to attrition, the outcome of interest is unobserved for a non-negligible share of the sample. We compare the effects estimated with our newly introduced MAR estimator with those estimated using several other mediation techniques. The MAR estimator for the total population yields the largest estimate of the indirect effect of absenteeism compared to other estimators. Yet, overall, the estimated indirect effects are small compared to the dominating direct effects and are not statistically significant. It appears that other causal mechanisms, unobserved in the data and entering the direct effect, are more important for explaining the positive effect of small kindergarten classes on math test scores.

Finally, Chapter 4, joint work with Martin Huber, investigates the sensitivity of average wage gap decomposition to methods resting on different assumptions regarding endogeneity of observed characteristics, sample selection into employment and estimators' functional form, to gain insight on the robustness of decomposition across identifying

assumptions. Literature on decomposition of wage gaps is concerned with splitting the difference in average wages between two groups into an explained part, attributed to differences in observed characteristics, and a remaining unexplained part linked to various unobserved factors; the latter is often interpreted as discrimination. Since the seminal contributions by Oaxaca (1973) and Blinder (1973) on the linear decomposition method, the field has developed, and new non-parametric approaches have been proposed (see for instance DiNardo et al., 1996; Barsky et al., 2002, among others). However, nearly all these methods fail to consider potential endogeneity that arises due to both (1) observed confounders that are defined prior to birth (e.g., parent's socio-economic status, religious affiliation), and (2) sample selection, as wages are only observed for the working population. There are only a few studies that control for both endogeneity and sample selection; one of them builds on the flexible causal mediation methods presented in Chapter 3 of this thesis. The current study compares the following estimators: Oaxaca-Blinder linear decomposition; semiparametric inverse probability weighting (IPW, see Hirano et al., 2003), which eases linearity but ignores endogeneity and sample selection, just as the Oaxaca-Blinder decomposition; IPW controlling for potential confounders at birth to mitigate endogeneity as in Huber (2015), but ignoring sample selection; and the approaches discussed in Chapter 3 of this thesis to tackle both endogeneity and sample selection. We decompose gender wage gap using data from the US National Longitudinal Survey of Youth 1979. Our findings suggest the wage gap components are not stable across methods. Even the estimate of the total wage gap varies depending on whether we account for sample selection or not. Furthermore, we also compare our preferred extensive specification, that includes not only the levels but also histories of mediator variables, to a more concise specification typically used in previous literature. To no surprise, the explained part of the wage gap decreases, and the unexplained component increases, when fewer mediator variable are included. Given the sensitivity of the wage gap components to methods and variable definitions, we recommend policy makers to be cautious when basing policies on the results of wage decompositions.

Chapter 1

Evaluating an Information Campaign about Rural Development Policies in FYR Macedonia

1.1 Introduction[1]

The agricultural sector plays an important role in the rural economy of the Western Balkans. In this paper, we focus on the former Yugoslav Republic (FYR) of Macedonia, where agriculture, together with forestry and fishing, accounts for about 15 percent of GDP and 17 percent of total employment (State Statistical Office of the Republic of Macedonia, 2015). While the agricultural sector is of importance and has naturally high development potential, it suffers from a problem common to many post-socialist countries – low productivity. To combat negative factors hindering rural growth and to increase agricultural competitiveness, environmental protection, and quality of life in rural areas, the National Strategy for Agriculture and Rural Development was adopted in 2007. The new strategy defines the country's long-term goals aligning Macedonian rural development policy with the common agricultural policy (CAP) of the European Union (EU), in particular with its second pillar, rural development programmes (RDPs) (Dimitrievski et al., 2014). RDPs are seven-year programs comprising various support measures such as financing of planning, training, and advice; annual management payments; and investment aid (European Commission, 2005; Dwyer and Powell, 2016). While EU member states must follow common strategic goals for rural development and agriculture, they adjust the design and implementation of RDPs to their country-specific contexts (Dwyer et al., 2012).

[1]This essay was written in co-authorship with Martin Huber, Ana Kotevska, and Aleksandra Martinovska Stojcheska. It was published as Huber et al. (2018).

Our study examines how a local campaign informing farmers about RDP measures affects their knowledge and interest in taking part in the program in FYR Macedonia. In the course of the campaign, a (randomly) selected group of farmers received a brochure describing the RDP measures and the application process. Based on evidence reported in previous literature (European Commission, 2013; IPARD II, 2015) and informal exchanges with agricultural specialists in FYR Macedonia (Prof. D.Dimitrievski, June 16, 2015, personal communication), we presume that providing information about existing RDP measures in person can increase farmers' awareness about the program and, hence, interest in participating. Dwyer and Powell (2016) emphasize the relevance of information-search cost, among other transaction costs, for RDP performance, pointing to a lack of research on "the costs arising from asymmetries in perception and understanding of programmes" (Dwyer and Powell, 2016, 548). Such asymmetries are possibly present in FYR Macedonia where RDP uptake is low, despite the availability of governmental means for rural development. Our interest lies in determining if providing farmers with information (hence lowering information-search cost and improving the understanding of procedures) affects their intention to participate in the program. According to policy recommendations drafted in Dwyer and Powell (2016), providing support and advice helping beneficiaries prepare and submit applications is crucial for effective use of funding.

Previous studies in development and agricultural economics focus on several aspects of information provision to farmers, including the role of media and extension services in agricultural information access (Hassan et al., 2010; Galadima, 2014), farmers' information needs (Lwoga et al., 2011), and their perceptions of the effectiveness of various information sources (Achuonjei et al., 2003). The majority of these investigations are descriptive and do not aim at estimating the size of information provision effects, while a (nonrandomized) survey is the most commonly employed method. While they collect useful information on farmers' attitudes and behavior, such surveys do not permit a causal interpretation of information provision effects on policy perception and participation. Another issue is limited generalizability, because all cited studies are conducted in developing countries of Africa and Asia, where political and economic background, agricultural practices, rural situations, and information provision might differ substantially from those in transition economies such as FYR Macedonia.

In its research design, our paper is related to a growing body of experimental literature on the effectiveness of randomized information campaigns in various fields of economics, e.g., public economics (Duflo and Saez, 2003; Chetty and Saez, 2013), labor economics (Altmann et al., 2015; Liebman and Luttmer, 2015), and environmental economics

(Ferraro and Miranda, 2013; Benders et al., 2006). Most of these investigations find small to moderate effects of information provision on the outcomes of interest (see, for instance, Chetty and Saez, 2013; Altmann et al., 2015). However, the effectiveness of randomized information campaigns depends ultimately on the field of study, the context, the exact implementation of an intervention, quality and quantity of provided information, and subjects' motivation (Saez, 2009; Feld et al., 2013; Altmann and Traxler, 2014).

This paper contributes to the literature in that it evaluates how information provision affects farmers' intention to participate in the RDP. To the best of our knowledge, no such study has yet been done in the context of transition economies, in the Western Balkans in particular. From a policy perspective, the paper is interesting as it could shed light on how to enhance RDP participation by lowering farmers' information acquisition costs and improving agricultural policy implementation in FYR Macedonia. If information provision does indeed increase farmers' intention to apply for the RDP, this provides policy makers with a relatively inexpensive tool to increase participation rates. Our study also hints at further potential reasons for nonparticipation that appear interesting from a policy perspective, namely: (1) the administrative burden of RDP projects as perceived by farmers, and (2) a specific financing scheme of some RDP measures requiring farmers to provide up to 50 percent of the total investment up front, to be reimbursed upon realized costs.

1.2 Institutional context

1.2.1 Challenges in rural areas of FYR Macedonia

FYR Macedonia is a small, landlocked, transitional economy in the Western Balkans region. The country experienced a sharp economic decline after the breakup of Yugoslavia in 1990 that affected all sectors, including agriculture, the main economic activity in rural areas. A number of socioeconomic issues still persist in rural Macedonia a quarter of century later, presenting a challenge for the successful implementation of rural development policies. These problems include farm fragmentation and small-scale private farming, leading to inefficient use of agricultural land (Dimitrievski et al., 2014), poor diversification of economic activities, insufficient investments in infrastructure, and limited access to markets and sources of finance (Kotevska et al., 2015). On the demographic side, the ongoing trend of out-migration from rural areas has led to a situation where

villages are left with a larger population of older and less-educated residents (European Commission, 2013). Unfavorable education structure, poor qualifications, and insufficient professional skills of the economically active population are considered to be among the factors limiting the potential of rural development (Kotevska et al., 2015). This further deepens the gap between urban and rural standards of living. Today, almost half of the country's poor population resides in rural areas (European Commission, 2013). Thus, the crucial question is of how the government can effectively use policy instruments, including the RDP, to address the problems of rural development and reverse the persistent negative trends.

1.2.2 Agricultural policy and RDP

After its independence from Yugoslavia, FYR Macedonia experienced turbulent agricultural policies with many reforms and ad hoc policy decisions. In 2005, the country received the status of an EU candidate. This new trend of European integration brought about changes in the national agricultural policy which had to be adjusted to the CAP. Therefore, FYR Macedonia focused on harmonization of the national policy for development of agriculture and rural areas. The rural development policy is to a large extent aligned with (the second pillar of) the CAP. It has four priority areas and instruments to support them: (1) increasing the competitiveness of the agricultural and forest holdings, (2) protecting and improving the environment and rural areas, (3) improving the quality of life and encouraging diversification of economic activities in rural areas, and (4) supporting local development (Dimitrievski et al., 2014, 128). In addition, rural development is financed by the EU via the Instrument for Pre-Accession Assistance for Rural Development (IPARD) (Dimitrievski et al., 2014), which is not investigated in this study.

After the 2007 introduction of the National Strategy for Agriculture and Rural Development, the Ministry of Agriculture, Forestry and Water Economy has been preparing and announcing annual programs for rural development. The rural development budget is planned on an annual basis and realized through up to eight calls per year. However, because investments require time to be organized and implemented, and due to limited institutional capacity, budget transfers planned for one year are often conducted only in successive years.

In the period from 2008 to 2014, projects of about EUR 31.4 million were funded under the national program for rural development (see Table 1.1). In the first few years of implementation, the budget was mainly used to increase competitiveness of agricultural holdings, mostly through farm modernization of primary producers. In 2014, a substantial increase in the budget was devoted to the agrifood processing sectors and for improving the quality of life and infrastructural improvement of rural areas. According to information provided by the Agency for Financial Support in Agriculture and Rural Development, in 2014, funds for increasing competitiveness were allocated to 700 applicants (farmers and companies) of relatively small investments averaging EUR 4,460, whereas funds for improving quality of life in rural areas were used by 80 municipalities, averaging EUR 64,470 (APM Database, 2015).

Table 1.1: Annual payments for structural and rural development in FYR Macedonia per priority area (2008-2014, *million EUR*)

	2008	2009	2010	2011	2012	2013	2014	Sum
Increasing the competitiveness of the agricultural and forest holdings:	2.2	1	6.2	0	3.2	0.9	6.2	19.6
Farm modernization	1.5	0.9	5.2	0	2.5	0	3	13.1
Agri-food support (processing, marketing)	0.6	0.1	0.9	-	0.7	0.9	3.2	6.5
Protecting and improving the environment and rural areas	0.4	0	0.2	-	-	-	0.8	1.4
Improving the quality of life and encouraging diversification of econ. activities in rural areas	-	0	0.6	0.1	0.2	1	8.5	10.4
Structural and Rural Development measures (Total)	2.6	1	6.9	0.1	3.4	2	15.5	31.4

Source: Own calculation based on data in the Macedonian APM database (APM Database, 2015).

1.3 Study design

Our study is based on an information campaign experiment conducted in the Southeast of FYR Macedonia in May – June 2015. A brochure was prepared for this purpose in cooperation with the Agency for Financial Support of Agriculture and Rural Development of the Republic of Macedonia. The assessment of the campaign's effectiveness to promote

interest in the RDP is motivated by the relatively low number of applications, despite the government's willingness to support the agricultural sector and the availability of funding.

The causal effect of information provision was intended to be evaluated by means of an experiment. We planned to randomly select 600 farmer households in the largest villages in the chosen region. Every second household on a list of households per village would be treated, while the remaining households would comprise the control group. The treatment probability would thus be asymptotically independent of farmers' characteristics. The treatment group would receive an information brochure on selected RDP measures delivered in person, whereas the control group would receive no such brochure. A survey would be conducted for the entire sample about two weeks later, collecting information on personal and farm characteristics, previous experiences with the RDP application and participation, awareness about the RDP and its potential benefits for the community and the farm, and, importantly, on the farmers' intention to apply for RDP measures and to cofinance RDP projects.

The actual implementation of the campaign deviated from the initial plan. Due to an unstable political situation and generally low levels of trust in the country, data collectors did not manage to fully follow the protocol. Reportedly, farmers were reluctant to communicate with strangers and accept brochures when the surveyors tried to approach the farmers at their homes. Therefore, instead of going to every second house when delivering the brochure, and going house to house to conduct the survey in preselected villages, the surveyors distributed them in several villages in public places, such as local shops, markets, pharmacies, fields, gardens, and water supply stations. They distributed the brochures in person and collected farmers' contact information to survey them $1 - 2$ weeks later. Reportedly, the brochures were more likely to be given to younger farmers, owners of small farms, and those who had not had experience with RDP participation, who were supposedly the types of farmers one predominantly meets in public places in rural areas. The face-to-face survey for the control group took place while the brochures were still being distributed to the treatment group. Once brochure dissemination was completed, the treatment group was surveyed. All treated individuals were interviewed, so there was no unit nonresponse. In the control group, an interviewer would go to the next available household in case of a refusal. The violations of the experimental design required the restriction of the evaluation sample to a specific subset of observations and the application of estimation methods that account for the fact that the intervention was not properly randomized.

The distributed brochure contains information about four selected RDP measures. The face-side of the brochure presents the title and the logo of the Agency for Financial Support in Agriculture and Rural Development, the phrase "Every year the Government of the Republic of Macedonia prepares financial support programs for rural development," and three major goals of the program: modernization and structural adjustment of the agrifood sector, support of economic activities related to nature protection and development of rural areas, and transition of national agricultural policy towards the EU CAP. The rest of the brochure describes selected RDP measures along with eligibility criteria, application processes, required documents, and contact details for the responsible authorities. The selected RDP measures include (1) *Support of young farmers* (Measure 112), (2) *Investments in farm modernization* (Measure 121), (3) *Investments in increasing the economic value of forestry* (Measure 122), and finally (4) *Support of economic associations of farms for joint agricultural activity* (Measure 131). Three of the four listed measures require cofinancing from the farmers' side. Measures 121 and 122 require 50 percent cofinancing by the farmer, whereas measure 131 requires up to 20 percent, depending on the submeasure (Zakon za Zemjodelstvo i Ruralen Razvoj [Law of Agriculture and Rural Development], 2010, 17 – 20). Importantly, the farmer must first personally finance the full amount of investment while actual RDP support is received upon the realized costs, if previously approved to be eligible. Measure 112 represents a grant of up to 600,000 Macedonian denars (EUR 9,760)[2] paid to a successful application in three installments over a three-year period (hence, cofinancing is not required). The brochure targets various groups of farmers and provides the most relevant information regarding RDP measures and the application process. If farmers wanted to obtain more details on the program, the contact information of the responsible authorities could be found on the back of the brochure.

1.4 Data and balancing tests

In our survey, cross-sectional data on 597 farmer households (represented by a household head), including 292 treated and 305 nontreated farmers, were collected. The dataset contains observations from 34 villages of the Southeastern region.

Respondents were asked about their attitudes and opinions about the RDP measured on a Likert scale ranging from 1 "strongly disagree" to 5 "strongly agree." The variables

[2]Based on the year-end 2014 exchange rate (National Bank of the Republic of Macedonia, 2017).

generated from these questions are used as outcomes in our analysis. One group of questions relates to farmers' willingness to apply and participate in the program in the near future (3 – 5 years): "How do you assess the possibility to use RDP support for your household (e.g., for mechanization, equipment purchases) in the next 3 – 5 years?" and "How do you assess your intention to use RDP support for your household in the next 3 – 5 years?" Another group of statements covers awareness and opinions about RDP application process and participation: "I have enough information to independently prepare the application (procedure and documents)," "I have enough knowledge and experience to independently prepare the application (procedure and documents)," "The RDP application (procedure and documents) is easy," and "The RDP increases the administrative work." Information on farmers' previous experiences with RDP was collected, including application for the program in the last three years, use of support in the last three years, and received value of support (in denars).

Background characteristics were also gathered, describing household size; household head's age, sex, educational attainment (primary education, high school, or college/university and higher), and experiences with farming activities, including number of years spent working on a farm, and the primary occupation (whether in agriculture or other industries). Information related to farming activities was available from the survey: farm profitability in the last three years (measured on a scale from 1 to 5: "very unprofitable," "moderately unprofitable," "break-even," "moderately profitable," "very profitable"), ease of getting a loan (1 to 5: "very difficult," "difficult," "medium," "easy," "very easy"), dependence on subsidies to break even financially (1 to 3: "not dependent," "slightly dependent," "very dependent"), frequency of cooperation with other agricultural producers (1 to 5: "never," "rarely," "not sure," "sometimes," "always"), share of agricultural production sold on a market, share of household income from farming, whether or not there are additional workers besides family members working on the farm, total farmed area (in hectares), and total livestock (in heads). Finally, the data contain binary indicators for receiving the brochure, reading it, and learning new facts about RDP measures.

Balancing t-tests comparing the mean values of the characteristics between the treatment and control groups revealed statistically significant (at the 5 percent level) differences in age, education, years in farming, having additional workers on the farm, the share of agricultural production sold on a market, farm profitability, farm capacity (in hectares), and some missing indicators, which points to a failure of randomization. For this reason, we use a restricted sample for our evaluation based on the information

about the brochure assignment process (i.e., brochures were more likely distributed to younger farmers, owners of small farms, and those who had not participated in the RDP previously) provided by the field personnel and reflected in the data. Specifically, we disregard observations from older age groups and only keep prime-age household heads that are up to 55 years old. Furthermore, we only include households that have not previously received RDP support and do not have any employees working on their farm.

As demonstrated in Table 1.2, which provides descriptive statistics and balancing t-tests for the covariates, the subsample is relatively well balanced in terms of mean values of a range of selected characteristics. Apart from primary education, farm profitability and a missing indicator for the share of agricultural production sold on a market, no mean is statistically significantly different across treatment states at the 5-percent level. We consider this subsample in our analysis of the brochure's effect outlined further below.

Table 1.2: Mean covariate values by treatment status in the selected subsample

Variables	Total subsample	Control (C)	Treatment (T)	Difference (T-C)	p-value
Age	44.611	45.703	43.904	-1.799	0.058
	(7.413)	(7.467)	(7.316)	[0.946]	
Male (binary)	0.755	0.723	0.776	0.053	0.345
	(0.431)	(0.450)	(0.419)	[0.056]	
Education: primary (binary)	0.078	0.139	0.038	-0.100	0.009
	(0.268)	(0.347)	(0.193)	[0.038]	
Education: high school (binary)	0.708	0.673	0.731	0.058	0.330
	(0.455)	(0.471)	(0.445)	[0.059]	
Education: college/ university	0.132	0.139	0.128	-0.010	0.812
(binary)	(0.339)	(0.347)	(0.335)	[0.044]	
Education missing (binary)	0.082	0.050	0.103	0.053	0.105
	(0.274)	(0.218)	(0.304)	[0.033]	
Household head's occupation:	0.514	0.535	0.500	-0.035	0.589
agriculture (binary)	(0.501)	(0.501)	(0.502)	[0.064]	
Household head's occupation	0.016	0.020	0.013	-0.007	0.674
missing (binary)	(0.124)	(0.140)	(0.113)	[0.017]	
Years in farming	22.006	22.356	21.779	-0.578	0.611
	(8.517)	(9.485)	(7.851)	[1.133]	
Household size	4.121	4.040	4.173	0.133	0.398
	(1.158)	(1.363)	(1.004)	[0.158]	

Continued on next page

Table 1.2 – continued from previous page

Variables	Total subsample	Control (C)	Treatment (T)	Difference (T-C)	*p*-value
Profitable farm[a]	3.549	3.426	3.628	0.202	0.008
	(0.572)	(0.638)	(0.511)	[0.075]	
Subsidy dependent[b]	2.078	2.168	2.019	-0.149	0.154
	(0.806)	(0.837)	(0.783)	[0.104]	
Subsidy dependent	0.004	0.000	0.006	0.006	0.319
missing (binary)	(0.062)	(0.000)	(0.080)	[0.006]	
Frequency of cooperation[c]	3.700	3.594	3.769	0.175	0.370
	(1.526)	(1.531)	(1.523)	[0.195]	
Frequency of cooperation	0.004	0.010	0.000	-0.010	0.318
missing (binary)	(0.062)	(0.100)	(0.000)	[0.010]	
Share of agricultural production	87.008	87.891	86.436	-1.445	0.488
sold on a market	(16.853)	(15.537)	(17.678)	[2.095]	
Share of agricult. production	0.016	0.000	0.026	0.026	0.045
sold missing (binary)	(0.124)	(0.000)	(0.159)	[0.013]	
Share of income from farming	51.490	53.297	50.321	-2.977	0.312
	(23.166)	(22.725)	(23.445)	[2.938]	
Share of income from farming	0.008	0.000	0.013	0.013	0.157
missing (binary)	(0.088)	(0.000)	(0.113)	[0.009]	
Capacity: farmed area (ha)	1.638	1.695	1.601	-0.094	0.508
	(1.097)	(1.129)	(1.078)	[0.142]	
Capacity: total livestock	1.115	1.184	1.071	-0.113	0.750
(number of heads)	(2.762)	(2.786)	(2.754)	[0.354]	
Number of observations	*257*	*101*	*156*	-	-

Notes: Standard deviations are in parentheses. Robust standard errors are in brackets. [a] *Profitable farm*: 1="very unprofitable"; 2="moderately unprofitable"; 3="break-even"; 4="moderately profitable"; 5="very profitable". [b] *Subsidy dependent*: 1="not dependent"; 2="slightly dependent"; 3="very dependent". [c] *Frequency of cooperation*: 1="never"; 2="rarely"; 3="not sure"; 4="sometimes"; 5="always".

The evaluation sample includes 257 observations, out of which 156 are treated and 101 comprise the control group. As can be seen from Table 1.2, farmers are, on average, about 45 years old, predominantly males, with a high school degree, who have spent almost half of their life working in farming. For half of the farmers, agriculture is the main occupation. They sell most of what they produce on the market, and more than half of their

income comes from farming. Farms in the sample are, on average, moderately profitable or break-even and somewhat dependent on subsidies. Table 1.3 provides additional insight into how our evaluation sample compares to the average farm household in the Southeast region and in the entire FYR Macedonia, in terms of characteristics available from the 2013 Farm Structure Survey. Household heads in the selected sample are typically younger, more educated, more likely to be female, and their household size tends to be larger, compared to the respective averages in the region and country. The average farm size in the sample is comparable to the regional and national averages but smaller in terms of total livestock.

Table 1.3: Mean values of characteristics at sample, regional, and national levels

Variables	Evaluation sample	Southeast region	FYR Macedonia
Average age	44.6	55.5	57.4
Male	75%	88%	89%
Education: no or incomplete primary	-	22%	12%
Education: primary	8%	34%	35%
Education: high school	71%	38%	47%
Education: college/ university	13%	6%	6%
Household size (number of members)	4.1	3.4	3.6
Average farm size (total ha/farm)	1.6	1.5	1.8
Capacity: total livestock (units/farm)	1.1	2.0	2.0
Number of individual farms	257	25,779	170,580

Source: Own 2015 survey and Farm Structure Survey 2013.

Item nonresponse was moderate. In 21 cases (8.2 percent) the educational level was not reported in the selected sample. The number of missing values in other covariates is even smaller. For the purpose of our analysis, we introduce binary indicators for missing values in covariates while replacing actual missing values with zeros.

1.5 Estimation methods

To evaluate the impact of the information brochure on farmers' willingness to apply and participate in the RDP, as well as on other outcome variables, four econometric methods are used: the simple difference in means, OLS, and two non-/semiparametric estimation

techniques, namely, propensity score matching and nonparametric multivariate reweighting (entropy balancing). Formally, we estimate regression specifications of the following kind:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \epsilon_i, \tag{1.1}$$

where the variable Y_i measures various outcomes, e.g., farmers' intention to apply for the RDP, for individual i. T_i is a binary indicator that takes a value of 1 if individual is "treated," i.e., received the information brochure, while β_1 is the coefficient of interest, as it captures the treatment effect. X_i is the vector of covariates used in the OLS regression, propensity score estimation, and entropy balancing.

First, we consider the simple differences in mean outcomes between treatment and control groups. If randomization had been successful, both groups would have been comparable in all their background characteristics (both observed and unobserved), and the differences in mean outcomes across treatment groups would have been unbiased estimates of the average casual effects of the intervention. However, the randomization was not successful, and even after restricting the original sample, some characteristics are not fully balanced across treatment states. For this reason, the simple difference is unlikely to provide an unbiased estimate of the average casual treatment effect.

As an alternative strategy, we control for a range of observed characteristics X_i in the estimation. We rely on the conditional independence assumption (CIA), which states that after conditioning on observed characteristics that jointly affect the treatment probability and the outcome, the independence of the treatment and the potential outcomes hold, such that there are no unobservables jointly affecting the treatment and the outcome (Imbens, 2004):

$$(Y(0), Y(1)) \perp T|X, \tag{1.2}$$

where $Y(0)$ and $Y(1)$ are potential outcomes under, respectively, nontreatment and treatment, T is a binary treatment indicator and X is the covariate set.[3]

The probability of receiving the brochure was reportedly negatively associated with farmers' age, farm capacity, and previous participation in the RDP. This is why it is important to control for these and related characteristics. Our dataset contains information

[3]The observed outcome is then defined as $Y_i = (1 - T_i) \cdot Y_i(0) + T_i \cdot Y_i(1)$, which can be rewritten in the form of equation (1.1). Unconfoundness is equivalent to $\epsilon_i \perp T_i|X_i$ (see Imbens, 2004).

about farmers' age. Farm capacity can be controlled by including variables such as farmed area and total livestock. Farmers who previously participated in the program are excluded from the evaluation sample.

However, we believe it is critical to account for additional characteristics that can be simultaneously related to the outcome variables and the treatment probability, because the brochures were more often distributed to relatively poorly informed farmers. Educational level is likely to affect farmers' awareness about the RDP and, hence, their potential interest in applying for agricultural support. As mentioned in a recent version of The National Programme for Agriculture and Rural Development, Macedonian small-scale farmers appear to have low educational levels (European Commission, 2013, 27). Because the brochure was more often distributed to the owners of smaller farms, it is possible that those who received it had lower educational levels. We also suspect that the relative importance of farming and farm profitability might have affected the probability of receiving the brochure and, at the same time, intention to participate. Individuals for whom farming is the main occupation and whose income is mostly generated by farming should be more interested in obtaining information about the RDP. For this reason, household's head occupation, the share of agricultural production sold on the market, and the share of income from farming are included in the regressions. Furthermore, farm profitability and subsidy dependence should be controlled, because some RDP measures require cofinancing. Given that it is easier for profitable and subsidy-independent farmers to cofinance a project, they might be more interested in learning about RDP measures and obtaining the brochure. Table 1.2 provides supporting evidence for this, because treated farmers are, on average, more likely to have profitable farms and be less subsidy dependent. Finally, we include an indicator for the frequency of cooperation with other farmers as a control variable. More cooperative farmers might be more socially open and active, which increases their chances of receiving the brochure and being interested in the RDP.

Our first approach to control for the observed confounders is a standard OLS regression of the outcome on a constant, the treatment indicator, and the covariates. However, an important drawback of OLS is that it assumes a linear relationship between regressors and the outcome variable, which may be violated in practice. Hence, we also apply more flexible semi- and nonparametric estimators, relying on less rigid functional form assumptions.

One of the most well-known approaches for the evaluation of treatment effects in nonrandomized studies is propensity score matching. The idea is to find for each treated

observation one or more nontreated units with a similar conditional treatment probability, i.e., propensity score. In a general form, the treatment effect ($\hat{\Delta}_{match}$) is defined as the average difference in the outcomes of the treated and the weighted nontreated matched units (see, for instance, Smith and Todd, 2005):

$$\hat{\Delta}_{match} = \frac{1}{N_1} \sum_{\{i:T_i=1\}} (Y_i - \sum_{\{i:T_j=0\}} W_{i,j} Y_j) \tag{1.3}$$

where $W_{i,j}$ is the weight given to the outcome of a nontreated observation j, when j is matched to a treated unit i, and N_1 is the number of treated observations. In this study, we conduct semiparametric kernel matching. First, the propensity score: $p(X) = \Pr(T = 1|X)$ is estimated in a probit regression (see Table A1.1 in Appendix 1 for the propensity score specification). Then, kernel regression of the outcome on the estimated propensity score among the nontreated is conducted to estimate the conditional mean outcome given the propensity score without treatment, $E(Y|T = 0, p(X)) =: m(0, p(X))$ (Huber et al., 2013). Formally,

$$\hat{m}(0, \hat{p}(X_i)) = \frac{\sum_{\{j:T_j=0\}} K(\hat{p}(X_i) - \hat{p}(X_j)/h) Y_j}{\sum_{\{j:T_j=0\}} K(\hat{p}(X_i) - \hat{p}(X_j)/h)}, \tag{1.4}$$

where $\hat{m}(0, \hat{p}(X_i))$ is an estimate of $m(0, \hat{p}(X_i))$, K is a kernel function and h is a bandwidth operator. In the estimations, the Epanechnikov kernel and a bandwidth of 0.6 is used.[4] Thereafter, the treatment effect for the treated is estimated by averaging the estimated function by the empirical distribution of $p(X)$ for the treated:

$$\hat{\Delta}_{kernelmatch} = \frac{1}{N_1} \sum_{\{i:T_i=1\}} (Y_i - \hat{m}(0, \hat{p}(X_i))). \tag{1.5}$$

Matching estimators rely on a common support assumption that ensures units with comparable characteristics exist in both treatment states. Figure 1.1 provides the distribution of the estimated propensity score before and after matching. The upper panel shows some non-overlapping areas in the distribution of the propensity score in the treated and nontreated groups prior to matching. Matching achieves a decent overlap in the propensity score distributions, as illustrated by the lower panel of Figure 1.1. Only four observations in the treatment group lie outside the common support and therefore

[4] These are the default options of the STATA command psmatch2 for the kernel type and bandwidth.

need to be excluded from propensity score matching. Additionally, Table 1.4 presents post-matching mean covariate values by treatment status, standardized differences, and percentage-reduction in standardized differences compared to the original (unmatched) sample, and balancing t-tests on the matched sample. Based on standardized differences and the percentage-reduction in standardized differences, we conclude that matching considerably improved balance in all characteristics (the average reduction in standardized differences was 75 percent), except for farm size.

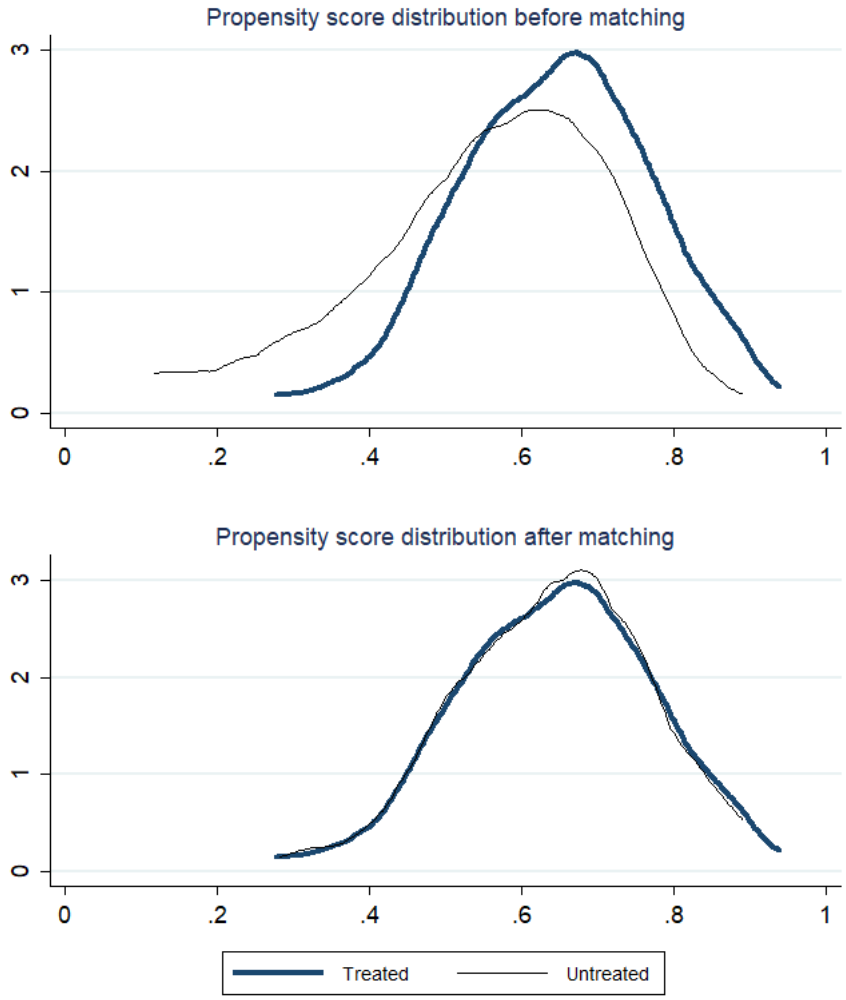

Figure 1.1: Density estimates of the estimated propensity score $\Pr(T = 1|X)$

Notes: The density estimations are based on pstest command in STATA. The bounds of the support of the propensity score are set to be 0 and 1.

Table 1.4: Covariate balance after propensity score matching

Variables	Treatment	Control	Std.diff. %	% reduction	t-value	p-value
Age	44.05	43.97	1.2	95.2	0.10	0.92
Male (binary)	0.78	0.78	0.1	99.5	0.01	0.99
Education: high school (binary)	0.75	0.75	-0.4	96.8	-0.04	0.97
Education: college/ university (binary)	0.12	0.13	-2.6	14.9	-0.23	0.82
Education missing (binary)	0.09	0.08	0.2	98.8	0.02	0.98
Household head's occupation: agriculture (binary)	0.50	0.52	-3.9	43.9	-0.34	0.74
Household head's occupation missing (binary)	0.01	0.01	0.4	92.4	0.04	0.97
Years in farming	21.69	21.38	3.6	46.2	0.32	0.75
Household size	4.16	4.15	1.1	89.7	0.10	0.92
Profitable farm[a]	3.62	3.58	7.2	79.6	0.67	0.50
Subsidy dependent[b]	2.03	2.03	0.2	99.1	0.01	0.99
Frequency of cooperation[c]	3.76	3.72	2.2	81.1	0.19	0.85
Share of agricult. production sold on a market	87.59	87.82	-1.4	84.3	-0.13	0.90
Share of income from farming	50.26	51.34	-4.7	63.9	-0.42	0.67
Capacity: farmed area (ha)	1.60	1.50	9.5	-11.2	0.85	0.40
Capacity: total livestock (number of heads)	1.06	1.11	-1.9	52.6	-0.17	0.86

Notes: "Std.diff.%" stands for standardized difference $\frac{100(\bar{x}_T - \bar{x}_C)}{\sqrt{(var(x_T)+var(x_C))/2}}$ (Rosenbaum and Rubin, 1985). "% reduction" is percentage reduction in the absolute value of standardized difference after matching as compared to before matching. "t-value" and "p-value" refer to two-sample t-tests for equality of means. [a] *Profitable farm*: 1="very unprofitable"; 2="moderately unprofitable"; 3="break-even"; 4="moderately profitable"; 5="very profitable". [b] *Subsidy dependent*: 1="not dependent"; 2="slightly dependent"; 3="very dependent". [c] *Frequency of cooperation*: 1="never"; 2="rarely"; 3="not sure"; 4="sometimes"; 5="always".

The next estimation technique employed in our analysis is entropy balancing,[5] a fully nonparametric multivariate reweighting method proposed by Hainmueller (2012). It does not rest on any propensity score model, but on user-specified initial base weights for nontreated observations. Reweighting is based on computing new weights in a way that

[5]The analysis is run in STATA using package "ebalance" developed by Hainmueller and Xu (2013).

the Kullback-Leibler divergence from the baseline weights is minimized, subject to the balancing constraints. Weights of the nontreated are defined in such a way that exact balance in prespecified covariate moments like the mean is attained for the reweighted nontreated group and the treated. Formally, the weights are chosen by minimizing the following loss function, while balancing the (multidimensional) vector of covariates X_i:

$$\min \sum_{\{i:T_i=0\}} h(\omega_i). \tag{1.6}$$

$$\sum_{\{i:T_i=0\}} \omega_i X_i = \frac{1}{N_1} \sum_{\{i:T_i=1\}} X_i \tag{1.7}$$

and the normalizing constraints

$$\sum_{\{i:T_i=0\}} \omega_i = 1 \tag{1.8}$$

$$\omega_i \geq 0 \; \forall \; i \quad \text{with} \quad T_i = 0, \tag{1.9}$$

where ω_i is a weight estimated for each nontreated observation i, and $h(\cdot)$ is a distance metric. Hainmueller (2012) uses the directed Kullback (1959) entropy divergence: $h(\omega_i) = \omega_i \log(\omega_i/q_i)$, where q_i is the initial base weight. The loss function $\sum_{\{i:T_i=0\}} h(\omega_i)$ measures the distance between the distribution of the estimated weights $\omega_1, \ldots, \omega_{N_0}$ and the initial base weights $q_1, \ldots, q_{N_0}$, where N_0 is the number of nontreated units. The distribution of the base weights is usually set to be uniform with $q_i = 1/N_0$. The constraint 1.7 balances the distribution of X_i between the treatment and the reweighted nontreated groups, so that the latter resembles the former in its covariate distribution. The normalizing constraints 1.8 and 1.9 force the weights to sum up to 1 and be nonnegative. The treatment effect on the treated can be estimated as the difference in mean outcomes between the treatment and the reweighted control groups:

$$\hat{\Delta}_{ebalance} = \frac{1}{N_1} \sum_{\{i:T_i=1\}} (Y_i - \frac{\sum_{\{i:T_i=0\}} Y_i \omega_i}{\sum_{\{i:T_i=0\}} \omega_i}). \tag{1.10}$$

Finally, it needs to be pointed out that our analysis relies implicitly on the stable unit treatment value assumption (SUTVA) that precludes any interaction, spillover, and general

equilibrium effects related to individual treatment assignment. However, it is possible that some study participants in the treated group spread information about the brochure in their villages, which would result in the contamination of the control group. In the event this happened, we estimate the lower bound of the absolute value of the treatment effect.

1.6 Results

This section summarizes our results by presenting the effect estimates for the subsample of farmers up to 55 years of age who do not employ additional workers and have not applied for the RDP in the last three years. The background characteristics are comparably well balanced for this group. A binary indicator for whether farmers have read the brochure or not suggests that only 5.8 percent of those who had received the information brochure did not read it, so that treatment noncompliance is low.

1.6.1 Main results

Table 1.5 presents the effects for the outcomes of interest. Column 2 reports the mean differences in outcomes across the treatment and control groups. The estimates based on OLS, kernel matching, and entropy balancing are provided in columns 3, 4, and 5.

Table 1.5: Treatment effects for the outcomes of interest

Outcome variables	Mean diff.	OLS	Match	ebalance
Panel A: Intention to apply for and use RDP support				
1) Farmer intends to apply for RDP in	0.087	0.005	0.041	0.017
one of the next calls	(0.093)	(0.087)	(0.095)	(0.114)
2) Possibility to use RDP in the next 3-5 years	-0.097	-0.204*	-0.148	-0.250*
	(0.107)	(0.106)	(0.116)	(0.142)
3) Intention to use RDP in the next 3-5 years	-0.071	-0.155	-0.113	-0.166
	(0.098)	(0.095)	(0.098)	(0.116)
Panel B: Judgements on information and application procedures				
4) Farmer has enough information to independently	0.215**	0.194**	0.210**	0.187*
prepare application	(0.091)	(0.092)	(0.099)	(0.106)
5) Farmer has enough knowledge and experience	0.153*	0.142*	0.150*	0.155*
to independently prepare application	(0.084)	(0.076)	(0.089)	(0.091)
6) RDP application (procedure and documents)	0.203**	0.156*	0.128	0.115

Continued on next page

Table 1.5 – continued from previous page

Outcome variables	Mean diff.	OLS	Match	ebalance
is easy	(0.094)	(0.093)	(0.099)	(0.097)
7) RDP increases administrative work	0.195***	0.153***	0.180***	0.177**
for household owners	(0.065)	(0.053)	(0.069)	(0.077)

Notes: Asymptotic standard errors in parentheses are robust to heteroskedasticity for the mean differences, OLS, and entropy balancing. Standard errors are based on 1999 bootstrap replications for the kernel matching estimation. Significance levels: *** 1%, ** 5%, * 10%. Ebalance: means are balanced. Sample sizes: for outcome variables 1-6 is 257 obs., for outcome variable 7 is 256 obs. All the outcome variables (except for *Possibility to use RDP...* and *Intention to use RDP...*) are measured on a five-point scale: 1="strongly disagree"; 2="disagree"; 3="don't know"; 4="agree"; 5="strongly agree". *Possibility to use RDP...* and *Intention to use RDP...* are measured as: 1= "very low", 2= "low", 3= "average", 4= "strong", 5= "very strong".

We find no statistical evidence that the brochure affected the farmers' intended uptake in the near future. For the outcome "Farmer intends to apply for the RDP in one of the next calls," the point estimates are close to zero and nonsignificant. Regarding the "Possibility to use the RDP in the next 3 – 5 years," the OLS and entropy balancing estimates are negative and statistically significant at the 10-percent level. Finally, the effect on the "Intention to use the RDP in the next 3 – 5 years" is not statistically significant.

The treatment effects for the outcome variables presented in Panel B of Table 1.5 might shed some light on why the brochure had mostly insignificant effects on the main outcomes of interest. We notice that the intervention had a positive and statistically significant effect on claiming to have sufficient information, as well as sufficient knowledge and experience to independently prepare the RDP application. Similarly, although with lower statistical significance, we find a positive treatment effect on the assessment of the application procedure as easy. The effect on associating the RDP with increasing administrative work for household owners is positive, relatively strong, and highly statically significant. This could be one reason why the intervention did not boost farmers' intention to use RDP support.

The brochure contained a brief description of bureaucratic procedures related to the application and the selection process. From this, treated farmers could have inferred high administrative costs of being involved in RDP projects. Local experts (namely National Extension Agency advisors) explained that farmers had often believed RDP participation required substantial administrative work, and only those farmers who had no other opportunities to finance their investments would turn to governmental aid. Similarly, a recent study by Dwyer and Powell (2016) reports that potential RDP applicants,

especially in new EU member states, are often discouraged by "what they perceive as costly application, negotiation or management processes" (Dwyer and Powell, 2016, 551). Taken together, this evidence suggests that information in the brochure might have reaffirmed pre-existing beliefs among farmers about the high administrative cost of RDP projects, and thus possibly discouraged their intention to participate.

1.6.2 Heterogeneity of treatment effects by farm profitability

In the next step, we consider the heterogeneity of treatment effects by farm profitability. As mentioned in the "Study design" section, most measures presented in the information brochure require cofinancing. Given that farmers must initially cofinance the project investment from their own means, and RDP support happens only after the costs are realized, it is likely that cofinancing is more feasible for profitable farmers compared to unprofitable ones. Profitable farmers have the opportunity to cofinance an RDP project, either from their own profits and savings or have an easier access to bank loans than unprofitable farmers. Figure A1.1 in Appendix 1 shows that although the majority of farmers in both groups find getting a loan difficult, a greater number of profitable farmers think obtaining a loan is easy compared to unprofitable ones. Thus, we would expect the brochure might have had differential effects by farm profitability.

The heterogeneity analysis is based on the evaluation sample, which contains 106 unprofitable and 151 profitable farms[6]. Table 1.6 presents the effects by farm profitability. Concerning Panel A, for unprofitable farmers, the effects on the reported possibility and intention to use the RDP in the household are negative and statistically significant in several cases, despite the small sample size. For profitable farmers, the impacts are never statistically significant. Turning to Panel B, we find that the brochure increased the profitable farmers' judgment about having enough information, as well as knowledge and experience to independently prepare the application. Both effects are highly significant and relatively strong. At the same time, the brochure had no statically significant effect on these outcome variables for unprofitable farmers. Another finding is that among unprofitable farmers, the intervention (statistically significantly) increased the perception that the RDP brings additional administrative work for the household; at the same time the impact is close to zero among profitable farmers.

[6]Farms that are reported to break even financially are included in the unprofitable group. In the group of unprofitable farmers, 56 received the brochure, and 50 did not; in the group of profitable farmers, 100 were treated, and 51 were not.

Table 1.6: Heterogeneity of treatment effects by farm profitability

Outcome variables	Mean diff.		OLS		Match		ebalance	
	Profit	Unprofit.	Profit	Unprofit.	Profit	Unprofit.	Profit	Unprofit.
Panel A: Intention to apply for and use RDP support								
1) Farmer intends to apply for RDP in	-0.007	0.045	0.116	-0.128	-0.062	-0.121	-0.105	-0.041
one of the next calls	(0.129)	(0.118)	(0.132)	(0.125)	(0.163)	(0.128)	(0.179)	(0.136)
2) Possibility to use RDP in the next 3-5 years	-0.067	-0.215	-0.034	-0.459***	-0.096	-0.667***	-0.137	-0.382*
	(0.141)	(0.163)	(0.143)	(0.167)	(0.158)	(0.237)	(0.208)	(0.202)
3) Intention to use RDP in the next 3-5 years	-0.077	-0.123	-0.041	-0.263*	-0.176	-0.309	-0.124	-0.222
	(0.132)	(0.149)	(0.135)	(0.151)	(0.153)	(0.196)	(0.214)	(0.159)
Panel B: Judgements on information and application procedures								
4) Farmer has enough information to independently	0.387***	0.032	0.400***	0.074	0.406***	-0.162	0.382**	0.010
prepare application	(0.117)	(0.140)	(0.121)	(0.140)	(0.144)	(0.181)	(0.158)	(0.174)
5) Farmer has enough knowledge and experience	0.352***	-0.026	0.293***	-0.019	0.313***	-0.167	0.387***	-0.104
to independently prepare application	(0.093)	(0.140)	(0.093)	(0.123)	(0.115)	(0.183)	(0.086)	(0.167)
6) RDP application (procedure and documents)	0.196	0.243*	0.145	0.153	0.167	0.113	0.329*	0.151
is easy	(0.122)	(0.146)	(0.133)	(0.146)	(0.114)	(0.203)	(0.173)	(0.194)
7) RDP increases administrative work	0.100	0.349***	0.031	0.297***	0.040	0.181	0.101	0.277**
for household owners	(0.065)	(0.115)	(0.058)	(0.095)	(0.102)	(0.131)	(0.097)	(0.129)

Notes: Asymptotic standard errors in parentheses are robust to heteroskedasticity for the mean differences, OLS, and entropy balancing. Standard errors are based on 1,999 bootstrap replications for the kernel matching estimation. Significance levels: *** 1%, ** 5%, * 10%. Ebalance: means are balanced. Sample sizes: for outcome variables 1-6 is 257 obs., for outcome variable 7 is 256 obs. All the outcome variables (except for *Possibility to use RDP...* and *Intention to use RDP...*) are measured on a five-point scale: 1="strongly disagree"; 2="disagree"; 3="don't know"; 4="agree"; 5="strongly agree". *Possibility to use RDP...* and *Intention to use RDP...* are measured as: 1= "very low", 2= "low", 3= "average", 4= "strong", 5= "very strong".

1.7 Conclusion

The present study was designed to determine the effects of a randomized information campaign on farmers' knowledge and intention to participate in the RDP in FYR Macedonia. Based on several reports and prior studies, the hypothesis was that a paucity of comprehensible information contributed to low application rates. We examined if by providing in-person information to farmers, thus lowering farmers' cost of information search, their interest in program participation could be piqued.

The results of our investigation indicate that although the information campaign raised farmers' knowledge about the Macedonian RDP, it did not increase their intention to participate in the program. Instead, it enhanced the perception that the RDP involvement required substantial administrative work from household owners. Furthermore, we found some heterogeneity in the effects by farm profitability. Whereas the information campaign appeared to increase knowledge among profitable farmers, it negatively affected the intention to use RDP support and increased perceived administrative burden among unprofitable farmers.

A caveat of the current study is that the intended randomization of the information brochure could not be properly implemented by the interviewers. We tackled this issue by controlling for observed covariates both in linear regression and nonparametric estimation. Notwithstanding potential limitations, the study's results suggest that the government should consider ways to improve RDP implementation and make it more accessible for Macedonian farmers, possibly by easing the administrative hurdle associated with program participation. Future research could investigate costs and benefits of modifying the financing mode of RDP measures to make them more affordable for break-even and unprofitable farmers.

Chapter 2

Combining Experimental Evidence with Machine Learning to Assess Anti-Corruption Educational Campaigns among Russian University Students

2.1 Introduction[1]

Young people, particularly students, are frequently observed to be the driving forces pushing for reforms that promote justice and fight corruption. The Rose Revolution in Georgia (2003), the Tulip Revolution in Kyrgyzstan (2005), the Arab Spring in Egypt (2011) and the student movements in Taiwan (2014), as well as the protests against corruption in Bulgaria (2013), Ukraine (2014), and Romania (2017), are just a few recent examples of student activism that resulted in social change (Altbach, 2016; Denisova-Schmidt et al., 2015; Klemenčič, 2014). In Russia, where the Putin generation is often viewed as infantile and apolitical (Kasamara and Sorokina, 2017; Volkov, 2017), the recently increased participation of youth in anti-corruption rallies is particularly interesting and controversial.

Corruption[2] has received substantial attention in Russia over the last decade, not only because of its detrimental effects on the national economy and society in general, but also because it became increasingly politicized. The Russian opposition movement has built an agenda around it, attracting a growing number of supporters, among them many high school and university students. Public anti-corruption rallies in March 2017 were

[1] This essay was written in co-authorship with Elena Denisova-Schmidt, Martin Huber, and Elvira Leontyeva. It was released as a SES Working paper, University of Fribourg (Huber et al., 2017). The descriptive findings were published in Solovyeva (2018).

[2] Corruption can be defined as both "the abuse of entrusted power for private gain" (Transparency International) and "the lack of academic integrity"; see recent discussions with examples in Denisova-Schmidt (2017, 2019), Denisova-Schmidt and de Wit (2017).

even described as "angry pupils' walks" in the media (Korostelev et al., 2017). On the other hand, opinion polls suggest that active participants in anti-corruption rallies are not representative of Russian youth. Less than 8% of people ages 18 – 24 have an interest in political issues and discuss them with friends or relatives, while only about 10% are ready to protest (Volkov, 2017). Overall, the stance of the Russian youth towards corruption issues is not clear, as no comprehensive study has yet scrutinized this problem on a grand scale.

This paper (a.) investigates the views of public university students in the Russian region of Khabarovsk on corruption and academic dishonesty during their studies and (b.) examines the effects of an educational campaign exposing students to various informational materials about corruption and its negative consequences. To this end, we surveyed a large sample of about 2,000 students and examined four different anti-corruption materials, namely, two videos produced by Transparency International Russia about the negative consequences of bribery and *reiderstvo* (a hostile corporate takeover) and two brochures, one a general anti-corruption brochure developed by the local authorities and the other a brochure addressing local corruption cases developed for students by the authors.

The results of our study suggest that, while various forms of dishonesty are prevalent among the surveyed students, corruption itself is predominantly viewed as something bad – "crime" and "evil" are the strongest associations expressed in the survey. The perception of corruption at the national level is more negative than at the individual level, which points to the possibility that some respondents have adapted to the situation and might use it for their own benefit. Interest in a roundtable discussion about corruption – a proxy for inclination towards anti-corruption activities – is strikingly low: only 5% of students agreed to join this event. This might be suggestive for young participants in anti-corruption rallies not being representative for the majority of Russian students, which would also be in line with national statistics showing low political activism among the youth (Volkov, 2017).

Concerning the effectiveness of the interventions, we find that although the effects of information exposure were not pronounced in the total sample, there were systematic patterns across subsamples defined by students' inclination to plagiarize when writing papers. One interesting result is that, while (some components of) our intervention promoted awareness of the negative consequences of corruption among students who frequently plagiarize, it led to more tolerant views on the impact of corruption on the Russian education and health systems among students who plagiarize less often. We also consider gender differences in attitudes towards informal academic practices and

corruption. Female students appeared to have stronger negative views on corruption,[3] but to be generally less responsive to interventions and more reluctant to participate in anti-corruption activities than males.

The fact that the interventions affect the participant groups differently has policy implications, as the same information might promote desired attitudes and behavior among some individuals while yielding unwanted results among others. Therefore, policy makers aiming to conduct large-scale anti-corruption campaigns should scrutinize the possibility of effect heterogeneity and target subgroups accordingly. In particular, our study suggests that anti-corruption information campaigns should be focused primarily on individuals who are more likely to be involved in wrongdoing, but not those who are distant from corrupt activities.

Our paper is related to a growing number of corruption studies using lab or field experiments for causal inference (see, for example, discussions in Armantier and Boly, 2011, 2013; Barr and Serra, 2010; Findley et al., 2014; Holmes, 2015; Serra and Wantchekon, 2012). One study that is particularly interesting in our context is that of John et al. (2014), whose findings in an experiment involving US students suggest that awareness about widespread dishonesty increases personal cheating activities while monetary incentives are rather unimportant. Also, Corbacho et al. (2016) find for an information experiment in Costa Rica that individuals who believe that everyone around them is corrupt and/or who have personal experience with corruption are more prone to corruption. Finally, our paper is related to Denisova-Schmidt et al. (2015) and Denisova-Schmidt et al. (2016), which investigate the effectiveness of an anti-corruption folder developed by Transparency International among students in Lviv, Ukraine and Khabarovsk, Russia, respectively. Similar to our comparison of "plagiarist" and "non-plagiarists", Denisova-Schmidt et al. (2015) separately consider students with and without experience in corrupt activities and also find that the intervention might increase tolerance for corrupt behavior. We improve upon these previous studies by considering more and different interventions (both brochures and videos), using a larger sample, and more thoroughly investigating effect heterogeneity. As a methodological advancement compared to other empirical studies in the field, we use machine learning approaches by Belloni et al. (2014), Athey and Imbens (2016), and Ludwig et al. (2017) for conducting robustness checks, finding effect heterogeneities, and conducting mutiple hypothesis testing, respectively.

[3]A large body of empirical literature suggests that women tend to be less corrupt; see Dimant and Tosato (2018); Dollar et al. (2001); Frank et al. (2011); Rivas (2013); Swamy et al. (2001).

The remainder of the paper is organized as follows: Section 2.2 explains the research design and presents the data along with descriptive statistics. Section 2.3 discusses the estimation methods applied in the study. The results are reported in Section 2.4. Section 2.5 concludes.

2.2 Research design and data

Our study is based on a large-scale randomized information campaign conducted among university students in the two cities of the Khabarovsk region – Khabarovsk and Komsomolsk-on-Amur. With populations of about 611,000 and 251,000 people (as of January 1, 2016; Federal State Statistics Service, 2016), respectively, both cities are among the largest urban centres in the Russian Far East. There are twelve universities in Khabarovsk and two in Komsomolsk-on-Amur, with a total of around 68,700 students in the Khabarovsk region in 2015 (Obrazovanie v Rossiiskoi Federatsii, 2014).

The sample of students was drawn from four large public universities in Khabarovsk and two in Komsomolsk-on-Amur, whose total student population accounted for over 70% of all students in the region in 2016 (according to our own calculations based on the online enrolment data from the participating universities). The survey was conducted in November and early December 2016 by a group of students previously instructed by our research team. The following research design was utilized: the interviewers approached students on campuses asking questions about their major, year and education scheme (full- or part-time, on-site or distance education). Only full-time, on-site students with majors in social, technical, and natural sciences or humanities were selected for the study. First-semester bachelor and diploma students were excluded, as they could lack sufficient experience and knowledge about university life. Students in other disciplines, e.g. medicine or theology, were not selected because of their small program sizes. Eligible individuals were asked to take part in a survey about attitudes towards corruption. The questionnaire included a range of questions about the students' motivation to join the university, their academic performance, previous experiences with informal practices[4], family background, and several demographic and socioeconomic characteristics. All the

[4]Here, "informal practices" refers to the practical norms that people often use in order to get things done.

interviews were conducted face-to-face and the interviewers filled out the questionnaire forms in Russian, the native language of all the persons involved.[5]

At one point during the interview, before being asked about their attitudes towards corruption and informal practices, every participant was randomized into one of the four interventions, henceforth also referred to as treatments, or a control group. Each treatment included exposure to one type of information materials about corruption and its negative consequences. The interviewer asked students to play a little game, with the subsequent question depending on the outcome of rolling a fair six-sided (cubical) die. The following assignment rule was applied: if 1 was rolled, the student received an official corruption-awareness brochure (henceforth called the "official brochure"). Rolling a 2 entailed a brochure prepared by our research team on the basis of the materials by Transparency International, a global anti-corruption NGO, and tailored to the student audience (henceforth called the "tailored brochure"). For a 3 or 4, a short video by Transparency International Russia about the negative consequences of bribery or about hostile corporate takeovers ("reiderstvo"), respectively, was shown; 5 and 6 entailed assignment to the control (or non-treated) group. The brochures were professionally printed and the video materials were shown on tablets brought along by the interviewers.

The official brochure was, in our opinion, overwhelming for readers, as it contained too much detailed information, as well as long, redundant definitions, and it was pedantically written and typed in a very small font. It included a portrait of the Russian president Vladimir Putin and his quotation about the fight against corruption, long definitions of corruption and anti-corruption activities, a list of laws and directives against corruption, some corruption-related statistics, examples of anti-corruption measures in the Khabarovsk region, an enumeration of punishments for corruption-related crimes, and a long list of contact information for various responsible authorities.

The tailored brochure was created by our research team with students in mind. We provided succinct and practical information, knowing the experiment participants would not have enough time to absorb less important details. Simple, everyday language was preferred over complex official formulations. The tailored brochure contained a short definition of corruption, a graph describing different types of corruption, some statistics, the negative consequences of bribery (a common corruption type), examples of recent corruption crimes in the Khabarovsk region, and a call for action.

[5]Two sensitive questions about the informal practices exercised by the students in their studies and whether they had encountered bribery at the university were asked on a separate card and filled out by the interviewees themselves.

The videos about the negative consequences of bribery and hostile corporate raiding were part of the "Ten Faces of Corruption" cartoon series developed by Transparency International Russia within the educational project "The Alphabet of a Corruption Fighter". The project targeted high-school and university students and attempted to clarify basic corruption-related concepts. The cartoons only offered video content without audio commentary. The characters were rats depicting the essence of various corrupt behaviors. The video about bribery (Transparency International Russia, 2015a) featured a suicide bomber rat giving a bribe to a security officer when boarding an airplane. The bomb then exploded in the air destroying the plane. The video about *reiderstvo* (Transparency International Russia, 2015b) showed rat police kicking out and arresting the director of a well-functioning cheese factory and overtaking his position.[6]

After the individuals assigned to the treatment groups had familiarized themselves with the respective information materials, the interviewers continued with questions about the informal practices used by students, their moral assessment of corruption, and whether corruption could be eradicated in Russia. At the end of the interview, students were invited to participate in a roundtable discussion taking place on International Anti-Corruption Day[7] (December 9, 2016) at the Pacific National University in Khabarovsk. Finally, respondents were asked whether they would take part in a similar survey next year. Interested students could leave their contact information. All of the post-intervention questions described above were used to construct outcome variables.

Despite the aim to randomize treatment assignment by rolling a die, the distribution of numbers 1 to 6 in the total sample is not perfectly uniform (as would be expected in case of proper randomization), as illustrated in Figure 2.1. In fact, Pearson's chi-squared test clearly rejects the uniform distribution at the 5% level of statistical significance.[8] The probabilities of the brochure treatments (treatments 1 and 2 in Figure 2.1) were higher compared to the video treatments (3 and 4) and the control group (5 and 6).

[6] *Reiderstvo*, or asset-grabbing, is the illicit acquisition of a business or part of a business in Russia; for more, see, for example, Louise Shelley and Judy Deane, `http://reiderstvo.org/`.

[7] The General Assembly of the United Nations introduced Anti-Corruption Day in 2005 in order "to raise awareness of corruption and of the role of the Convention [against Corruption, resolution 58/4] in combating and preventing it" `http://www.un.org/en/events/anticorruptionday/background.shtml`.

[8] The test statistic and the critical value are equal to 21.08 and 9.24, respectively.

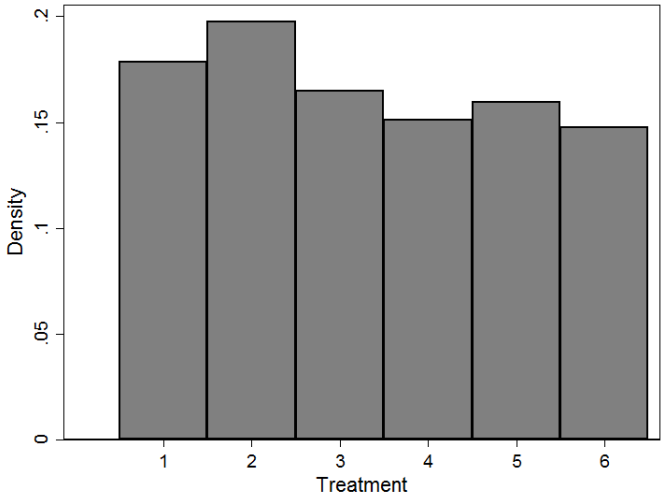

Figure 2.1: Treatment distribution in the total sample

Despite such imbalances in treatment assignment, the average values of the covariates measured in the survey prior to treatment are balanced across the treatment states similarly to a successfully randomized experiment. F-tests conducted for each of the 87 observed covariates revealed hardly any statistically significant (at the 5% level) differences across treatment groups; see Table A2.1 in Appendix 2. One exception was the indicator for having a family with both parents with a p-value of 0.04. For four further covariates – namely, the indicators for a family with no parents, father's occupation: househusband or a retiree, having a Unified State Exam (USE) score of more than 250 (highest quantile), and having a job related to students' education – differences were statistically significant at the 10% level. Given the large number of covariates tested, we are not concerned by these few rejections. Nevertheless, as a robustness check, we ran the main estimations presented in Section 2.4 on the subsample of students surveyed by the interviewers for whom proper treatment randomization (i.e., uniformly distributed numbers 1 to 6) could not be rejected at the 10% level when conducting F-tests separately for each interviewer. Neither covariate balance nor treatment effect estimates in this subsample differed to an important extent from our main results based on the full sample.

Our final sample is comprised of 2,003 individuals, 75% (1,501) of whom study in Khabarovsk and 25% (502) in Komsomolsk-on-Amur. Table 2.1 shows the means and the standard deviations for selected covariates[9] for the 1,741 respondents without any missing

[9]The full list of covariates can be found in Table A2.1 in Appendix 2.

values in these variables. The typical respondent is about 20 years old and just over half of the sample (54%) is female. About one third of the individuals reported to spend on average less than 10,000 rubles (USD 155)[10] a month, while 55% of the respondents have average monthly expenditures between 10,000 and 20,000 rubles (USD 155-310), and 12% spend more than 20,000 rubles. The university education of slightly more than half of the students is state-financed. About 37% of the survey participants study humanities, 31% major in social sciences, 25% are in technical sciences, and 8% specialize in natural sciences.

Concerning previous experiences with wrongdoing and corruption, the self-assessed use of connections is more common than bribery for solving problems. Yet the incidence of additional payments in school prior to tertiary education (e.g., fees for construction, maintenance and school repairs, guarding, etc.) is non-negligible and higher than gift-giving to teachers.[11] Strikingly, about 34% of the participants claimed to have encountered forms of wrongdoing (e.g., bribes, gifts, and help from on-site proctors) during the USE, while 21% encountered some wrongdoing in the university admission process (e.g., cases of admission commissions, instances of preferential admissions). Reportedly, the incidence of bribery at universities after admission appears to be less of an issue. Concerning the use of informal practices by respondents while studying, by far the most popular practice is partial plagiarism when writing papers, followed by crib sheets and copying from others at exams. The least common form of academic dishonesty is asking professors for preferential treatment (e.g., easing requirements, exemption from exams, etc.).

Item non-response is low in our data. In about 4% of the observations, the students' year of birth is missing. Non-response in other demographic, socioeconomic, or individual characteristics is even rarer. About 3% of the students were reluctant to reveal their own informal practices (concerning the question "How often do you use the following practices...?") and whether they encountered bribery at the university. In the estimation part of our analysis, observations with missing values in the covariates are kept in

[10] Based on the average of daily exchange rates from the Russian Central Bank in the period January 1 to November 1, 2016.

[11] Primary and secondary education is predominantly public and tuition-free in Russia. However, informal payments at schools are widespread and range from covering basic maintenance of a school building and the provision of school guarding to some excessive school needs. While voluntary additional school payments have been ruled legal, the fees are often coercive in reality. Also, gift-giving to teachers can be voluntary or forced by parental committees or even the teachers themselves. Our data do not allow the distinguishing between the two types in both the cases of additional school fees and gift-giving to teachers.

Table 2.1: Summary statistics for selected covariates

Variables	Mean	SD
Age	19.99	1.23
Gender: female (binary)	0.54	0.50
Monthly spending: <10k rub (binary)	0.33	0.47
Monthly spending: 10–20k rub (binary)	0.55	0.50
Monthly spending: >20k rub (binary)	0.12	0.33
Education is state financed (binary)	0.53	0.50
Major: humanities (binary)	0.37	0.48
Major: social sciences (binary)	0.31	0.46
Major: technical sciences (binary)	0.25	0.43
Major: natural sciences (binary)	0.08	0.27
Average grade (1=satisfactory...5=excellent)	3.26	1.12
Family or friends solved problems using connections (1=never...5=system.)	2.34	1.04
Family or friends solved problems using bribes (1=never...5=system.)	1.92	0.98
Frequency of giving gifts to teachers at school (1=never...5=system.)	2.80	1.08
Frequency of paying additional fees at school (1=never...5=system.)	3.22	1.20
Encountered (personally/friends/relatives) wrongdoing at USE (binary)	0.34	0.47
Encountered (personally/friends/relatives) wrongdoing at university admission (binary)	0.21	0.41
Encountered bribery at university (1=never...5=system.)	1.55	0.86
How often do you use the following practices? (1=never...5=system.)		
Use crib sheets at exams	2.90	1.17
Submit papers downloaded from the internet	2.25	1.26
Buy papers from friends or specialized firms	1.85	1.15
Write papers plagiarizing some chapters from the internet	3.27	1.20
Copy from other students during exams or tests	2.85	1.17
Deceive professors about study problems	1.95	1.09
Ask professors for preferential treatment	1.63	0.95

the data. Missing values in covariates are replaced with zeros while dummy variables indicating missing observations are generated.

2.3 Methods

Two econometric methods are employed to evaluate the effects of the anti-corruption information materials on the outcomes of interest. Our first strategy is to take differences in mean outcome values between each of the treatment groups and the control group. This yields unbiased estimates of the causal treatment effects if randomization was successful, meaning that any observed and unobserved pre-treatment characteristics are comparable across the treatment groups.

Although the observed pre-treatment characteristics are well balanced in the sample, a few minor differences are still present. As a robustness check, our second strategy aims at controlling for such differences. Specifically, our goal is to control for the confounders of both treatment assignment and outcome of interest in a flexible functional way, potentially allowing interactions as well as higher order terms of confounders to enter both the treatment and outcome equations. To this end, we apply the method of Belloni et al. (2014) to select confounders as well as non-linear functions thereof based on LASSO regression, a machine learning approach permitting variable selection in high dimensional data. More concisely, this so-called post-double-selection method relies on a two-step, LASSO-based variable selection of control variables that are either predictive for the treatment or the outcome (or both). Thereafter, the treatment effects of interest are estimated by an OLS regression of the outcome on the treatment indicators and the selected controls. In our study, we generated higher order terms up to the third order and interaction terms up to the second order for all covariates using the "Generate.Powers" command in the "LARF" package by An and Wan (2016) for the statistical software "R". We added these terms to the list of potential controls for the two-step LASSO procedure and estimated the treatment effects using the "rlassoEffects" command with its default options in the R package "hdm" by Spindler et al. (2016).

Our investigation goes beyond the analysis of treatment effects in the total population and explores the effect heterogeneity of the intervention. We opted for a data-driven rather than ad-hoc approach for finding the most substantial effect heterogeneities in an "honest" way, preventing inferential multiple testing issues related to "snooping" for subgroups with significant effects. This technique builds on modification of a popular method of regression trees, yet another machine learning approach. While the regression tree method partitions the sample in such a way that an outcome is best predicted,[12] the method used in our

[12]That is to say the splitting minimizes the out-of-sample mean squared error.

analysis, the causal tree approach by Athey and Imbens (2016), recursively searchers for sample splits that maximize the mean squared treatment effect.[13]

Specifically, we use the "causalTree" package by Athey et al. (2016) for finding covariates and their values to split our sample on. To this end, we apply the so-called "honest estimation" that uses only one part of a sample ("training data") for subgroup partitioning, while the other part of the sample ("test data") is used to estimate treatment effects within the defined subgroups. This approach, common in machine learning literature, is known as "sample splitting" and used to prevent the aforementioned inference problems.[14] Since our analysis considers more than one treatment and several outcomes, the honest spitting is conducted separately for each treatment-outcome combination, resulting in 124 regression trees. We find the most frequent predictors (and their levels) among those suggested by the recursive partitioning algorithm for the first-level (primary) splitting. At the next step, we generate binary indicators for the most important predictors and use them for splitting the total sample. In the effect-heterogeneity analysis, the average treatment effects are then estimated separately in each of the constructed subsamples.

Our information intervention can potentially affect a number of outcomes of interest rather than just one key outcome. A concern related to estimating the treatment effects for a large number of outcome variables is known as the multiple testing problem, which is an increase in the rate of false "discoveries" of statistically significant effects in multiple simultaneous statistical tests. The issue is that a declared confidence level applies to each test considered individually, and as the number of tests increases, the expected number of incorrect rejections of the null hypothesis also increases (compared to each test considered individually).

We therefore conduct joint hypothesis tests to find whether there are statistically significant treatment effects on groups of outcomes defined by specific questions asked in the survey. To this end, we employ the multiple testing procedure by Ludwig et al. (2017) based on machine learning. The question underlying this test is whether treatment status is predictable from outcomes. Applying sample splitting methods, the test compares the goodness of prediction of treatment status in the original sample with that in a sample where the original treatment status is randomly permuted (i.e., observations are

[13] This is equivalent to finding the largest effect heterogeneities across subgroups.

[14] Our sample is set to be split randomly, such that half of all observations are in the training dataset and the rest are in the testing dataset. To limit the complexity of trees, we apply cross-validation and pruning by specifying a complexity parameter equal the minimum cross validation error that penalizes model complexity; furthermore, the minimum leaf (i.e., subgroup) size is set to 25 observations.

randomly classified as treated or non-treated). If the prediction in the original sample is significantly better than in the permuted one, this is viewed as evidence of a treatment effect on a group of outcomes. We first run the multiple outcome testing for logically grouped outcomes separately for each treatment using the full sample. The multiple significance testing is later repeated in each subgroup defined based on the causal tree procedure.

2.4 Results

We subsequently present the findings, first for the total sample and later on for specific subsamples. Table 2.2 reports the effect estimates based on differences in means in the total sample. Column 2 presents the mean outcomes in the control group. The third column contains the estimated treatment effects of the official corruption-awareness brochure. Columns 4 and 5 give the heteroskedasticity robust standard errors and the p-values, respectively. The estimates for the brochure developed by our team and the videos about the negative consequences of bribery and a hostile corporate raid, i.e. *reiderstvo*, are presented in columns 6 – 8, 9 – 11, and 12 – 14, respectively.

Looking at the control means, we find that informal practices were judged to be quite prevalent among the surveyed students. The use of crib sheets during exams, partial plagiarism from the internet, and copying from other students during exams were thought to occur rather often, as their control means are close to 4 on a scale from 1 (never) to 5 (systematically). What stands out when inspecting the moral assessment of corruption is that "crime" and "evil" were the strongest associations with corruption, whereas defining corruption as a necessity was the least popular option. Interestingly, students perceived corruption's impact on an aggregate level (i.e., its effects on the Russian economy, politics, education and health systems, and police) more negatively, on average, than on a personal level (i.e., on students' career opportunities, quality of life, education, health, and safety). As far as participation in future corruption-awareness activities is concerned, the students expressed very little interest: only 5% agreed to join a roundtable discussion about corruption, and 12% were willing to take part in a next-year survey about corruption.

Inspecting the treatment effects, we find only a handful of them to be statistically significantly different from zero. The official brochure increased the perceived frequency of students copying from others during exams (significant at the 10% level) and the tendency to tolerate informal academic practices in several cases: when a course was considered

"useless", when students worked, and when it was hard to learn the material (the effects are statistically significant at the 1, 10, and 5% levels, respectively). Also, students who received the official brochure were more likely to agree that corruption is a means of income (significant at the 5% level) and to more positively perceive corruption impact on the Russian health system and police (significant at the 5% and 10% levels, respectively).

The tailored brochure was found to strengthen the perceived frequency of students plagiarizing some chapters from the internet when writing papers (significant at the 1% level) and to decrease the tendency to never accept academic cheating (significant at the 10% level). The anti-bribery video slightly increased the reported frequency of students submitting papers downloaded from the internet (significant at the 10% level).

Interestingly, almost all the presented information materials seemed to lower students' interest in the roundtable on corruption and the next survey round, although only the effects of the official brochure and the anti-bribery video were statistically significant (at the 5-10% level).

As a robustness check, we apply the post-double-selection method by Belloni et al. (2014) to control for the covariates and their transformations when estimating treatment effects on individual outcomes. As shown in Table A2.2 in Appendix 2, the effects are very similar in terms of size and significance to the mean difference estimates. Thus, our results are robust to the inclusion of these background characteristics.

Since we consider a large number of individual outcomes, it is important to verify whether treatment effects are jointly significant for groups of outcomes defined by the questions asked in the survey, using the method of Ludwig et al. (2017). The p-values from the tests presented in Table 2.3 indicate that the multiple outcome tests fail to reject the null hypotheses of no treatment effects in the considered outcome groups. In other words, the tests do not provide supporting evidence that there are differences in the outcome distribution between the treatment and control groups. Therefore, we cannot rule out that the few statistically significant effects on individual outcomes are in fact spurious, which implies that the treatments were not effective for the total sample.

Table 2.2: Effects in the total sample

Outcome	Control mean	Official brochure			Tailored brochure			Video: bribery			Video: *reiderstvo*		
		Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.
How often do you think students use the following practices? (1=never... 5=systematically)													
Use crib sheets at exams	3.93	-0.02	0.06	0.77	-0.02	0.06	0.70	0.01	0.06	0.87	-0.02	0.07	0.75
Submit papers downloaded from the internet	3.49	0.02	0.07	0.81	-0.01	0.07	0.93	0.13	0.07	0.08	0.00	0.08	0.96
Buy papers	3.21	0.11	0.07	0.13	0.07	0.07	0.36	0.08	0.08	0.28	0.08	0.08	0.33
Write papers plagiarizing some chapters from the internet	3.75	0.09	0.07	0.17	0.18	0.06	0.00	0.09	0.07	0.20	0.08	0.07	0.23
Copy from other students during exams or tests	3.74	0.12	0.07	0.07	0.09	0.06	0.16	0.11	0.07	0.11	0.06	0.07	0.42
Deceive professors about study problems	3.10	-0.04	0.08	0.62	0.06	0.08	0.47	-0.01	0.08	0.92	0.06	0.08	0.45
Ask professors preferential treatment	2.47	-0.04	0.08	0.57	0.12	0.08	0.13	0.07	0.08	0.39	0.04	0.08	0.61
When do you think these practices are acceptable? (1=definitely no... 5=definitely yes)													
When a course is useless	2.63	0.27	0.09	0.00	0.03	0.08	0.72	-0.01	0.09	0.89	0.03	0.09	0.75
When students work	2.98	0.16	0.08	0.06	0.01	0.08	0.90	0.03	0.09	0.76	-0.08	0.09	0.37
If it is hard to learn material	2.71	0.17	0.08	0.04	-0.01	0.08	0.88	0.04	0.08	0.61	-0.09	0.09	0.31
Always acceptable	2.11	0.09	0.07	0.24	-0.02	0.07	0.80	0.09	0.08	0.23	0.00	0.08	0.99
Never acceptable	3.00	-0.13	0.09	0.16	-0.16	0.09	0.07	-0.09	0.09	0.33	-0.01	0.10	0.92
What does corruption mean to you? (1= definitely no... 5= definitely yes)													
Necessity	1.92	0.10	0.07	0.19	0.03	0.07	0.69	0.04	0.08	0.59	0.03	0.08	0.72
Means of income	2.85	0.19	0.09	0.03	0.05	0.09	0.61	0.13	0.09	0.15	-0.06	0.10	0.50
Crime	4.08	-0.02	0.08	0.84	-0.06	0.07	0.38	0.06	0.08	0.45	-0.01	0.08	0.90
Means to solve problems	3.07	0.08	0.08	0.32	0.05	0.08	0.53	-0.02	0.09	0.82	-0.09	0.09	0.31
Compensation for low salaries	2.59	0.13	0.09	0.14	0.09	0.09	0.28	-0.01	0.09	0.90	-0.03	0.09	0.77
Evil	3.83	0.00	0.08	0.99	0.01	0.08	0.93	-0.01	0.09	0.91	-0.11	0.10	0.25
In your view, how does corruption affect...? (1=strictly negative... 5=fully positive)													
Your career opportunities	2.34	0.03	0.07	0.68	-0.05	0.07	0.47	-0.06	0.08	0.44	0.01	0.08	0.92
Your quality of life	2.39	0.03	0.07	0.68	-0.07	0.07	0.29	-0.04	0.07	0.56	-0.01	0.08	0.90
Your education	2.22	-0.01	0.07	0.94	-0.02	0.07	0.78	-0.09	0.07	0.20	-0.03	0.08	0.68
Your health	2.28	0.02	0.07	0.78	-0.05	0.07	0.46	-0.02	0.07	0.83	-0.10	0.07	0.19
Your safety	2.09	-0.04	0.07	0.53	-0.03	0.07	0.63	-0.09	0.07	0.18	-0.09	0.07	0.20
Russian economy	1.52	0.06	0.05	0.30	-0.01	0.05	0.86	0.01	0.05	0.84	0.07	0.06	0.27
Russian politics	1.58	0.03	0.06	0.56	-0.01	0.05	0.82	-0.01	0.05	0.80	0.03	0.06	0.65
Russian education	1.55	0.08	0.05	0.14	0.02	0.05	0.74	-0.01	0.05	0.82	0.04	0.06	0.50
Russian health system	1.54	0.13	0.06	0.03	0.01	0.05	0.84	0.03	0.06	0.61	0.05	0.06	0.42
Russian police	1.44	0.10	0.06	0.08	-0.01	0.05	0.82	0.01	0.06	0.83	0.06	0.06	0.28
Can corruption be eradicated in Russia? (1=definitely no... 5=definitely yes)	2.52	-0.06	0.07	0.38	-0.10	0.07	0.13	-0.12	0.07	0.10	0.00	0.08	1.00
Take part in roundtable? (0=no, 1=yes)	0.05	-0.02	0.01	0.15	-0.02	0.01	0.21	-0.02	0.01	0.04	0.00	0.01	0.86
Take part in survey next year? (0=no, 1=yes)	0.12	-0.04	0.02	0.08	-0.03	0.02	0.12	-0.04	0.02	0.07	-0.02	0.02	0.29

Notes: 'Effect' represents the difference between the mean outcome value in each treatment group and the control mean, 'se' provides asymptotic standard error robust to heteroskedasticity, and '*p*-v.' stands for *p*-value.

Table 2.3: Multiple outcomes test in full sample

Question group	Official brochure	Tailored brochure	Video: bribery	Video: *reiderstvo*
How often do you think students use the following [corrupt] practices?	0.97	0.17	0.69	0.33
When do you think these [corrupt] practices are acceptable?	0.30	0.77	0.58	0.40
What does corruption mean to you?	0.94	0.55	0.50	0.88
In your view, how does corruption affect aspects of your life?	0.61	0.65	0.69	0.51
In your view, how does corruption affect public spheres in Russia?	0.97	0.76	0.65	0.45
Interest in anti-corruption activities	0.91	0.95	0.65	0.94

Note: The p-values of the joint significance tests are presented.

Heterogeneity of effects

The recursive partitioning algorithm of Athey and Imbens (2016) applied to our data indicates that treatment effects differ most commonly across students who never, seldom, or sometimes wrote papers by plagiarizing some chapters from the internet versus those who did it often or systematically (10 primary-level splits).[15]

Before discussing individual treatment effects in the subsamples, we examine p-values using the joint significance test of Ludwig et al. (2017) for groups of outcomes (based on specific surey questions) in Table 2.4.[16] In the subgroup of 463 students who tended to plagiarize more often, the test finds a jointly statically significant (at the 5% level) effect on the outcomes based on question "In your view, how does corruption affect aspects of your life?" (Table 2.4, Panel A). Among 513 students who never, seldom, or sometimes plagiarize, the tailored brochure had a jointly statistically significant (at the 10% level) effect on the outcomes related to students' interest in anti-corruption activities. The anti-bribery video statistically significantly (at the 5% level) affected the

[15]There are 124 combinations of treatment-outcomes in total but, given our specification of the recursive partitioning, splits are not found for some combinations of treatment-outcomes, which results in 95 primary-level splits.

[16]Note that only the test subsample of 1,002 observations is used for the multiple outcome testing and estimation of individual treatment effects in the subsamples defined by the recursive partitioning algorithm. See Section 2.3 for details.

group of outcomes based on the question "In your view, how does corruption affect public spheres in Russia?" (Table 2.4, Panel B).

Table 2.4: Multiple outcomes test: Subgroups based on plagiarism in studies

Question group	Official brochure	Tailored brochure	Video: bribery	Video: *reiderstvo*
Panel A: Students who often/systematically write papers plagiarizing from the internet				
How often do you think students use the following [corrupt] practices?	0.35	0.82	0.71	0.50
When do you think these [corrupt] practices are acceptable?	0.80	0.97	0.32	0.59
What does corruption mean to you?	0.53	0.66	0.53	0.92
In your view, how does corruption affect aspects of your life?	0.72	0.03	0.69	0.93
In your view, how does corruption affect public spheres in Russia?	0.23	0.62	0.42	0.59
Interest in anti-corruption activities	0.28	0.80	0.40	0.51
Panel B: Students who never/seldom/sometimes write papers plagiarizing from the internet				
How often do you think students use the following [corrupt] practices?	0.18	0.50	0.58	0.14
When do you think these [corrupt] practices are acceptable?	0.30	0.57	0.12	0.12
What does corruption mean to you?	0.44	0.67	0.28	0.59
In your view, how does corruption affect aspects of your life?	0.87	0.42	0.34	0.39
In your view, how does corruption affect public spheres in Russia?	0.32	0.68	0.01	0.29
Interest in anti-corruption activities	0.28	0.06	0.10	0.71

Note: The p-values of the joint significance tests are presented.

Next, Tables 2.5 and 2.6 present the results for individual outcomes in, respectively, the groups of students who frequently wrote papers while plagiarizing some chapters from the internet and those who did so less frequently. Comparing the mean outcome values among non-treated students in both groups, it is striking how those who tended to plagiarize more reported a higher frequency of various informal practices among students, showed more acceptance of dishonesty, and were more skeptic about the possibility of eradicating

Table 2.5: Effects among students who often/systematically write papers plagiarizing some chapters from the internet

Outcome	Control mean	Official brochure			Tailored brochure			Video: bribery			Video: *reiderstvo*		
		Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.
How often do you think students use the following practices? (1=never... 5=systematically)													
Use crib sheets at exams	4.09	-0.21	0.12	0.07	-0.23	0.12	0.05	-0.20	0.12	0.11	-0.06	0.11	0.59
Submit papers downloaded from the internet	3.71	-0.07	0.14	0.60	-0.08	0.13	0.56	-0.04	0.15	0.81	0.04	0.14	0.76
Buy papers	3.37	-0.11	0.15	0.44	-0.27	0.14	0.06	-0.02	0.16	0.91	-0.04	0.15	0.77
Write papers plagiarizing some chapters from the internet	4.14	0.06	0.11	0.61	0.04	0.11	0.74	-0.23	0.14	0.11	0.02	0.12	0.84
Copy from other students during exams or tests	3.98	0.01	0.14	0.93	0.16	0.13	0.23	0.01	0.14	0.93	0.22	0.14	0.11
Deceive professors about study problems	3.23	-0.43	0.16	0.01	-0.13	0.16	0.40	-0.20	0.18	0.24	0.00	0.16	0.98
Ask professors for preferential treatment	2.51	-0.29	0.14	0.04	-0.21	0.17	0.22	-0.23	0.17	0.18	-0.09	0.16	0.57
When do you think these practices are acceptable? (1=definitely no... 5=definitely yes)													
When useless course	2.85	-0.04	0.17	0.83	-0.14	0.17	0.42	-0.08	0.20	0.71	0.02	0.18	0.90
When students work	3.13	0.27	0.17	0.11	-0.16	0.18	0.36	-0.11	0.20	0.60	-0.13	0.18	0.47
If it is hard to learn material	2.80	0.02	0.17	0.90	-0.22	0.17	0.20	-0.02	0.19	0.91	-0.31	0.17	0.07
Always acceptable	2.16	-0.22	0.16	0.17	-0.34	0.15	0.02	-0.11	0.18	0.53	-0.32	0.16	0.05
Never acceptable	2.82	-0.13	0.18	0.50	-0.09	0.18	0.60	0.02	0.21	0.91	-0.06	0.18	0.75
What does corruption mean to you? (1= definitely no... 5= definitely yes)													
Necessity	1.95	0.12	0.17	0.46	-0.14	0.14	0.33	-0.13	0.17	0.42	-0.13	0.15	0.39
Means of income	2.88	0.20	0.19	0.29	0.03	0.19	0.87	-0.21	0.21	0.33	-0.22	0.19	0.25
Crime	4.06	-0.10	0.16	0.52	-0.07	0.14	0.65	-0.01	0.18	0.96	-0.12	0.15	0.41
Means to solve problems	3.21	0.20	0.16	0.22	0.07	0.18	0.69	-0.30	0.20	0.14	-0.14	0.19	0.47
Compensation for low salaries	2.69	0.18	0.19	0.36	-0.13	0.20	0.50	-0.33	0.20	0.10	-0.32	0.20	0.11
Evil	3.75	-0.13	0.18	0.47	0.06	0.16	0.70	0.03	0.18	0.87	-0.32	0.21	0.12
In your view, how does corruption affect...? (1=strictly negative... 5=fully positive)													
Your career opportunities	2.43	-0.29	0.15	0.05	-0.46	0.15	0.00	-0.11	0.18	0.55	-0.25	0.16	0.11
Your quality of life	2.42	-0.08	0.15	0.60	-0.28	0.15	0.06	-0.04	0.18	0.81	-0.08	0.16	0.60
Your education	2.23	-0.28	0.14	0.05	-0.20	0.14	0.17	-0.15	0.16	0.33	-0.16	0.15	0.30
Your health	2.34	-0.06	0.16	0.72	-0.20	0.16	0.20	0.05	0.17	0.75	0.01	0.15	0.95
Your safety	2.09	-0.10	0.15	0.52	-0.14	0.15	0.37	-0.03	0.15	0.82	-0.18	0.14	0.20
Russian economy	1.54	0.07	0.12	0.58	-0.14	0.10	0.17	-0.10	0.12	0.38	-0.09	0.11	0.45
Russian politics	1.58	0.07	0.12	0.56	-0.14	0.11	0.19	-0.12	0.12	0.31	-0.10	0.12	0.39
Russian education	1.56	0.04	0.11	0.70	-0.01	0.11	0.93	-0.02	0.12	0.87	-0.14	0.11	0.20
Russian health system	1.54	0.12	0.12	0.34	-0.03	0.11	0.80	-0.01	0.12	0.93	-0.05	0.11	0.68
Russian police	1.44	0.03	0.11	0.76	-0.02	0.11	0.86	0.02	0.12	0.88	-0.04	0.11	0.73
Can corruption be eradicated in Russia? (1=definitely no... 5=definitely yes)	2.29	0.09	0.13	0.52	-0.20	0.13	0.14	-0.07	0.16	0.66	-0.08	0.15	0.59
Take part in roundtable? (0=no, 1=yes)	0.03	0.00	0.03	0.84	-0.01	0.02	0.55	-0.02	0.02	0.31	0.00	0.03	0.92
Take part in survey next year? (0=no, 1=yes)	0.08	0.01	0.04	0.79	-0.06	0.03	0.06	-0.02	0.04	0.58	-0.02	0.04	0.68

Notes: 'Effect' represents the difference between the mean outcome value in each treatment group and the control mean, 'se' provides asymptotic standard error robust to heteroskedasticity, and '*p*-v.' stands for *p*-value.

Table 2.6: Effects among students who never/seldom/sometimes write papers plagiarizing some chapters from the internet

Outcome	Control mean	Official brochure			Tailored brochure			Video: bribery			Video: *reiderstvo*		
		Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.
How often do you think students use the following practices? (1=never... 5=systematically)													
Use crib sheets at exams	3.76	-0.02	0.14	0.89	0.15	0.12	0.20	0.17	0.14	0.23	0.05	0.15	0.72
Submit papers downloaded from the internet	3.37	0.06	0.15	0.66	0.15	0.14	0.29	0.22	0.15	0.14	0.01	0.16	0.93
Buy papers	3.19	0.02	0.15	0.91	0.14	0.14	0.34	0.14	0.15	0.35	0.08	0.16	0.61
Write papers plagiarizing some chapters from the internet	3.57	0.03	0.14	0.85	0.22	0.13	0.08	0.30	0.13	0.02	0.01	0.15	0.93
Copy from other students during exams or tests	3.72	0.10	0.13	0.44	0.22	0.13	0.08	0.21	0.14	0.14	0.00	0.16	0.98
Deceive professors about study problems	3.08	-0.15	0.17	0.39	0.32	0.16	0.04	0.02	0.17	0.90	0.18	0.17	0.29
Ask professors for preferential treatment	2.50	-0.21	0.15	0.18	0.26	0.15	0.09	-0.14	0.16	0.36	0.16	0.17	0.33
When do you think these practices are acceptable? (1=definitely no... 5=definitely yes)													
When useless course	2.48	0.48	0.17	0.01	0.20	0.15	0.20	0.09	0.16	0.58	0.15	0.18	0.39
When students work	2.82	0.08	0.16	0.61	0.15	0.16	0.33	0.04	0.17	0.80	0.11	0.18	0.54
If it is hard to learn material	2.63	0.09	0.16	0.56	0.07	0.16	0.64	-0.01	0.15	0.96	0.08	0.18	0.65
Always acceptable	2.10	0.31	0.15	0.04	0.10	0.13	0.44	-0.02	0.14	0.90	0.12	0.16	0.46
Never acceptable	2.94	-0.25	0.17	0.15	-0.29	0.17	0.08	-0.40	0.18	0.03	-0.10	0.20	0.61
What does corruption mean to you? (1= definitely no... 5= definitely yes)													
Necessity	1.88	0.13	0.14	0.36	0.20	0.14	0.13	-0.10	0.12	0.43	0.17	0.15	0.27
Means of income	2.76	0.25	0.18	0.15	0.35	0.17	0.04	0.17	0.17	0.32	-0.10	0.18	0.57
Crime	4.01	0.06	0.16	0.68	0.07	0.15	0.64	0.20	0.16	0.21	0.24	0.15	0.12
Means to solve problems	2.95	0.08	0.16	0.60	0.21	0.15	0.17	0.02	0.17	0.92	-0.26	0.18	0.15
Compensation for low salaries	2.48	-0.12	0.16	0.47	0.24	0.15	0.12	-0.04	0.16	0.83	0.00	0.16	0.98
Evil	3.81	0.00	0.17	0.98	-0.10	0.17	0.54	-0.17	0.18	0.32	-0.02	0.18	0.93
In your view, how does corruption affect...? (1=strictly negative... 5=fully positive)													
Your career opportunities	2.29	-0.02	0.14	0.90	-0.01	0.13	0.97	-0.15	0.14	0.28	-0.08	0.14	0.57
Your quality of life	2.33	0.04	0.14	0.76	0.05	0.13	0.70	-0.19	0.14	0.18	-0.03	0.15	0.86
Your education	2.20	0.16	0.14	0.24	0.19	0.14	0.18	-0.16	0.13	0.22	-0.02	0.15	0.89
Your health	2.19	-0.11	0.14	0.42	0.06	0.13	0.64	-0.14	0.14	0.32	-0.14	0.15	0.33
Your safety	2.07	-0.17	0.14	0.22	-0.09	0.13	0.52	-0.31	0.14	0.02	-0.25	0.16	0.11
Russian economy	1.53	0.03	0.09	0.74	0.16	0.09	0.10	0.03	0.09	0.79	0.08	0.12	0.49
Russian politics	1.58	0.07	0.10	0.49	0.10	0.09	0.29	-0.03	0.10	0.76	0.01	0.11	0.89
Russian education	1.59	0.06	0.10	0.53	0.21	0.10	0.04	-0.07	0.10	0.49	0.02	0.11	0.85
Russian health system	1.61	0.12	0.11	0.29	0.23	0.11	0.03	-0.01	0.11	0.95	-0.04	0.12	0.73
Russian police	1.50	-0.01	0.11	0.94	-0.01	0.11	0.91	-0.25	0.10	0.01	-0.11	0.12	0.39
Can corruption be eradicated in Russia? (1=definitely no... 5=definitely yes)	2.55	-0.20	0.14	0.13	-0.23	0.13	0.07	-0.11	0.15	0.45	-0.02	0.15	0.89
Take part in roundtable? (0=no, 1=yes)	0.04	-0.03	0.03	0.32	-0.05	0.02	0.02	-0.03	0.03	0.34	0.00	0.03	0.89
Take part in survey next year? (0=no, 1=yes)	0.11	-0.05	0.04	0.21	-0.08	0.04	0.05	-0.05	0.04	0.23	-0.04	0.05	0.37

Notes: 'Effect' represents the difference between the mean outcome value in each treatment group and the control mean, 'se' provides asymptotic standard error robust to heteroskedasticity, and '*p*-v.' stands for *p*-value.

corruption in Russia.[17] Focusing on general patters of treatment effects, the tailored brochure appears to have impacted students in both groups more than other information materials, yielding a larger number of statistically significant effects. At the same time, the video treatments seemed to be least effective in changing students' attitudes.

Another observation worth mentioning is that students who plagiarized frequently reacted to the provided information materials as expected by our research team. Specifically, both brochures seemed to negatively affect students' perception of the impact of corruption on all the listed aspects of personal life, including career, quality of life, education, health, and safety. Yet, only the first three of these outcomes were affected statistically significantly (at the 1-10% levels). The perceived frequencies of using crib sheets at exams, deceiving professors about study problems, and asking professors for preferential treatment were reduced by the official brochure (significant at the 5-10% levels), whereas the tailored brochure reduced the reported use of crib sheets at exams and buying papers (statistically significant at the 10% level). The tailored brochure tended to lower the acceptance of informal practices in various situations; however, only its negative effect on the "unconditional" acceptance (i.e., "always acceptable") was statistically significant (at the 5% level). The *reiderestvo* video negatively affected the "unconditional" acceptance of information academic practices and in cases when the course material is hard to learn (both effects are statistically significant at the 10% level).

For the group of students who plagiarized less frequently, some information materials tended to raise the reported frequency of informal academic practices and their acceptance, lower the belief that corruption can be eradicated in Russia, and reduce the interest in corruption-awareness activities. Yet, only few of the treatment effects are found to be statistically significant. Considering patterns of statistically significant effects, the tailored brochure increased the reported frequency of informal practices such as plagiarizing chapters from the internet, copying from others at exams, deceiving professors about study problems, and asking them for preferential treatment (statistically significant at the 5-10% levels). Contrary to our expectations, the tailored brochure led to a more positive perception of the impact of corruption on education and health system in Russia (both are statistically significant at the 5% level). Furthermore, the brochure's negative effects on students' interest in the roundtable discussion on corruption and the next round of the survey were also statistically significant (at the 5% level), further lowering students' interest. Lastly, the official brochure increased the unconditional

[17] Most of these differences between the two subgroups are statically significant at the 5% or 10% levels.

acceptance of informal academic practices and when the course was considered "useless" (statistically significant at the 5% level).

Additionally, we consider effect heterogeneities by gender, comparing the groups of 911 male and 1,092 female students. Although not suggested by the recursive partitioning algorithm, this is nevertheless common in the literature (see for example Swamy et al., 2001; and Jetter and Walker, 2015). The p-values from the multiple outcome tests are presented in Appendix 2 Table A2.3. The test of Ludwig et al. (2017) did not point to jointly statistically significant effects for female students (Panel B). Among male students (Panel A), however, the official brochure affected the acceptance of informal academic practices and their opinion about the impact of corruption on public spheres in Russia (significant at the 5% level), and the *reiderstvo* video affected the reported frequency of informal academic practices and the perceived impact of corruption on public sectors in the country (significant at the 10% level).

Tables A2.4 (for males) and A2.5 (for females) report the effects on individual outcome variables. Considering mean outcome values in the respective control groups, females appeared to have stronger negative opinions about the influence of corruption on their lives and to be more reluctant to participate in future corruption-awareness activities than males. Male students exposed to the official brochure demonstrated more tolerance towards informal academic practices, but only the effects on the acceptance of these practices in cases when the course was considered "useless" and when students worked were statistically significantly positive (at the 5-10% levels). The official brochure also led to a more positive perception of corruption on the global level, particularly on the Russian health system and police (statistically significant at the 1 and 10% levels). Furthermore, male students were significantly (at the 10% level) dissuaded by all the treatments (but the *reiderstvo* video) from participation in corruption-awareness activities, whereas the participation propensity of females remained unaffected. Overall, with fewer significant treatment effects, female students were, on average, less responsive to the intervention than males.

2.5 Conclusion

In this paper, we examined the attitudes of Russian students towards dishonest academic practices and corruption and used an experimental design to investigate the effects of an educational campaign consisting of four distinct interventions: two brochures (one officially provided by the local authorities and one particularly tailored to students)

and two videos (about bribery and hostile corporate takeovers) informing students about corruption and its negative consequences. The results suggest that various forms of academic cheating are quite common at Russian universities. At the same time, the attitudes towards corruption are generally negative among the surveyed students. Corruption is believed to have particularly detrimental consequences at the aggregate (national) level, while its effects at the individual level are viewed somewhat less negatively.

Even though the effects of the interventions were not too pronounced in the total sample, we found interesting patterns of impacts, partly going in opposite directions, in subsamples defined along students' plagiarizing behavior and gender. One interesting result is that the interventions promoted awareness of the negative consequences of corruption among students who plagiarize, while they led to more tolerance towards academic dishonesty and more pragmatic attitudes towards corruption among "non-plagiarists". Furthermore, while female students had a more negative opinion about corruption than males, they were generally less responsive to interventions. This demonstrates that information campaigns may affect various groups substantially differently. While the attitudes and behavior of some individuals might be slanted in the desired direction, the very same information can produce detrimental effects by increasing the awareness of corruption among other groups of individuals. Thus, it appears critical for policy makers to reflect on population heterogeneity before conducting large-scale educational campaigns in order to avoid undesired effects.

Comparing the effectiveness across the four interventions, we conclude that the brochures appeared to have more impact on students than the videos. However, only in the subsample of students who frequently plagiarized did the brochures sway students' attitudes towards informal academic practices and corruption in the desired direction. Hence, both the content and the form of information materials appear to matter. When preparing an educational campaign, policy makers should think carefully about tailoring the information materials to the respective target audience in order to maximize effectiveness.

Chapter 3

Direct and Indirect Effects under Sample Selection and Outcome Attrition

3.1 Introduction[1]

Following the seminal papers of Judd and Kenny (1981), Baron and Kenny (1986), and Robins and Greenland (1992), the evaluation of direct and indirect effects, also known as mediation analysis, is widespread in social sciences, see for instance the applications in MacKinnon (2008). The aim is to disentangle the total causal effect of a treatment on an outcome of interest into an indirect component operating through one or several intermediate variables, i.e. mediators, as well as a direct component. As example, consider the effect of educational interventions on health, where part of the effect might be mediated by health behaviors, see Brunello et al. (2016), or personality traits, see Conti et al. (2016). While earlier studies on mediation typically rely on tight linear models, the more recent literature considers more flexible and possibly nonlinear specifications. A large number of contributions assumes a 'sequential conditional independence' assumption, implying that the assignment of the treatment and the mediator is conditionally exogenous given observed covariates and given the treatment and the covariates, respectively. For examples, see Pearl (2001), Robins (2003), Petersen et al. (2006), van der Weele (2009), Flores and Flores-Lagunes (2009), Imai et al. (2010), Hong (2010), Albert and Nelson (2011), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt et al. (2012), Zheng and van der Laan (2012), and Huber (2014a), among many others.

Our main contribution is the extension of such mediation models to account for issues of outcome nonresponse and sample selection, implying that outcomes are only observed for a subset of the initial population or sample of interest. These problems frequently occur in empirical applications as, for instance, wage gap decompositions, where wages

[1]This essay was written in co-authorship with Martin Huber. It was released as a SES Working paper, University of Fribourg (Huber and Solovyeva, 2018a).

are only observed for those who select themselves into employment. In a range of studies evaluating total (rather than direct and indirect effects), sample selection is modelled by a so-called missing at random (MAR) restriction, which assumes conditional exogeneity of sample selection given observed variables, see for instance Rubin (1976), Little and Rubin (1987), Robins et al. (1994), Robins et al. (1995), Carroll et al. (1995), Shah et al. (1997), Fitzgerald et al. (1998), Abowd et al. (2001), and Wooldridge (2002, 2007). In contrast, so-called sample selection or nonignorable nonresponse models permit sample selection to be related to unobservables. Unless strong parametric assumptions are imposed (see for instance Heckman 1976, 1979; Hausman and Wise, 1979; and Little, 1995), identification requires an instrumental variable (IV) for sample selection (e.g. Das et al., 2003; Newey, 2007; and Huber, 2012, 2014b).

In this paper, we combine the identification of average natural direct and indirect effects based on sequential conditional independence with specific MAR or IV assumptions about sample selection. We show under which conditions the parameters of interest in the total as well as the selected population (whose outcomes are actually observed) are identified by inverse probability weighting[2] (IPW) based on particular propensity scores for treatment and selection. Under MAR, effects in the total population are obtained through reweighting by the inverse of the selection propensity given observed characteristics. If selection is related to unobservables, we make use of a control function that can be regarded as a nonparametric version of the inverse Mill's ratio in Heckman-type selection models. Under specific conditions, reweighting observations by the inverse of the selection propensity given observed characteristics and the control function identifies the effects in the selected and the total population. To convey the intuition of our identification results, we provide a brief simulation study, in which the finite sample properties of semiparametric IPW estimation with probit-based propensity scores is investigated.

As an empirical illustration, we evaluate the average natural direct and indirect effects of Program STAR, an educational experiment in Tennessee, U.S., which randomly assigned children to small classes in kindergarten and primary school. The positive impact of STAR classes on academic achievement has been demonstrated, for example, in Krueger (1999), but less is known about the underlying causal mechanisms. We consider absenteeism in kindergarten as potential mediator of the overall effect. The outcome of interest is the score on a standardized math test in the first grade of primary school, which is unobserved for a non-negligible share of students in the data due to attrition. We apply one of our

[2]The idea of using inverse probability weighting to control for selection problems goes back to Horvitz and Thompson (1952).

proposed IPW-based estimators to account for outcome attrition and compare the results to several alternative mediation estimators that make no corrections for sample selection.

The remainder of this paper is organized as follows. Section 3.2 discusses the parameters of interest, the assumptions, and the nonparametric identification results based on inverse probability weighting. Section 3.3 outlines estimation based on the sample analogs of the identification results. Section 3.4 presents a simulation study. Section 3.5 provides an application to Project STAR data. Section 3.6 concludes.

3.2 Identification

3.2.1 Parameters of interest

We would like to disentangle the average treatment effect (ATE) of a binary treatment variable D on an outcome variable Y into a direct effect and an indirect effect operating through the mediator M, which has bounded support and may be a scalar or a vector and discrete and/or continuous. To define the effects of interest, we use the potential outcome framework, see Rubin (1974), which has been applied in the context of mediation analysis by Rubin (2004), Ten Have et al. (2007), and Albert (2008), among others. $M(d), Y(d, M(d'))$ denote the potential mediator state as a function of the treatment and potential outcome as a function of the treatment and the potential mediator, respectively, under treatments $d, d' \in \{0, 1\}$. Only one potential outcome and mediator state, respectively, is observed for each unit, because the realized mediator and outcome values are $M = D \cdot M(1) + (1 - D) \cdot M(0)$ and $Y = D \cdot Y(1, M(1)) + (1 - D) \cdot Y(0, M(0))$.

The ATE is given by $\Delta = E[Y(1, M(1)) - Y(0, M(0))]$. To disentangle the latter, note that the (average) natural direct effect (using the denomination of Pearl, 2001)[3] is identified by exogenously varying the treatment but keeping the mediator fixed at its potential value for $D = d$:

$$\theta(d) \quad = \quad E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}, \tag{3.1}$$

Equivalently, by exogenously shifting the mediator to its potential values under treatment and non-treatment but keeping the treatment fixed at $D = d$, the (average) natural indirect

[3]Robins and Greenland (1992) and Robins (2003) refer to this parameter as the total or pure direct effect and Flores and Flores-Lagunes (2009) as net average treatment effect.

effect[4] is obtained:

$$\delta(d) \quad = \quad E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}. \tag{3.2}$$

The ATE is the sum of the direct and indirect effects defined upon opposite treatment states:

$$\begin{aligned}
\Delta &= E[Y(1, M(1)) - Y(0, M(0))] \\
&= E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0) \\
&= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1).
\end{aligned} \tag{3.3}$$

This follows from adding and subtracting $E[Y(0, M(1))]$ or $E[Y(1, M(0))]$, respectively. The notation $\theta(1), \theta(0)$ and $\delta(1), \delta(0)$ points to possible effect heterogeneity w.r.t. the potential treatment state, implying the presence of interaction effects between the treatment and the mediator. However, the effects cannot be identified without further assumptions, as either $Y(1, M(1))$ or $Y(0, M(0))$ is observed for any unit, whereas $Y(1, M(0))$ and $Y(0, M(1))$ are never observed.

In contrast to natural effects, which are functions of the potential mediators, the so-called controlled direct effect is obtained by setting the mediator to a predetermined value m, rather than $M(d)$:

$$\gamma(m) \quad = \quad E[Y(1, m) - Y(0, m)], \quad m \text{ in the support of } M. \tag{3.4}$$

Whether $\theta(d)$ or $\gamma(m)$ is of primary interest depends on the research question at hand. The controlled direct effect may provide policy guidance whenever mediators can be externally prescribed, as for instance in a sequence of active labor market programs assigned by a caseworker, where D and M denotes assignment of the first and second program, respectively. This allows analysing the direct effect of the first program under alternative combinations of program prescriptions. In contrast, the natural direct effect assesses the effectiveness of the first program given the status quo decision to participate in the second program in the light of participation or non-participation in the first program. We refer to Pearl (2001) for further discussion of what he calls the descriptive and prescriptive natures of natural and controlled effects.

[4]Robins and Greenland (1992) and Robins (2003) refer to this parameter as the total or pure indirect effect and Flores and Flores-Lagunes (2009) as mechanism average treatment effect.

Our identification results will make use of a vector of observed covariates, denoted by X, that may confound the causal relations between D and M, D and Y, and M and Y. A further complication in our evaluation framework is that Y is assumed to be observed for a subpopulation, i.e. conditional on $S = 1$, where S is a binary variable indicating whether Y is observed/selected, or not. We therefore also define the direct and indirect effects among the selected population:

$$\theta_{S=1}(d) = E[Y(1, M(d)) - Y(0, M(d))|S = 1],$$
$$\delta_{S=1}(d) = E[Y(d, M(1)) - Y(d, M(0))|S = 1],$$
$$\gamma_{S=1}(m) = E[Y(1, m) - Y(0, m)|S = 1].$$

Empirical examples with partially observed outcomes include wage regressions, with S being an employment indicator, see for instance Gronau (1974), or the evaluation of the effects of policy interventions in education on test scores, with S being participation in the test, see Angrist et al. (2006). Throughout our discussion, S is allowed to be a function of D, M, and X, i.e. $S = S(D, M, X)$. However, S must neither be affected by nor affect Y.[5] S is therefore not a mediator, as selection per se does not causally influence the outcome. An example for such a set up in terms of nonparametric structural models is given by

$$Y = \phi(D, M, X, U), \quad S = \psi(D, M, X, V), \tag{3.5}$$

where U, V are unobserved characteristics and ϕ, ψ are general functions.[6]

3.2.2 Assumptions and identification results under MAR

This section presents identifying assumptions that formalize the sequential conditional independence of D and M as imposed by Imai et al. (2010) and many others as well as a MAR restriction on Y that implies that S is related to observables.[7]

[5] See for instance Imai (2009) for an alternative set of restrictions, assuming that selection is related to the outcome but is independent of the treatment conditional on the outcome and other observable variables.

[6] Note that $Y(d, M(d')) = \phi(d, M(d'), X, U)$, which means that fixing the treatment and the potential mediator yields the potential outcome.

[7] We also implicitly impose the Stable Unit Treatment Value Assumption (SUTVA, see Rubin, 1990) stating that the potential mediators and outcomes for any individual are stable in the sense that their values do not depend on the treatment allocations in the rest of the population.

Assumption 1 (conditional independence of the treatment):
(a) $Y(d,m) \perp D|X=x$, (b) $M(d') \perp D|X=x$ for all $d,d' \in \{0,1\}$ and m in the support of M.

By Assumption 1, there are no unobservables jointly affecting the treatment, on the one hand, and the mediator and/or the outcome, on the other hand, conditional on X. In observational studies, the plausibility of this assumption crucially hinges on the richness of the data, while in experiments, it is satisfied if the treatment is randomized within strata defined by X or randomized independently of X.[8]

Assumption 2 (conditional independence of the mediator):
$Y(d,m) \perp M|D=d',X=x$ for all $d,d' \in \{0,1\}$ and m,x in the support of M,X.

By Assumption 2, there are no unobservables jointly affecting the mediator and the outcome conditional on D and X. Assumption 2 only appears realistic if detailed information on possible confounders of the mediator-outcome relation is available in the data (even in experiments with random treatment assignment) and if post-treatment confounders of M and Y can be plausibly ruled out when controlling for D and X.[9]

Assumption 3 (conditional independence of selection):
$Y \perp S|D=d,M=m,X=x$ for all $d \in \{0,1\}$ and m,x in the support of M,X.

By Assumption 3, there are no unobservables jointly affecting selection and the outcome conditional on D,M,X, such that outcomes are are missing at random (MAR) in the denomination of Rubin (1976). Put differently, selection is assumed to be selective w.r.t. observed characteristics only.

Assumption 4 (common support):
(a) $\Pr(D=d|M=m,X=x)>0$ and (b) $\Pr(S=1|D=d,M=m,X=x)>0$ for all $d \in \{0,1\}$ and m,x in the support of M,X.

Assumption 4(a) is a common support restriction requiring that the conditional probability to be treated given M,X, henceforth referred to as propensity score, is larger than zero in either treatment state. It follows that $\Pr(D=d|X=x)>0$ must hold, too. By Bayes' theorem, Assumption 4(a) implies that $\Pr(M=m|D=d,X=x)>0$, or in the case of M being continuous, that the conditional density of M given D,X is larger than zero.

[8] In the latter case, even the stronger condition $\{Y(d',m),M(d),X\} \perp D$ holds.

[9] Several studies in the mediation literature discuss identification in the presence of post-treatment confounders of the mediator that may themselves be affected by the treatment. See for instance Robins and Richardson (2010), Albert and Nelson (2011), Tchetgen Tchetgen and VanderWeele (2014), Imai and Yamamoto (2011), and Huber (2014a).

Conditional on X, M must not be deterministic in D, as otherwise identification fails due to the lack of comparable units in terms of the mediator across treatment states. Assumption 4(b) requires that for any combination of D, M, X, the probability to be observed is larger than zero. Otherwise, the outcome is not observed for some specific combinations of these variables implying yet another common support issue.

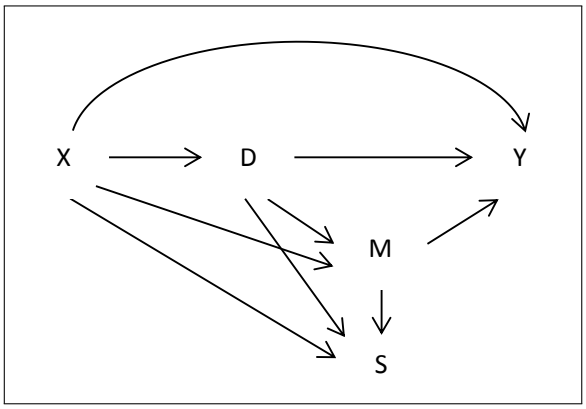

Figure 3.1: Causal framework under MAR

Figure 3.1 illustrates the causal framework underlying our assumptions by means of a causal graph, see for instance Pearl (1995), in which each arrow represents a potential causal effect. Further (unobserved) variables that only affect one of the variables explicitly displayed in the system are kept implicit. For instance, there may be unobservable variables U that affect the outcome, but do not influence D, M, or S; otherwise, there would be confounding.

Under Assumptions 1 to 4, potential outcomes as well as direct and indirect effects in the total population are identified based on weighting by the inverse of the treatment and selection propensity scores.

Theorem 1:

(i) Under Assumptions 1, 2, 3, and 4, for $d \in \{0, 1\}$,

$$\begin{aligned} E[Y(d, M(1-d))] &= E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|M,X) \cdot \Pr(S=1|D,M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right], \\ E[Y(d, M(d))] &= E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|X) \cdot \Pr(S=1|D,M,X)}\right]. \end{aligned} \tag{3.6}$$

(ii) Under Assumptions 1(a), 2, 3, and 4, and M following a discrete distribution,

$$E[Y(d,m)] = E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\} \cdot S}{\Pr(D=d|X) \cdot \Pr(M=m|D,X) \cdot \Pr(S=1|D,M,X)}\right]. \tag{3.7}$$

Proof: See Appendix 3.0.1

Using the results of Theorem 1, it can be shown that the direct and indirect effects are identified by

$$\begin{aligned}
\theta(d) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X) \cdot S}{\Pr(D=d|X) \cdot \Pr(S=1|D,M,X)}\right], \\
\delta(d) &= E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|M,X) \cdot \Pr(S=1|D,M,X)} \cdot \left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right], \\
\gamma(m) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|X)}\right) \cdot \frac{I\{M=m\} \cdot S}{\Pr(M=m|D,X) \cdot \Pr(S=1|D,M,X)}\right].
\end{aligned}$$

These expressions are related to the IPW-based identification in Huber (2014a) for the case with no missing outcomes with the difference that here, multiplication by $S/\Pr(S=1|D,M,X)$ is included to account for sample selection. Furthermore, our results fit into the general framework of Wooldridge (2002), who considers the IPW-based M-estimation of missing data models. Finally, for the identification of $\gamma(m)$, Assumption 1 can be relaxed to Assumption 1(a) because (in contrast to $\theta(d), \delta(d)$) the distribution of the potential mediator $M(d)$ need not be identified.

3.2.3 Assumptions and identification results under selection related to unobservables

In the following discussion, we consider the case that selection is related to both observables and unobservables that are associated with the outcome. Assumptions 3 and 4 are therefore replaced. Rather, we assume that an instrumental variable for S is available to tackle sample selection.

Assumption 5 (Instrument for selection):

(a) There exists an instrument Z that may be a function of D, M, i.e. $Z = Z(D,M)$, is conditionally correlated with S, i.e. $E[Z \cdot S|D,M,X] \neq 0$, and satisfies (i) $Y(d,m,z) = Y(d,m)$ and (ii) $\{Y(d,m), M(d')\} \perp Z(d'',m')|X=x$ for all $d,d',d'' \in \{0,1\}$ and z,m,m',x in the support of Z,M,X,

(b) $S = I\{V \leq \Pi(D,M,X,Z)\}$, where Π is a general function and V is a scalar (index

of) unobservable(s) with a strictly monotonic cumulative distribution function conditional on X,
(c) $V \perp (D, M, Z)|X$.

Assumption 5 no longer imposes the independence of Y and S given observed characteristics. As the unobservable V in the selection equation is allowed to be associated with unobservables affecting the outcome, Assumptions 1 and 2 generally do not hold conditional on $S = 1$ due to the endogeneity of the post-treatment variable S. In fact, $S = 1$ implies that $\Pi(D, M, X, Z) > V$ such that conditional on X, the distribution of V generally differs across values of D, M. This entails a violation of the sequential conditional independence assumptions on D, M given $S = 1$ if potential outcome distributions differ across values of V. We therefore require an instrumental variable denoted by Z, which is allowed to be affected by D and M, but must not affect Y or be associated with unobservables affecting M or Y conditional on X, as invoked in (5a).[10] We apply a control function approach based on this instrument,[11] which requires further assumptions.

By the threshold crossing model postulated in 5(b), $\Pr(S = 1|D, M, X, Z) = \Pr(V \leq \Pi(D, M, X, Z)) = F_V(\Pi(D, M, X, Z))$, where $F_V(v)$ denotes the cumulative distribution function of V evaluated at v. We will henceforth use the notation $p(W) = \Pr(S = 1|D, M, X, Z)$ with $W = D, M, X, Z$ for the sake of brevity. Again by Assumption 5(b), the selection probability $p(W)$ increases strictly monotonically in Π conditional on X, such that there is a one-to-one correspondence between the distribution function F_V and specific values v given X. For X fixed, the identification of F_V by $p(W)$ is 'as good as good as' identifying V. By Assumption 5(c), V is independent of (D, M, Z) given X, implying that the distribution function of V given X is (nonparametrically) identified. Figure 3.2 illustrates the causal framework underlying Assumptions 1, 2, and 5 by means of a causal graph.

By comparing individuals with the same $p(W)$, we control for F_V and thus for the confounding associations of V with (i) D and $\{Y(d, m), M(d')\}$ and (ii) M and $Y(d, m)$ that occur conditional on $S = 1$. In other words, $p(W)$ serves as control function where

[10] As an alternative set of IV restrictions in the context of selection, d'Haultfoeuille (2010) permits the instrument to be associated with the outcome, but assumes conditional independence of the instrument and selection given the outcome.

[11] Control function approaches have been applied in semi- and nonparametric sample selection models, e.g., Ahn and Powell (1993), Das et al. (2003), Newey (2007), and Huber (2012, 2014b) as well as in nonparametric instrumental variable models, see, for example, Newey et al. (1999), Blundell and Powell (2004), and Imbens and Newey (2009).

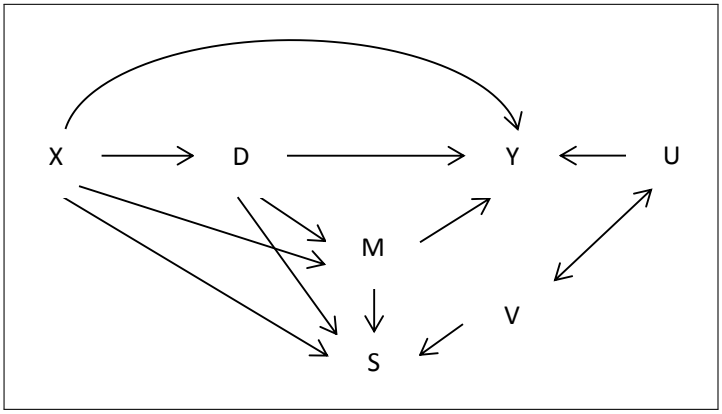

Figure 3.2: Causal framework under selection on unobservables

the exogenous variation comes from Z. More concisely, it follows from our assumptions for any bounded function g that

$$E\left[g(Y(d,m))|D,M,X,p(W),S=1\right]=E\left[g(Y(d,m))|D,M,X,F_V,S=1\right]$$
$$=E\left[g(Y(d,m))|D,X,F_V,S=1\right]=E\left[g(Y(d,m))|X,F_V,S=1\right].$$

The first equality follows from $p(W)=F_V$ under Assumption 5, the second from the fact that when controlling for F_V, conditioning on $S=1$ does not result in an association between $Y(d,m)$ and M given D,X such that $Y(d,m)\perp M|D,X,p(W),S=1$ holds by Assumptions 2 and 5. The third equality follows from the fact that when controlling for F_V, conditioning on $S=1$ does not result in an association between $Y(d,m)$ and D given X such that $Y(d,m)\perp D|X,p(W),S=1$ holds by Assumptions 1 and 5. Similarly,

$$E\left[g(M(d))|D,X,p(W),S=1\right]=E\left[g(M(d))|D,X,F_V,S=1\right]=E\left[g(M(d))|X,F_V,S=1\right]$$

follows from the fact that when controlling for F_V, conditioning on $S=1$ does not result in an association between $M(d)$ and D given X such that $M(d)\perp D|X,p(W),S=1$ holds by Assumptions 1 and 5. These results will be useful in the proofs of Theorems 2 and 3 (see Appendix 3.0.2).

Furthermore, identification requires the following common support assumption, which is similar to Assumption 4(a), but in contrast to the latter also includes $p(W)$ as a conditioning variable.

Assumption 6 (common support):
$\Pr(D=d|M=m,X=x,p(W)=p(w),S=1)>0$ for all $d\in\{0,1\}$ and m,x,z in the

support of M, X, Z.

By Bayes' theorem, Assumption 6 implies that the conditional density of $p(W) = p(w)$ given $D, M, X, S = 1$ is larger than zero. This means that in fully nonparametric contexts, the instrument Z must in general be continuous and strong enough to importantly shift the selection probability $p(W)$ conditional on D, M, X in the selected population. Assumptions 1, 2, 5, and 6 are sufficient for the identification of mean potential outcomes as well as direct and indirect effects in the selected population.

Theorem 2:
(i) Under Assumptions 1, 2, 5, and 6 for $d \in \{0, 1\}$,

$$\begin{aligned} E[Y(d, M(1-d))|S=1] &= E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X,p(W))} \cdot \frac{\Pr(D=1-d|M,X,p(W))}{\Pr(D=1-d|X,p(W))}\middle| S=1\right], \\ E[Y(d, M(d))|S=1] &= E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|X,p(W))}\middle| S=1\right]. \end{aligned} \tag{3.8}$$

(ii) Under Assumptions 1(a), 2, 5, and 6, and M following a discrete distribution,

$$E[Y(d,m)|S=1] = E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\}}{\Pr(D=d|X,p(W)) \cdot \Pr(M=m|D,X,p(W))}\middle| S=1\right] \tag{3.9}$$

Proof: See Appendix 3.0.2.

Therefore, the direct and indirect effects are identified by

$$\begin{aligned} \theta_{S=1}(d) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X,p(W))} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X,p(W))}\right) \cdot \frac{\Pr(D=d|M,X,p(W))}{\Pr(D=d|X,p(W))}\middle| S=1\right], \\ \delta_{S=1}(d) &= E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X,p(W))} \cdot \left(\frac{\Pr(D=1|M,X,p(W))}{\Pr(D=1|X,p(W))} - \frac{1-\Pr(D=1|M,X,p(W))}{1-\Pr(D=1|X,p(W))}\right)\middle| S=1\right], \\ \gamma_{S=1}(m) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|X,p(W))} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|X,p(W))}\right) \cdot \frac{I\{M=m\}}{\Pr(M=m|D,X,p(W))}\middle| S=1\right]. \end{aligned}$$

In nonparametric models that allow for general forms of effect heterogeneity related to unobservables, direct and indirect effects can generally only be identified among the selected population. The reason is that effects among selected observations cannot be extrapolated to the non-selected population if the effects of D and M interact with unobservables that are distributed differently across $S = 1, 0$. The identification of effects in the total population therefore requires additional assumptions. In Assumption 7 below, we impose homogeneity in the direct and indirect effects across selected and non-selected populations conditional on X, V. A sufficient condition for effect homogeneity is the separability of observed and unobserved components in the outcome variable, i.e. $Y = \eta(D, M, X) + \nu(U)$, where η, ν are general functions and U is a scalar or vector of

unobservables. Furthermore, common support as postulated in Assumption 6 needs to be strengthened to hold in the entire population. In addition, the selection probability $p(w)$ must be larger than zero for any w in the support of W; otherwise, outcomes are not observed for some values of D, M, X. Assumption 8 formalizes these common support restrictions.

Assumption 7 (conditional effect homogeneity):
$E[Y(1,m) - Y(0,m)|X = x, V = v, S = 1] = E[Y(1,m) - Y(0,m)|X = x, V = v]$ and $E[Y(d, M(1)) - Y(d, M(0))|X = x, V = v, S = 1] = E[Y(d, M(1)) - Y(d, M(0))|X = x, V = v]$, for all $d \in \{0,1\}$ and m, x, v in the support of M, X, V.

Assumption 8 (common support):
(a) $\Pr(D = d|M = m, X = x, p(W) = p(w)) > 0$ and (b) $p(w) > 0$ for all $d \in \{0,1\}$ and m, x, z in the support of M, X, Z.

While the mean potential outcomes in the total population remain unknown even under Assumptions 7 and 8, the effects of interest are nevertheless identified by the separability of U.

Theorem 3:
(i) Under Assumptions 1, 2, 5, 6, 7, and 8 for $d \in \{0,1\}$,

$$\theta(d) = E\left[\left(\frac{Y \cdot D}{\Pr(D = 1|M, X, p(W))} - \frac{Y \cdot (1 - D)}{1 - \Pr(D = 1|M, X, p(W))}\right) \cdot \frac{\Pr(D = d|M, X, p(W)) \cdot S}{\Pr(D = d|X, p(W)) \cdot p(W)}\right] \tag{3.10}$$

$$\delta(d) = E\left[\frac{Y \cdot I\{D = d\} \cdot S}{\Pr(D = d|M, X, p(W)) \cdot p(W)} \cdot \left(\frac{\Pr(D = 1|M, X, p(W))}{\Pr(D = 1|X, p(W))} - \frac{1 - \Pr(D = 1|M, X, p(W))}{1 - \Pr(D = 1|X, p(W))}\right)\right].$$

(ii) Under Assumptions 1(a), 2, 5, 6, 7, and 8, and M following a discrete distribution,

$$\gamma(m) = E\left[\left(\frac{Y \cdot D}{\Pr(D = 1|X, p(W))} - \frac{Y \cdot (1 - D)}{1 - \Pr(D = 1|X, p(W))}\right) \cdot \frac{I\{M = m\} \cdot S}{\Pr(M = m|D, X, p(W)) \cdot p(W)}\right]. \tag{3.11}$$

Proof: See Appendix 3.0.3.

3.2.4 Extensions to further populations, parameters, and variable distributions

This section briefly sketches how the identification results can be extended to further populations of interest, policy-relevant parameters, and richer distributions of the treatment and/or the mediator. First and in analogy to the concept of weighted treatment effects in Hirano et al. (2003), direct and indirect effects can be identified for particular target populations by reweighting observations according to the distribution of X in the target population. To this end, we define $\omega(X)$ to be a well-behaved weighting function depending on X. Including $\frac{\omega(X)}{E[\omega(X)]}$ in the expectation operators presented in the theorems above yields the parameters of interest for the target population. As an important example, consider $\omega(X) = \Pr(D = 1|X)$. For some well-behaved function $f(Y, D, M, S, X, Z)$ of the observed data,

$$\begin{aligned} &E\left[\tfrac{\omega(X)}{E[\omega(X)]} \cdot f(Y, D, M, S, X, Z)\right] = E\left[\tfrac{\Pr(D=1|X)}{\Pr(D=1)} \cdot f(Y, D, M, S, X, Z)\right] \\ &= E\left[\tfrac{\Pr(D=1|X)}{\Pr(D=1)} \cdot f(Y, D, M, S, X, Z)\right] = E\left[f(Y, D, M, S, X, Z)|D = 1\right], \end{aligned} \tag{3.12}$$

i.e. the expected value of that function among the treated is identified. Likewise, defining $\omega(X) = 1 - \Pr(D = 1|X)$ gives the expected value among the non-treated. Any of the expressions in the expectation operators of the theorems may serve as $f(Y, D, M, S, X, Z)$ in (3.12).[12]

Second, the identification results may be extended to well-behaved functions of Y, rather than Y itself. For instance, replacing Y by $I\{Y \leq a\}$, the indicator function that Y is not larger than some value a, everywhere in the theorems permits identifying distributional features or effects. The inversion of potential outcome distribution functions allows identifying quantile treatment effects.

[12] For instance, the weighted versions of the parameters identified in Theorem 1 correspond to

$$\begin{aligned} E_\omega[Y(d, M(1-d))] &= E\left[\frac{\omega(X)}{E[\omega(X)]} \cdot \frac{Y \cdot I\{D = d\} \cdot S}{\Pr(D = d|M, X) \cdot \Pr(S = 1|D, M, X)} \cdot \frac{\Pr(D = 1 - d|M, X)}{\Pr(D = 1 - d|X)}\right], \\ E_\omega[Y(d, M(d))] &= E\left[\frac{\omega(X)}{E[\omega(X)]} \cdot \frac{Y \cdot I\{D = d\} \cdot S}{\Pr(D = d|X) \cdot \Pr(S = 1|D, M, X)}\right], \\ E_\omega[Y(d, m)] &= E\left[\frac{\omega(X)}{E[\omega(X)]} \cdot \frac{Y \cdot I\{D = d\} \cdot I\{M = m\} \cdot S}{\Pr(D = d|X) \cdot \Pr(M = m|D, X) \cdot \Pr(S = 1|D, M, X)}\right]. \end{aligned}$$

.

Third, our framework can be adapted to allow for multiple or multivalued (rather than binary) treatments. If D is multivalued discrete, the derived expressions may be applied under minor adjustments. For instance, for any $d \neq d'$ in the discrete support of D, the expression for potential outcomes in Theorem 1 becomes

$$E[Y(d, M(d'))] = E\left[\frac{Y \cdot I\{D = d\} \cdot S}{\Pr(D = d|M, X) \cdot \Pr(S = 1|D, M, X)} \cdot \frac{\Pr(D = d'|M, X)}{\Pr(D = d'|X)}\right]$$

under appropriate common support conditions. If D is continuous, any indicator functions for treatment values, which are only appropriate in the presence of mass points, need to be replaced by kernel functions, while treatment propensity scores need to be substituted by conditional density functions. In analogy to Hsu et al. (2018b), who consider mediation analysis with continuous treatments in the absence of sample selection, the expression for potential outcomes in Theorem 1 becomes

$$\begin{aligned} E[Y(d, M(d'))] &= \lim_{h \to 0} E\left[\frac{Y \cdot \omega(D; d, h) \cdot S}{E[\omega(D; d, h)|M, X] \cdot \Pr(S = 1|D, M, X)}\right. \\ &\times \left.\frac{E[\omega(D; d', h)|M, X]}{E[\omega(D; d', h)|X]}\right]. \end{aligned}$$

The weighting function $\omega(D; d) = K\left((D - d)/h\right)/h$, with K being a symmetric second order kernel function assigning more weight to observations closer to d and h being a bandwidth operator. For h going to zero, i.e. $\lim_{h \to 0}$, $E[\omega(D; d', h)|X]$ and $E[\omega(D; d', h)|M, X]$ correspond to the conditional densities of D given X and given M, X, respectively, also known as generalized propensity scores. We refer to Hsu et al. (2018b) for more discussion on direct and indirect effects of continuous treatments and how estimation may proceed based on generalized propensity scores. We also note that in the context of controlled direct effects, such kernel methods not only allow for a continuous treatment, but (contrarily to our theorems) also for a continuous mediator.

3.3 Estimation

The parameters of interest can be estimated using the normalized versions of the sample analogs of the IPW-based identification results in Section 4.2. This implies that the weights of the observations used for the computation of mean potential outcomes add up to unity, as advocated in Imbens (2004) and Busso et al. (2009). For instance, the normalized sample

analogs of the results in Theorem 1, part (i) are given by

$$\hat{\mu}_{1,M(0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot D_i \cdot S_i}{\hat{p}(M_i, X_i) \cdot \hat{\pi}(D_i, M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{D_i \cdot S_i}{\hat{p}(M_i, X_i) \cdot \hat{\pi}(D_i, M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)},$$

$$\hat{\mu}_{0,M(1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot (1-D_i) \cdot S_i}{(1-\hat{p}(M_i, X_i)) \cdot \hat{\pi}(D_i, M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i) \cdot S_i}{(1-\hat{p}(M_i, X_i)) \cdot \hat{\pi}(D_i, M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)},$$

$$\hat{\mu}_{1,M(1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot D_i \cdot S_i}{\hat{p}(X_i) \cdot \hat{\pi}(D_i, M_i, X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{D_i \cdot S_i}{\hat{p}(X_i) \cdot \hat{\pi}(D_i, M_i, X_i)},$$

$$\hat{\mu}_{0,M(0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \cdot (1-D_i) \cdot S_i}{(1-\hat{p}(X_i)) \cdot \hat{\pi}(D_i, M_i, X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i) \cdot S_i}{(1-\hat{p}(X_i)) \cdot \hat{\pi}(D_i, M_i, X_i)}.$$

i indexes observations in an i.i.d. sample of size n and $\hat{\mu}_{d,M(d')}$ is an estimate of $\mu_{d,M(d')} = E[Y(d, M(d'))]$ with $d, d' \in \{1, 0\}$. $\hat{p}(M_i, X_i)$, $\hat{p}(X_i)$ are estimates of the treatment propensity scores $\Pr(D = 1|M_i, X_i)$, $\Pr(D = 1|X_i)$, respectively, while $\hat{\pi}(D_i, M_i, X_i)$ is an estimate of the selection propensity score $\Pr(S = 1|D, M, X)$. Direct and indirect effect estimates are obtained by $\hat{\theta}(d) = \hat{\mu}_{1,M(d)} - \hat{\mu}_{0,M(d)}$ and $\hat{\delta}(d) = \hat{\mu}_{d,M(1)} - \hat{\mu}_{d,M(0)}$.

When propensity scores are estimated parametrically, e.g. based on probit models as in the simulations and application below, then $\hat{\mu}_{d,M(d')}, \hat{\theta}(d), \hat{\delta}(d)$ satisfy the sequential GMM framework discussed in Newey (1984), with propensity score estimation representing the first step and parameter estimation the second step. This approach is $\sqrt{n}$-consistent and asymptotically normal under standard regularity conditions. When the propensity scores are estimated nonparametrically, $\sqrt{n}$-consistency and asymptotic normality can be obtained if the first step estimators satisfy particular regularity conditions. See Hsu et al. (2018a), who consider series logit estimation of the propensity scores, however, for the case without sample selection. Furthermore, the bootstrap is consistent for inference as the proposed IPW estimators are smooth and asymptotically normal.

3.4 Simulation study

This section provides a brief simulation study, in which we investigate the finite sample properties of estimation of natural direct and indirect effects based on the sample analogs

of Theorems 1 to 3. To this end, the following data generating process is considered:

$$
\begin{aligned}
&Y = 0.5D + M + 0.5DM + X - \alpha DU + U, \quad Y \text{ is observed if } S = 1,\\
&S = I\{0.5D - 0.5M + 0.25X + Z + V > 0\},\\
&M = 0.5D + 0.5X + W, \quad D = I\{0.5X + Q > 0\},\\
&Z = 0.25X - 0.25M + R,\\
&X, U, V, W, Q, R \sim \mathcal{N}(0,1), \text{ independently of each other.}
\end{aligned}
$$

The outcome Y is a linear function of the observed variables D, M, X and an unobserved term U, and is only observed if the selection indicator S – which depends on D, M, X, an instrument Z, and an unobservable V – is equal to one. α gauges the interaction of D and U in the outcome equation. For $\alpha \neq 0$, the treatment effect is heterogeneous in U such that Assumption 7 is violated. W and R denote the unobservables in the linearly modelled mediator M and instrument Z, respectively. Any unobservable as well as the observed covariate X are standard normally distributed independent of each other. In this framework, the assumptions underlying Theorem 1 are satisfied.

We run 5,000 Monte Carlo simulations with sample sizes $n = 1000, 4000$ and consider estimation of the natural direct and indirect effects in the total population ($\theta(d)$, $\delta(d)$) based on three different estimators: (i) normalized IPW as suggested in Huber (2014a) among the selected ('IPW w. $S = 1$') that controls for X but ignores selection bias, (ii) normalized IPW based on Theorem 1 assuming MAR ('IPW MAR'), and (iii) normalized IPW based on Theorem 3 ('IPW IV'). We estimate the treatment and selection propensity scores by probit and apply a trimming rule that discards observations with $\hat{p}(M, X)$ smaller than 0.05 or larger than 0.95 or with $\hat{\pi}(D, M, X)$ smaller than 0.05 to prevent exploding weights due to small denominators. Trimming hardly affects IPW estimator (i), but reduces the variance of estimation based on Theorems 1 and 3 in several cases.

Table 3.1 reports the simulations results under $\alpha = 0.25$,[13] namely the bias, standard deviation (std), and the root mean squared error (RMSE) of the various estimators for the natural direct and indirect effects in the total population. Ignoring selection (IPW w. $S = 1$) yields biased estimates of the direct effects under either sample size, while biases are generally small for estimation based on Theorem 1. Interestingly, the latter result also holds for estimation related to Theorem 3, where the selection process accounts for the same observed factors as under the correct MAR assumption, plus the control function.

[13] Results are very similar when $\alpha = 0$ and therefore omitted.

Table 3.1: Simulations under selection on observables, total population

	$\hat{\theta}(1)$			$\hat{\theta}(0)$			$\hat{\delta}(1)$			$\hat{\delta}(0)$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse	bias	std	rmse
	$\alpha = 0.25$, $n = 1000$											
IPW w. $S = 1$	-0.16	0.14	0.21	-0.17	0.16	0.23	-0.01	0.15	0.15	-0.02	0.11	0.12
IPW MAR	0.03	0.28	0.28	0.01	0.20	0.20	-0.03	0.13	0.14	-0.05	0.14	0.15
IPW IV	-0.01	0.30	0.30	-0.02	0.31	0.31	-0.02	0.18	0.18	-0.03	0.15	0.15
	$\alpha = 0.25$, $n = 4000$											
IPW w. $S = 1$	-0.16	0.07	0.18	-0.17	0.08	0.19	0.00	0.08	0.08	-0.01	0.06	0.06
IPW MAR	0.01	0.15	0.15	0.01	0.10	0.10	-0.02	0.07	0.07	-0.03	0.08	0.09
IPW IV	-0.01	0.15	0.15	-0.02	0.16	0.16	-0.01	0.09	0.09	-0.02	0.08	0.08

Note: 'std' and 'rmse' report the standard deviation and root mean squared error, respectively.

Even though including the control function is not required for consistency, it does not jeopardize identification either, even if Assumption 7 requiring $\alpha = 0$ is not satisfied,[14] as reflected in the low biases. However, accounting for this unnecessary variable entails an increase of the standard deviation in some cases. In general, the estimators based on Theorems 1 and 3 are (due to the estimation of the sample selection propensity score) less precise than IPW without selection correction in the selected sample. The proposed methods become relatively more competitive in terms of the RMSE as the sample size increases and gains in bias reduction become relatively more important compared to losses in precision.

As a modification to our initial setup, we introduce a correlation between U and V, which implies that the assumptions underlying Theorem 1 no longer hold, while those of Theorem 2 are satisfied and those of Theorem 3 are satisfied when $\alpha = 0$:

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \text{ where } \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

Table 3.2 reports the results for the estimation of natural effects in the total population under $\alpha = 0$ and 0.25 using the same methods as before. Non-negligible biases occur not only when ignoring sample selection ('IPW w. $S = 1$'), but also when selection is assumed to be related to observables only (IPW MAR). When $\alpha = 0$, estimation based on Theorem 3 (IPW IV) is close to being unbiased and dominates the other methods in

[14]Note that in spite of $\alpha = 0.25$, estimation based on (the incorrect) Theorem 3 is consistent because the distribution of U is not associated with S conditional on D, M, X.

Table 3.2: Simulations with selection on unobservables, total population

	$\hat{\theta}(1)$			$\hat{\theta}(0)$			$\hat{\delta}(1)$			$\hat{\delta}(0)$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse	bias	std	rmse
	$\alpha = 0$, $n = 1000$											
IPW w. $S = 1$	-0.28	0.13	0.31	-0.27	0.16	0.32	0.07	0.16	0.18	0.07	0.12	0.14
IPW MAR (Th. 1)	-0.09	0.30	0.31	-0.11	0.21	0.24	0.06	0.14	0.15	0.04	0.15	0.16
IPW IV (Th. 3)	0.02	0.32	0.32	-0.01	0.31	0.31	-0.02	0.18	0.18	-0.05	0.16	0.16
	$\alpha = 0$, $n = 4000$											
IPW w. $S = 1$	-0.28	0.07	0.29	-0.28	0.08	0.29	0.08	0.08	0.12	0.09	0.06	0.11
IPW MAR (Th. 1)	-0.11	0.16	0.20	-0.11	0.10	0.15	0.06	0.07	0.09	0.06	0.09	0.11
IPW IV (Th. 3)	0.01	0.17	0.17	-0.01	0.16	0.16	-0.02	0.09	0.09	-0.04	0.08	0.09
	$\alpha = 0.25$, $n = 1000$											
IPW w. $S = 1$	-0.37	0.13	0.39	-0.35	0.15	0.38	0.05	0.16	0.16	0.07	0.12	0.14
IPW MAR (Th. 1)	-0.20	0.30	0.36	-0.20	0.21	0.28	0.03	0.14	0.14	0.04	0.15	0.16
IPW IV (Th. 3)	-0.14	0.32	0.34	-0.16	0.31	0.35	-0.02	0.18	0.18	-0.05	0.16	0.16
	$\alpha = 0.25$, $n = 4000$											
IPW w. $S = 1$	-0.38	0.07	0.38	-0.36	0.08	0.36	0.06	0.08	0.10	0.09	0.06	0.11
IPW MAR (Th. 1)	-0.22	0.16	0.27	-0.20	0.10	0.22	0.04	0.07	0.08	0.06	0.09	0.11
IPW IV (Th. 3)	-0.14	0.16	0.22	-0.16	0.16	0.23	-0.01	0.09	0.09	-0.04	0.08	0.09

Note: 'std' and 'rmse' report the standard deviation and root mean squared error, respectively.

terms of RMSE under the larger sample size ($n = 4000$). When $\alpha = 0.25$, however, also the latter approach is biased due to the violation of Assumption 7. Therefore, Table 3.3 considers the estimation of natural effects among the selected population only ($\theta_{S=1}(d)$, $\delta_{S=1}(1)$) in the presence of the D-U-interaction effect. We investigate the performance of estimation based on Theorem 2 ('IPW IV w. $S = 1$'), as well as of IPW among the selected ignoring selection. While the latter approach is biased, the former is close to being unbiased, but less precise. Under the larger sample size, our approach dominates both in terms of unbiasedness and RMSE.[15]

Table 3.3: Simulations with selection on unobservables, selected population ($S = 1$)

	$\hat{\theta}_{S=1}(1)$			$\hat{\theta}_{S=1}(0)$			$\hat{\delta}_{S=1}(1)$			$\hat{\delta}_{S=1}(0)$		
	bias	std	rmse	bias	std	rmse	bias	std	rmse	bias	std	rmse
	$\alpha = 0.25$, $n = 1000$											
IPW w. $S = 1$	-0.11	0.13	0.17	-0.09	0.15	0.17	0.05	0.16	0.16	0.07	0.12	0.14
IPW IV w. $S = 1$ (Th. 2)	0.00	0.21	0.21	-0.03	0.23	0.23	0.02	0.17	0.17	-0.01	0.12	0.12
	$\alpha = 0.25$, $n = 4000$											
IPW w. $S = 1$	-0.12	0.07	0.14	-0.10	0.08	0.12	0.06	0.08	0.10	0.09	0.06	0.11
IPW IV w. $S = 1$ (Th. 2)	0.01	0.10	0.10	-0.02	0.11	0.12	0.03	0.08	0.08	-0.00	0.06	0.06

Note: 'std' and 'rmse' report the standard deviation and root mean squared error, respectively.

[15]Results are very similar when setting $\alpha = 0$ and therefore omitted.

3.5 Empirical application

This section illustrates the evaluation of direct and indirect treatment effects in the presence of sample selection using data from Project STAR (Student-Teacher Achievement Ratio), an educational experiment conducted from 1985 to 1989 in Tennessee, USA. In the experiment, a cohort of students entering kindergarten and their teachers were randomly assigned within their school to one of three class types: small (13 – 17 students), regular (22 – 26 students), or regular with an additional teacher's aid. Students were supposed to remain in the assigned class type through third grade, returning to regular classes afterwards. The goal of Project STAR was to investigate the impact of class size on academic achievement measured by standardized and curriculum-based tests in mathematics, reading, and basic study skills. Numerous studies found positive effects of reduced class size on academic performance both short- (Folger and Breda, 1989; Finn and Achilles, 1990; Krueger, 1999), mid- (Finn et al., 1989), long-term (Nye et al., 2001; Krueger and Whitmore, 2001), and even on later-life outcomes (Chetty et al., 2011). While benefits of small classes are well documented, the causal mechanisms underlying the effect are less well-understood. Finn and Achilles (1990) argue that the impact is likely driven by classroom processes related to higher teacher morale and satisfaction translated to students, increased teacher-student interactions and time for individual attention, and student involvement in learning activities.

We investigate whether the effect of reduced class size on academic performance is mediated by the number of days absent from school. There might be several explanations for why class size affects days of absence. A smaller concentration of children in a classroom may be related to reduced transmission of infectious diseases and hence absenteeism.[16] Increased student involvement and closer teacher-student relationships in smaller classes may represent further channels making children and their parents more engaged and less likely to miss classes. As for the link between school absence and academic performance, a number of studies demonstrated a negative association between the two, see for instance Gershenson et al. (2017), Gottfried (2009), and Morrissey et al. (2014).

We compare results using the IPW MAR estimator ('IPW MAR' in Table 3.5) based on Theorem 1 (relying on Assumptions 1 through 4) in Section 3.2 to three

[16] Odongo et al. (2017) find a positive correlation between school size and communicable disease prevalence rates in Kenya. We are, however, not aware of any such study considering class (rather than school) size.

previously developed mediation estimators that ignore sample selection:[17] (i) a linear mediation estimator allowing for treatment-mediator interactions but neither accounting for observed pre-treatment confounders, nor selection, which is numerically equivalent to the decomposition of Blinder (1973) and Oaxaca (1973) ('Lin w. $S = 1$, no X');[18] (ii) a semiparametric IPW-based analog of the linear mediation estimator not accounting for confounding considered in Huber (2015) ('IPW w. $S = 1$, no X'); and (iii) the IPW estimator suggested in Huber (2014a) that incorporates observed pre-treatment covariates X but ignores sample selection when estimating the effect for the total population ('IPW w. $S = 1$'). We apply the same trimming rule as in the simulations presented in Section 3.4, which discards observations with treatment propensity scores $\hat{p}(M, X)$ smaller than 0.05 or larger than 0.95 or with $\hat{\pi}(D, M, X)$ smaller than 0.05. However, no observations are dropped for any IPW method, as such extreme propensity scores do not occur in our sample.

The treatment (D) is a binary indicator which is one if a child entering kindergarten was enrolled in a small class and zero otherwise.[19] The outcome (Y) is the first grade score in the Stanford Achievement Test (SAT) in mathematics. For IPW MAR estimation, a selection indicator S for missing outcomes is generated and all observations in our evaluation sample are preserved, such that effects are estimated for the entire population. In the case of the remaining three estimators, the evaluation is based on the data with non-missing Y, such that estimation relies on the selected sample only. The mediator (M) is the number of days a child was absent during the kindergarten year. Observed covariates (X) consist of child's race, gender, year of birth, and free lunch status as a proxy for socio-economic status. They are controlled for in the 'IPW w. $S = 1$' and 'IPW MAR' estimators. Even if these variables are initially balanced due to the random assignment of D, they might confound M and Y, implying that they are imbalanced when conditioning on the mediator for estimating direct and indirect effects.[20]

[17] We do not consider IPW IV estimation based on Theorems 2 and 3, as our data do not contain credible instruments.

[18] See Huber (2015) on the equivalence of conventional wage gap decompositions and a simple mediation model.

[19] Following Chetty et al. (2011), we consider regular class size with and without additional teaching aid to be one treatment.

[20] For example, Ready (2010) reports a stronger negative impact of absenteeism on early literacy outcomes for students with lower socioeconomic status, which implies that socioeconomic status and absenteeism interact in explaining the outcome. If socioeconomic status in addition affects absenteeism, it is a confounder of the association between absenteeism and the literacy outcomes.

We restrict the initial sample of 11,601 children to 6,325 observations who were part of Project STAR in kindergarten such that their treatment status was observed.[21] About 30% of participants in the kindergarten year were randomized into small classes. Table 3.4 presents summary statistics for the variables included in our empirical illustration for individuals without any missing values in the covariates by treatment status ($d = 0$ is for children randomized into the regular-size classes and $d = 1$ is for children in small classes). It shows a positive and statistically significant association between reduced class size and the average score in the standardized math test. Furthermore, children in small classes are, on average, about 0.7 days less absent and this difference is significant at the 5% level. There are no statistically significant differences in students' gender, race,[22] and free lunch status across treatment states due to treatment randomization. The sample is not perfectly balanced in terms of students' years of birth: children born in 1978 and 1980 are less likely to be in small classes (differences are statistically significant at the 1 and 10% levels, respectively), while those born in 1979 are more likely to be in small classes (significant at the 5% level). There is substantial attrition: math SAT scores in the first grade are observed for only 70% of program participants in the kindergarten year. The number of missing values in other key variables is much smaller. In the estimations, observations with missing values in M or X are dropped, which concerns all in all 83 cases, or about 1% of the sample.

Table 3.4: Mean covariate values by treatment status

Variable	Total	$d = 0$	$d = 1$	Difference	p-value
Student's gender: male	0.51	0.51	0.51	0.00	0.96
	[0.50]	[0.50]	[0.50]	(0.01)	
Student's race: white	0.67	0.67	0.68	0.01	0.42
	[0.47]	[0.47]	[0.47]	(0.02)	
Free lunch	0.48	0.49	0.47	-0.02	0.25
	[0.50]	[0.50]	[0.50]	(0.02)	
Born 1978	0.01	0.01	0.00	-0.01	0.00
	[0.08]	[0.09]	[0.05]	(0.00)	
Born 1979	0.23	0.22	0.25	0.03	0.04
	[0.42]	[0.42]	[0.43]	(0.01)	

Continued on next page

[21] 5,276 students joined the program in subsequent years. About 2,200 entered the experiment in the first grade, 1,600 in the second and 1,200 in the third grade.

[22] Less than 1% of students in the sample are Asian, Hispanic, Native American or other race. In our analysis, they are included in one group with black students.

Table 3.4 – continued from previous page

Variable	Total	$d=0$	$d=1$	Difference	p-value
Born 1980	0.76	0.77	0.74	-0.02	0.09
	[0.43]	[0.42]	[0.44]	(0.01)	
Born 1981	0.00	0.00	0.00	0.00	0.87
	[0.03]	[0.03]	[0.03]	(0.00)	
Kindergarten days absent	10.51	10.72	10.01	-0.71	0.02
	[9.76]	[9.95]	[9.29]	(0.31)	
Math SAT grade 1	534.54	531.52	541.25	9.73	0.00
	[43.83]	[42.92]	[45.10]	(2.14)	

Note: Standard deviations are in squared brackets. Cluster-robust standard errors are in parentheses.

Table 3.5 provides point estimates ('est.'), cluster-robust standard errors ('s.e.') based on blockbootstrapping the effects 1999 times, and p-values for the total treatment effect, as well as natural direct and indirect effects under treatment and non-treatment ($\hat{\theta}(1)$, $\hat{\theta}(0)$, $\hat{\delta}(1)$, $\hat{\delta}(0)$) for the four estimators.

Table 3.5: Effects of small class size in kindergarten on the math SAT in grade 1

	Total effect			$\hat{\theta}(1)$			$\hat{\theta}(0)$			$\hat{\delta}(1)$			$\hat{\delta}(0)$		
	est.	s.e.	p-val	est.	s.e.	p-val	est.	s.e.	p-val	est.	s.e.	p-val	est.	s.e.	p-val
IPW MAR	8.74	2.37	0.00	8.52	2.36	0.00	7.75	2.70	0.00	0.99	0.79	0.21	0.23	0.13	0.09
Lin w. $S=1$, no X	9.73	2.16	0.00	9.46	2.17	0.00	9.55	2.15	0.00	0.27	0.18	0.12	0.18	0.13	0.16
IPW w. $S=1$, no X	9.73	2.16	0.00	9.55	2.15	0.00	9.43	2.18	0.00	0.30	0.21	0.16	0.18	0.13	0.15
IPW w. $S=1$	9.20	2.14	0.00	9.01	2.14	0.00	8.77	2.19	0.00	0.43	0.32	0.18	0.19	0.14	0.18

Note: Cluster-robust standard errors ('s.e.') and p-values ('p-val') for the point estimates ('est.') are obtained by bootstrapping the latter 1999 times.

The total average effect of small class assignment is very similar across all the methods and highly statistically significant, amounting to an increase of almost 10 points. Furthermore, we find that if anything, the contribution of the indirect effects due to reduced days of absence is positive, but rather modest, ranging 0.18 to 0.99 points across different methods and treatment states. The IPW MAR estimator yields the largest indirect effects (amounting to 3 – 11% of the total effect), and the indirect effect on the non-treated group is statistically significant at the 10% level. It is thus the direct effects, which are highly statistically significant for any method, that mostly drive the total effect. IPW MAR yields direct effect estimates of 8.52 points under treatment and 7.75 points under non-treatment, which is slightly smaller than those of the other estimators

exploiting the subsample with non-missing outcomes only (ranging from 9.01 to 9.55 points under treatment and from 8.77 to 9.55 points under non-treatment). We therefore conclude that causal mechanisms not observed in the data (possibly including teacher motivation and individual teacher-student interaction) and entering the direct effect are much more important than absenteeism for explaining the effect of small kindergarten classes on math performance.

3.6 Conclusion

In this paper, we proposed an approach for disentangling a total causal effect into a direct component and a indirect effect operating through a mediator in the presence of outcome attrition or sample selection. To this end, we combined sequential conditional independence assumptions about the assignment of the treatment and the mediator with either selection on observables/missing at random or instrumental variable assumptions on the outcome attrition process. We demonstrated the identification of the parameters of interest based on inverse probability weighting by specific treatment, mediator, and/or selection propensity scores and outlined estimation based on the sample analogs of these results. We also provided a brief simulation study and an empirical illustration based on the Project STAR experiment in the U.S. to evaluate the direct and indirect effects of small classes in kindergarten on math test scores in first grade.

Chapter 4

On the Sensitivity of Wage Gap Decompositions

4.1 Introduction[1]

A vast empirical literature is concerned with the analysis and decomposition of gender wage gaps. Blinder (1973) and Oaxaca (1973) (see also Duncan, 1967) suggested a linear method allowing disentangling the total gap into an explained part that is linked to differences in observed characteristics, for instance education, and an unexplained part that is linked to unobserved factors, for instance discrimination. Several studies proposed non-parametric decomposition methods dropping the linearity assumptions, see for instance DiNardo et al. (1996), Barsky et al. (2002), Frölich (2007), Mora (2008), and Ñopo (2008). Finally, another branch of the literature suggested decomposition methods at quantiles (rather than means) of the wage distribution, see for instance Juhn et al. (1993), DiNardo et al. (1996), Machado and Mata (2005), Melly (2005), Firpo et al. (2007), Chernozhukov et al. (2009), and Firpo et al. (2009).

The aforementioned methods ignore the potential endogeneity of the observed characteristics, which are typically 'bad controls' in the sense of Angrist and Pischke (2009) as they are determined later in life, i.e. after gender. This implies that the explained and unexplained parts do not correspond to the true causal mechanisms related to observed and unobserved factors, respectively, through which gender influences wage. For this reason, policy conclusions – for instance about the magnitude of discrimination – are difficult to derive from such conventional decompositions, see Kunze (2008), Huber (2015), and Yamaguchi (2014) for related criticisms. Using an approach that comes from the literature on nonparametric causal mediation analysis (see for instance Robins and Greenland, 1992 and Pearl, 2001), Huber (2015) controls for observed confounders at birth as one possible approach to improve upon the endogeneity issue. However, a further threat to identification is sample selection (see Heckman, 1976b and Heckman, 1979) due

[1] This essay was written in co-authorship with Martin Huber. It was released as a SES Working paper, University of Fribourg (Huber and Solovyeva, 2018b).

to the fact that wages are only observed for those who work. For this reason, Neuman and Oaxaca (2003) and Neuman and Oaxaca (2004) combine classic decompositions with Heckman-type sample selection correction.[2] Alternatively, Maasoumi and Wang (2016) apply the copula approach of Arellano and Bonhomme (2010) to model the joint distribution of the quantile of the wage distribution and selection. In the presence of panel data, Blau and Kahn (2006) and Olivetti and Petrongolo (2008)[3] consider proxying non-observed wages by the observed wage in the closest period.[4] Finally, few studies aim at controlling for both endogeneity and sample selection. García et al. (2001) combine instrumental variable regression to control for the endogeneity of one of the observed characteristics (education) with Heckman-type sample selection correction in a parametric framework. The more flexible causal mediation method by Huber and Solovyeva (2018a) aims at tackling endogeneity by conditioning on observed potential confounders and sample selection by controlling for the selection probability based on observables and/or instruments.

In this paper, we investigate the sensitivity of average wage gap decompositions to various methods ignoring and considering endogeneity and sample selection, to provide insights on the robustness of decompositions across identifying assumptions. To this end, we consider US wage data collected in the year 2000 coming from the National Longitudinal Survey of Youth 1979 (NLSY). The latter is a panel study of young individuals in the US aged 14 to 22 years in 1979. The analysed estimators include the Oaxaca-Blinder decomposition; semiparametric inverse probability weighting (IPW, see Hirano et al., 2003), which eases linearity but ignores endogeneity and sample selection just as the Oaxaca-Blinder decomposition; IPW controlling for potential confounders at birth to mitigate endogeneity as in Huber (2015) but ignoring sample selection; and the approaches proposed in Huber and Solovyeva (2018a) to tackle both endogeneity and sample selection.

We find that the explained and unexplained wage gap components are generally not stable across methods. Even the total gap estimates differ non-negligibly between methods ignoring and controlling for sample selection. Although we do not claim that

[2]See also the method of Machado (2017), which permits arbitrary unobserved heterogeneity in the selection process.

[3]Olivetti and Petrongolo (2008) also estimate the Manski bounds (Manski, 1989) on the distribution of wages, using the actual and the imputed wage distributions. Bičakova (2014) derives bounds on gender unemployment gaps.

[4]As an alternative use of panel data, Lemieux (1998) combines fixed effect estimation with decomposition methods and allows for heterogeneity of the return to fixed effects across groups. However, this strategy depends on individuals switching groups, which is rarely the case for gender.

any of the estimators is capable of fully tackling identification concerns, our results cast doubts about the usefulness of standard decompositions used in the vast majority of empirical studies, which ignore endogeneity and sample selection altogether. We also investigate the robustness of our findings w.r.t. the definition of the observed characteristics. In our main specification, we include both levels as well as histories of such characteristics (e.g., current occupation as well as years in current occupation). In a robustness check, we only keep the levels and omit histories (as it appears to be the convention in many decompositions) and find this to reduce the explained and increase the unexplained component across our estimators. In light of the sensitivity of some of our results w.r.t. methods and variable definitions, we advise caution when basing policy recommendations (which typically require a proper identification of the causal mechanisms underlying the wage gap) on the outcomes of wage decompositions. This seems important given that the empirical literature on wage decompositions appears to have paid comparably little attention to identification issues that may jeopardize the interpretability of the parameters of interest.

Goraus et al. (2015) provide a further study systematically investigating the robustness of wage gap decompositions across specifications, considering the Polish Labor Force Survey. The authors compare estimates of the unexplained component across parametric and nonparametric methods for both means and quantiles. They also analyze issues of common support (or overlap) in observed characteristics across females and males and selection into employment based on Heckman-type sample selection corrections. Their results suggest that enforcing versus not enforcing common support in the characteristics has a non-negligible impact on the estimates. Also our IPW procedures enforce common support by specific trimming rules to ensure the comparability of observations across gender and employment states in terms of observables. The sample selection corrections, on the other hand, barely affect estimates of the unexplained component in Goraus et al. (2015). We also find that our weighting-based sample selection corrections change the unexplained component moderately when compared to IPW controlling for potential confounders alone, while more variation is observed for the total wage gap and the explained component. We point out that one major distinction of our study and Goraus et al. (2015) is that they do not consider methods that control for confounders at birth to tackle the endogeneity of the observed characteristics.

The remainder of this paper is organized as follows. Section 4.2 formally discusses the econometric parameters of interest and the identifying assumptions required for the various methods considered to consistently decompose wage gaps into observed and unobserved

causal mechanisms. Section 4.3 discusses the NLSY data, sample definition, and descriptive statistics. Section 4.4 presents and interprets the estimation results. Section 4.5 concludes.

4.2 Identification

Fortin et al. (2011) pointed out that while it is standard in econometrics to first discuss identification and then introduce appropriate estimators, most studies in the field of wage gap decompositions jump directly to estimation without clarifying identification first. Here, we first define what, in our opinion, should be the parameters of interest to be able to derive useful policy recommendations. To this end, let G denote a binary group dummy for gender, Y the outcome of interest (e.g., log wage) and X the vector of observed characteristics (e.g., education, work experience, occupation, industry, and others). We assume that G causally precedes X, which appears intuitive, as gender is determined even prior to birth, while X is determined by decisions later in life. G might influence Y 'indirectly' via its effect on X, i.e. by a causal mechanism related to observed characteristics. For instance, gender may have an effect on wage because females and males select themselves into different occupations. G might affect Y also 'directly', i.e. through factors not observed by the researcher, such that they do not appear in X. For instance, gender could have an impact on the perception of individual traits by decision makers in the labor market (see Greiner and Rubin, 2011), which in turn may entail discriminatory behavior. A graphical representation of this causal framework is given in Figure 4.1, where arrows represent causal effects: G influences Y either through X or 'directly'.

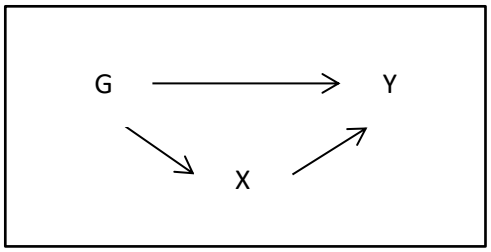

Figure 4.1: A graphical representation of the decomposition under Assumption 1

For a formal definition of the causal mechanisms running through observed characteristics X and unobserved factors as parameters of interest, we denote by $Y(g)$ and $X(g)$ the potential outcomes and characteristics when exogenously setting gender G

to a specific g, with $g \in \{1, 0\}$.[5] $E(X(1)) - E(X(0))$ gives the average causal effect of G on X (represented by the arrow of G to X in Figure 4.1), so to speak the 'first stage' of the indirect effect. $E(Y(1)) - E(Y(0))$, on the other hand, gives the total average causal effect of G on Y, represented by the sum of direct and indirect (i.e. operating through X) effects. Following the causal mediation literature (see Robins and Greenland, 1992 and Pearl, 2001), we further refine the potential outcome notation to be able to distinguish between the causal mechanisms in Figure 4.1: Let $Y(g) = Y(g, X(g))$, to make explicit that the potential outcome is affected by the group variable both directly and indirectly via $X(g)$. This permits rewriting the total effect of G on Y as $E(Y(1)) - E(Y(0)) = E[Y(1, X(1))] - E[Y(0, X(0))]$ and more importantly, it allows disentangling the latter into the causal mechanisms of interest. That is, the difference in potential outcomes due to a switch from $X(1)$ to $X(0)$ while keeping gender fixed at $G = 1$ yields the indirect effect denoted by ψ, while varying gender and fixing characteristics at $X(0)$ gives the direct effect η. Both together add up to the total causal effect:

$$E[Y(1, X(1))] - E[Y(0, X(0))] = \underbrace{E[Y(1, X(1))] - E[Y(1, X(0))]}_{\psi} + \underbrace{E[Y(1, X(0))] - E[Y(0, X(0))]}_{\eta}. \tag{4.1}$$

We now introduce the first identifying assumption considered in our empirical analysis, which rules out endogeneities of G, X and sample selection issues.

Assumption 1 (sequential independence):
(a) $\{Y(g', x), X(g)\} \perp G$ for all $g', g \in \{0, 1\}$ and x in the support of X,
(b) $Y(g', x) \perp X | G = g$ for all $g', g \in \{0, 1\}$ and x in the support of X,
(c) $Y(g, X)$ is linear X for $g \in \{0, 1\}$,
(d) $\Pr(G = 1 | X = x) > 0$ for all x in the support of X,

where '$\perp$' denotes statistical independence. Under Assumption 1(a), G is as good as randomly assigned, i.e. there are no factors confounding G on the one hand and Y and/or X on the other hand. Under Assumption 1(b), observed characteristics like education are as good as randomly assigned within gender, i.e. given G, so that there are no factors confounding X and Y. Assumption 1(c) imposes potential outcomes to be linear in X.

[5] See for instance Rubin (1974) for an introduction to the potential outcome framework.

Finally, Assumption 1(d) is a common support restriction. It implies that the conditional probability (the so-called propensity score) to belong to the reference group ($G = 1$), e.g., males, is larger than zero for any value in the support of X, such that for each female observation ($G = 0$), there exists a male who is comparable w.r.t. X.

The Oaxaca-Blinder decomposition consistently estimates ψ and η under Assumptions 1(a)-1(c). To see this, note that under Assumption 1(a), $E(X(g)) = E(X|G = g)$. Under Assumptions 1(a), 1(b), and 1(c), $E[Y(g, x)] = E(Y|G = g, X = x) = c_g + x\beta_g$, where c_g denotes a gender-specific constant and β_g denotes a vector of gender-specific coefficients on X in the respective female or male population. Finally, by iterated expectations, $E[Y(g, X(g'))] = c_g + E(X|G = g')\beta_g$ for $g, g' \in \{0, 1\}$. Therefore,

$$\psi = E[Y(1, X(1))] - E[Y(1, X(0))] = [E(X|G = 1) - E(X|G = 0)]\beta_1, \quad (4.2)$$

$$\eta = E[Y(1, X(0))] - E[Y(0, X(0))] = c_1 - c_0 + E(X|G = 0)(\beta_1 - \beta_0). \quad (4.3)$$

The left hand expressions in (4.2) and (4.3) correspond to the probability limits of the explained and unexplained components, respectively, in the Oaxaca-Blinder decompositions. For (4.2) and (4.3) to hold, Assumptions 1(a) and 1(b) could be relaxed to mean independence, while full independence needs to be maintained for decompositions of quantiles.[6]

Nonparametric approaches do not rely on the linearity assumption 1(c), but instead require common support as postulated in Assumption 1(d). This becomes obvious from considering the denominators of the following expressions based on inverse probability weighting (IPW) by the propensity score, which identify the parameters of interest as discussed in Huber (2015):

$$\psi = E\left[\frac{Y \cdot G}{\Pr(G = 1)}\right] - E\left[\frac{Y \cdot G}{\Pr(G = 1|X)} \cdot \frac{1 - \Pr(G = 1|X)}{1 - \Pr(G = 1)}\right], \quad (4.4)$$

$$\eta = E\left[\frac{Y \cdot G}{\Pr(G = 1|X)} \cdot \frac{1 - \Pr(G = 1|X)}{1 - \Pr(G = 1)}\right] - E\left[\frac{Y \cdot (1 - G)}{1 - \Pr(G = 1)}\right]. \quad (4.5)$$

(4.5) is identical to the identification result for the average treatment effect on the non-treated (see Hirano et al., 2003 for IPW-based treatment evaluation in subgroups based on reweighting), even though the causal framework differs. In classic treatment evaluation, one typically controls for pre-treatment (or pre-group) variables to tackle the

[6]However, analogous results to (4.2) and (4.3) cannot be applied to quantile decompositions, because the law of iterated expectations does not apply, see Fortin et al. (2011).

endogeneity of the treatment (or group). Here, X are post-group variables such that conditioning allows separating the indirect causal mechanism via X from the direct one related to unobservables. Obviously, this is only feasible if neither G nor X given G are endogenous as postulated in Assumption 1. In the empirical application presented in Section 4.4, we consider both the Oaxaca-Blinder decomposition and estimation based on the sample analogues of (4.4) and (4.5).

In a next step, we ease Assumption 1 by assuming that the identifying restrictions need not hold unconditionally but conditional on a set of observed covariates measured at birth and denoted by W. This allows for endogeneity of X, as long as it can be tackled by W. The dashed arrow going from W to G in Figure 4.2 even points to the possibility of an endogenous G. This may appear unnecessary when assuming gender to be randomly assigned by nature. However, specific interventions like selective abortions could in principle jeopardize randomization, which is permitted in Assumption 2 below as long as W captures all confounding.

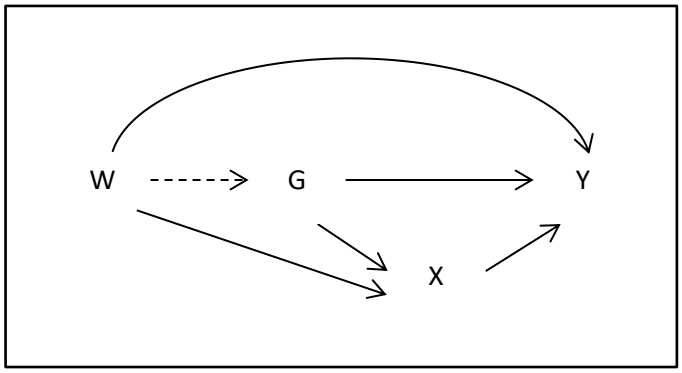

Figure 4.2: A graphical representation of the decomposition under Assumption 2

Assumption 2 (sequential conditional independence):
(a) $\{Y(g', x), X(g)\} \perp G|W$ for all $g', g \in \{0, 1\}$ and x in the support of X,
(b) $Y(g', x) \perp X|G = g, W = w$ for all $g', g \in \{0, 1\}$ and x, w in the support of X, W,
(c) $\Pr(G = 1|X = x, W = w) > 0$ and $0 < \Pr(G = 1|W = w) < 1$ for all x, w in the support of X, W.

Identical or similar conditions as Assumption 2 have been frequently applied in the literature on causal mediation analysis, see for instance Pearl (2001), and Imai et al. (2010). Assumptions 2(a) and (b) imply that after controlling for W, no unobserved variables confound either G and Y, G and X, or X and Y given G. Assumption 2(c)

is a refined common support restriction, requiring that the conditional probability of belonging to the reference group given X, W is larger than zero, while the conditional probability given W must neither be zero nor one. The latter implies that for each female in the population, there exists a comparable observation in terms of W among males and vice versa. Under Assumption 2, it follows from the results of IPW-based identification of direct and indirect effects in Huber (2014a) that

$$\psi = E\left[\frac{Y \cdot G}{\Pr(G=1|W)}\right] - E\left[\frac{Y \cdot G}{\Pr(G=1|X,W)} \cdot \frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right], \quad (4.6)$$

$$\eta = E\left[\frac{Y \cdot G}{\Pr(G=1|X,W)} \cdot \frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right] - E\left[\frac{Y \cdot (1-G)}{1-\Pr(G=1|W)}\right]. \quad (4.7)$$

Estimation of (ethnic) wage gaps based on (4.6) and (4.7) has been considered in Huber (2015), and is also among the methods investigated in our empirical application presented further below.

The approaches discussed so far abstract from sample selection stemming from the fact that wages are only observed for individuals in employment and that the decision to work is unlikely to be random. However, the previous sets of assumptions, even if satisfied in the total population, do not hold in the working subpopulation if selection into employment is related to factors that also affect the outcome, for instance, ability. To improve upon this problem both notationally and methodologically, we introduce a binary selection indicator S which is equal to one if an individual is employed, such that the wage outcome Y is observed in the data and zero otherwise. We maintain that G, X, W are observed for all individuals and note that each of these variables might affect S, which can be considered as yet another outcome variable.

Using the results of Huber and Solovyeva (2018a), one may combine Assumption 2 with specific restrictions on the nature of selection into employment. The first approach of Huber and Solovyeva (2018a) assumes selection to be related to the observed variables G, X, W only.

Assumption 3 (Selection on observables):
(a) $Y \perp S|G=g, X=x, W=w$ for all $g \in \{0,1\}$ and x, w in the support of X, W,
(b) $\Pr(S=1|G=g, X=x, W=w) > 0$ for all $g \in \{0,1\}$ and x, w in the support of X, W.

By Assumption 3(a), there are no unobservables confounding S and Y conditional on G, X, W, so that outcomes are missing at random (MAR) in the denomination of Rubin

(1976). The common support restriction implies that conditional on the values of G, X, W in their joint support, the probability to be observed is larger than zero; otherwise, no outcome is observed for some specific combinations of these variables and identification fails. Figure 4.3 presents a graphical illustration of the decomposition with selection on observables.

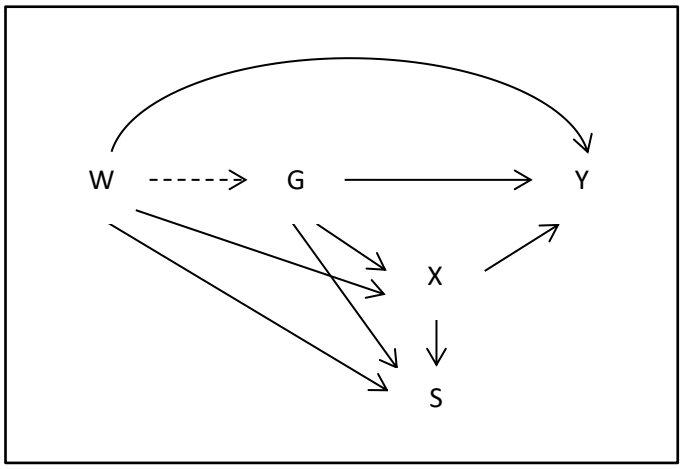

Figure 4.3: A graphical representation of the decomposition under Assumption 3

Under Assumptions 2 and 3, the parameters of interest are identified by the following IPW expression, which fits the general framework of IPW-based M-estimation of missing data models in Wooldridge (2002):

$$\begin{aligned}
\psi &= E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|W) \cdot \Pr(S=1|G,X,W)}\right] \\
&- E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|X,W) \cdot \Pr(S=1|G,X,W)} \cdot \frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right], \qquad (4.8) \\
\eta &= E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|X,W) \cdot \Pr(S=1|G,X,W)} \cdot \frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right] \\
&- E\left[\frac{Y \cdot (1-G) \cdot S}{(1-\Pr(G=1|W)) \cdot \Pr(S=1|G,X,W)}\right]. \qquad (4.9)
\end{aligned}$$

Alternatively to Assumption 3, Huber and Solovyeva (2018a) present a control function approach for the case that selection is related to unobservables affecting the outcome. This requires an instrument for selection, denoted by Z, which affects selection but is not directly associated with the outcome. Figure 4.4 provides a graphical representation of mediation with selection on unobservables and an instrument for selection. $\mathcal{E}$, V, and U denote

unobserved variables that affect the instrument for selection Z, the selection indicator S, and the outcome Y, respectively.

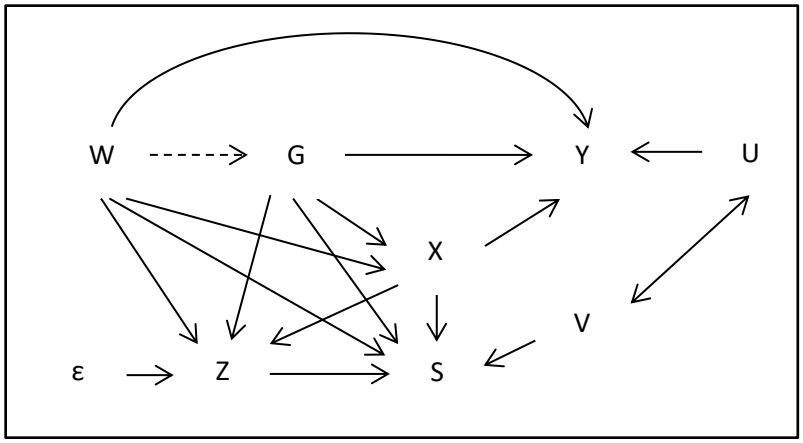

Figure 4.4: A graphical representation of the decomposition under Assumption 4

Assumption 4 (Instrument for selection):
a) There exists an instrument Z that may be a function of G, X, i.e. $Z = Z(G, X)$, is conditionally correlated with S, i.e. $E[Z \cdot S|G, X, W] \neq 0$, and satisfies (i) $Y(g, x, z) = Y(g, x)$ and (ii) $\{Y(g, x), X(g')\} \perp Z(g'', x')|W = w$ for all $g, g', g'' \in \{0, 1\}$ and z, x, x', w in the support of Z, X, W,
(b) $S = I\{V \leq \Pi(G, X, W, Z)\}$, where Π is a general function and V is a scalar (index of) unobservable(s) with a strictly monotonic cumulative distribution function conditional on W,
(c) $V \perp (G, X, Z)|W$,
(d) $E[Y(1, x) - Y(0, x)|W = w, V = v, S = 1] = E[Y(1, x) - Y(0, x)|W = w, V = v]$ and $E[Y(g, X(1)) - Y(g, X(0))|W = w, V = v, S = 1] = E[Y(g, X(1)) - Y(g, X(0))|W = w, V = v]$, for all $g \in \{0, 1\}$ and x, w, v in the support of X, W, V,
(e) $\Pr(G = 1|X = x, W = w, p(Q) = p(q)) > 0$, $0 < \Pr(G = 1|W = w, p(Q) = p(q)) < 1$, and $p(q) > 0$ for all $g \in \{0, 1\}$ and x, w, z in the support of X, W, Z.

In contrast to Assumption 3(a), the unobservable V in the selection equation is now allowed to be associated with unobservables U affecting the outcome. Therefore, the distribution of V generally differs across values of G, X conditional on W, which entails confounding. Identification hinges on exogenous shifts in the conditional selection probability $p(Q) = \Pr(S = 1|G, X, W, Z)$ based on instrument Z, with $Q = (G, X, W, Z)$ for the sake of brevity. By using $p(Q)$ as additional control variable in the decompositions, one controls

for the distribution of V and thus, for the confounding associations of V with (i) G and $\{Y(g,x), X(g')\}$ and (ii) X and $Y(g,x)$ that occur conditional on $S=1$.

Z and S have to satisfy particular conditions. Z must not affect Y or be associated with unobservables affecting X or Y conditional on W, as invoked in Assumption 4(a). By the threshold crossing model in Assumption 4(b), $p(Q)$ identifies the distribution function of V given W. Assumption 4(c) implies the (nonparametric) identification of the distribution of V, as the latter is independent of (G,X,Z) given W. Assumption 4(d) imposes homogeneity of the observed and unobserved causal mechanisms across employed and non-employed populations conditional on W,V. Without this restriction, wage decompositions can merely be conducted for the employed but not the total population, as effects might be heterogeneous in unobservables, see also the discussion in Newey (2007). A sufficient condition for effect homogeneity in unobservables is separability of observed and unobserved components in the outcome variable, i.e. $Y=\eta(G,X,W)+\nu(U)$, where η,ν are general functions and U is a scalar or vector of unobservables. Finally, the first part of Assumption 4(e) strengthens the previous common support assumption 2(c) to also hold when including $p(Q)$ as additional control variable. The second part requires the selection probability $p(Q)$ to be larger than zero for any combination of values in the support of G,X,W,Z to ensure that outcomes are observed for all values occurring in the population. Under Assumptions 2 and 4, the causal mechanisms are identified by the following expressions:

$$\begin{aligned}
\psi &= E\left[\frac{Y\cdot G\cdot S}{\Pr(G=1|W,p(Q))\cdot p(Q)}\right] \\
&- E\left[\frac{Y\cdot G\cdot S}{\Pr(G=1|X,W,p(Q))\cdot p(Q)}\cdot\frac{1-\Pr(G=1|X,W,p(Q))}{1-\Pr(G=1|W,p(Q))}\right], \qquad (4.10)\\
\eta &= E\left[\frac{Y\cdot G\cdot S}{\Pr(G=1|X,W,p(Q))\cdot p(Q)}\cdot\frac{1-\Pr(G=1|X,W,p(Q))}{1-\Pr(G=1|W,p(Q))}\right] \\
&- E\left[\frac{Y\cdot (1-G)\cdot S}{(1-\Pr(G=1|W,p(Q)))\cdot p(Q)}\right]. \qquad (4.11)
\end{aligned}$$

4.3 Data

Our data come from the National Longitudinal Survey of Youth 1979 (NLSY79), a panel survey of young individuals who were aged 14 to 22 years at the first wave in 1979.[7] Conducted annually until 1994, it then became biannual. The data contain a wealth of individual characteristics, including rich information relevant for labor market decisions, such as education, occupation, work experience and more. We estimate decompositions for wages reported in the year 2000 when respondents were 35 – 43 years old. After excluding 1,351 observations from the total NLSY79 sample in 2000 due to various data issues,[8] our evaluation sample consists of 6,658 individuals (3,162 men and 3,496 women). Table A4.1 in Appendix 4 provides descriptive statistics (mean values, mean differences, and respective p-values based on two-sample t-tests) for the key variables in our analysis. The group variable G is equal to zero for female and one for male respondents, such that male wages are regarded as reference wages, as it is frequently the case in the decomposition literature.[9] The outcome variable of interest (Y) is the log average hourly wage in the past calendar year reported in 2000. The selection indicator S is equal to one for individuals who indicated to have worked at least 1,000 hours in the past calendar year. This is the case for 87% of males and 70% of females.

The set of post-group characteristics X, which potentially mediate the effect of gender on wages, consists of individual variables reported in or constructed with reference to 1998: marital status, years in marriage, the region of residence and how many years an individual has been residing in that region, an indicator for living in an urban area (SMSA) and the number of years living in an urban area, education level, indicators for the year when first worked, number of jobs ever had, tenure with the current employer (in weeks), industry and the number of years working there, occupation and the number

[7] The NLSY79 data consist of three independent probability samples: a cross-sectional sample (6,111 subjects, or 48%) representing the non-institutionalized civilian youth; a supplemental sample (42%) oversampling civilian Hispanic, black, and economically disadvantaged nonblack/non-Hispanic young people; and a military sample (10%) comprised of youth serving in the military as of September 30, 1978 (Bureau of Labor Statistics, U.S. Department of Labor, 2001).

[8] Specifically, we excluded 502 persons who reported to have worked 1,000 hours or more in the past calendar year, but whose average hourly wages in the past calendar year were either missing or equal to zero. We also dropped 54 working individuals with average hourly wages of less than \$1 in the past calendar year. Furthermore, 608 observations with missing values in mediators (see Table A4.1 in Appendix 4 for the full list of mediators) and 186 observations with missing values in the instruments for selection – the number of young children and the employment status of the respondent's mother back when the respondent was 14 years old – were excluded.

[9] We refer to Sloczynski (2013) for a discussion of reference group choice in the potential outcome framework.

of years working in that occupation, whether employed in 1998 and total years of employment. Further characteristics are the form of employment (whether full-time), the share of full-time employment in employment years in 1994–98, total weeks of employment, the number of weeks unemployed and the number of weeks out of the labor force, and whether health problems prevented work. Moreover, several higher-order (squared and cubed) and interaction terms are included to make the propensity score specification more flexible. p-values of the two-sample t-tests in Table A4.1 in Appendix 4 reveal that women in our sample differ significantly (at the 5% level) from men in a range of variables. For instance, males have on average more labor market experience, while females have a higher average level of education. Important differences also arise in other factors related to labor market performance (e.g., industry, occupation, employment form, etc.).

Although X includes and even surpasses the set of variables conventionally used in wage decompositions, further potentially important characteristics mediating the effect of gender on wage are not considered. For instance, risk preferences, attitudes towards competition and negotiations, and other socio-psychological factors (see Bertrand 2011 and Azmat and Petrongolo 2014) are not available in our data. Their effects thus contribute to the unexplained component.

Potential confounders W related to factors determined at or prior to birth include race, religion, year of birth, birth order, parental place of birth (in the US or abroad), and parental education. We acknowledge that further confounders not available in our data but correlated with G, X, and/or Y likely exist. For instance, see Cobb-Clark (2016) for a review of biological factors, such as sensory functioning (e.g., time-space perceptions), emotions, and levels of sex hormones, potentially linking gender with labor market behavior and outcomes. In particular, some studies relate higher levels of prenatal testosterone to stronger preference for risk (Garbarino et al., 2011) and sorting into traditionally male-dominated occupations (Manning et al. 2010 and Nye and Orel 2015). Therefore, we do not claim that controlling for W fully tackles endogeneity bias. Nevertheless, we are interested in the sensitivity of decompositions w.r.t. to the inclusion and exclusion of W, even if these variables only comprise a subset of the actual confounders.

Finally, we define the number of children in 1999 younger than 6 and 15 years old, respectively, as instruments Z for selection into our employment indicator S. Such instruments based on the number of children in a household have been widely used as

instruments for labor supply in the empirical labor market literature, see for instance Mulligan and Rubinstein (2008). We, however, note that the validity of this approach is not undisputed, as the number of children might be correlated with unobservables also affecting the wage outcome, like relative preference for family and working life. For this reason, Huber and Mellace (2014) provided a method to partially test instrument validity, namely a joint test for the exclusion restriction and additive separability of the unobservable V in the selection equation. They applied them to children-based instruments for female labor supply in four data sets but found no statistical evidence for the violation of the IV assumptions. As a word of caution, however, their tests cannot detect all possible violations of instrument validity even asymptotically, as they rely on a partial identification approach. Even though concerns about the instruments may therefore remain, it is our aim to verify how sensitive decompositions are across different methods, also w.r.t. modelling selection based on instruments commonly used in the literature. In a robustness check, we consider an indicator for the respondent's mother working for pay back when the respondent was 14 years old as an additional instrument for selection. This, however, yields very similar point estimates based on (4.10) and (4.11) as when using the children-based instruments alone, see the discussion below.

4.4 Empirical results

We decompose the gender wage gap based on the five approaches outlined in Section 4.2. Table 4.1 provides the estimated effects (est.) along with standards errors (s.e.) and p-values (p-val) using 999 bootstrap replications. It also shows the shares (% tot.) of the explained and unexplained components in the total gender wage gap. The last two columns (Trimmed obs., %) indicate, respectively, the number and the share of units dropped in the IPW estimations due to a trimming rule that discards observations with extreme propensity scores larger than 0.99 and/or smaller than 0.01. This is done to prevent the assignment of very large weights to specific observations (due to small denominators in IPW) as a consequence of insufficient common support across gender or selection into employment.

Our main specification includes the full list of post-group characteristics (X) presented in Table A4.1 in in Appendix 4 as well as several higher-order and interaction terms.[10]

[10]The included higher-order terms are marriage history squared and cubed, tenure squared and cubed, and years in current occupation squared and cubed. The interaction terms are between binary indicators for region in 1998 and urban residency, first job before 1975, first job in 1976-79, industry indicators, and employment in 1998; between education indicators and occupation indicators, years in current occupation,

The standard Oaxaca-Blinder decomposition (Oaxaca-Bl.) based on (4.2) and (4.3) as well as IPW (IPW no W) based on (4.4) and (4.5) invoke Assumption 1 and thus neither control for the potential endogeneity of X nor for selection. Therefore, estimations are conducted in the subsample with $S = 1$. Under Assumption 2, IPW is based on (4.6) and (4.7) and includes potential confounders W listed in Table A4.1 in Appendix 4 (IPW with W) to tackle endogeneity. Under Assumption 3, IPW based on (4.8) and (4.9) uses these covariates to control for both endogeneity and selection (IPW MAR). Finally, under Assumption 4, IPW based on (4.10) and (4.11) in addition utilizes a combination of the number of children younger than 6 and 15 years old as instruments (Z) for selection into employment (IPW IV).

Table 4.1: Gender wage gap decomposition based on NLSY79: main specification

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)				Trimmed	
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	% tot.	obs.	%
Oaxaca-Bl.	0.299	0.019	0.000	0.083	0.021	0.000	28%	0.215	0.024	0.000	72%	0	0%
IPW no W	0.293	0.019	0.000	0.118	0.030	0.000	40%	0.176	0.031	0.000	60%	28	0%
IPW with W	0.264	0.017	0.000	0.096	0.028	0.001	36%	0.168	0.030	0.000	64%	28	0%
IPW MAR	0.365	0.035	0.000	0.219	0.033	0.000	60%	0.147	0.035	0.000	40%	90	1%
IPW IV	0.141	0.324	0.665	-0.005	0.102	0.964	-3%	0.145	0.328	0.658	103%	584	9%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications. The trimming rule discards observations with propensity scores (specific to each estimator) below 0.01 or above 0.99.

When applying the classic Oaxaca-Blinder decomposition, 28% (0.083) of the total gender wage gap[11] of 0.299 is attributed to differences in the included post-group characteristics X, while about 72% (0.215) remains unexplained. All estimates are highly statistically significant.[12] In contrast to the Oaxaca-Blinder decomposition, IPW without W does not impose linearity of Y in X given G but instead requires an estimate of the propensity score $\Pr(G = 1|X)$, which is obtained by logit regression. Figures A4.1 to A4.9 and Tables A4.2 and A4.3 in Appendix 4 present, respectively, histograms and summary statistics (minimum, mean, and maximum) of the within-group propensity scores used in our IPW-based estimations.[13] Figure A4.1 suggests a decent overlap in the distribution of

and the employment indicator 1998; and between tenure and the urban indicator, occupation indicators, years in current occupation, and the full-time employment indicator in 1998.

[11] Among the methods considered, the differences in the estimates of the total wage gap are statistically significant at the 10% level between the Oaxaca-Blinder and the IPW MAR estimators, Oaxaca-Blinder and IPW IV, IPW without and with controlling for W, IPW without W and IPW IV, and IPW MAR and IPW IV.

[12] The regression-based Oaxaca-Blinder estimator does not rely on common support, see the discussion in Section 4.2, and therefore does not require trimming observations with extreme propensity score values.

[13] Table A4.5 in Appendix 4 additionally provides the number and the share of trimmed observations for each propensity score.

estimates of $\Pr(G = 1|X)$, implying common support in observed characteristics across females and males over most of the support of X. Applying a trimming rule that excludes observations with propensity scores below 0.01, we drop 28 units from the sample. Compared to the Oaxaca-Blinder decomposition, the explained component is slightly larger and the unexplained component is somewhat smaller, while total wage gap remains almost unchanged. For IPW including potential confounders W, Figures A4.2 and A4.3 in Appendix 4 display the histograms of the logit-based estimates of $\Pr(G = 1|W)$ and $\Pr(G = 1|X, W)$ and point to decent common support w.r.t. either propensity score. Therefore, (only) the same 28 observations as for IPW are without controls dropped from the sample. Controlling for W leads to moderately smaller estimates of the total wage gap as well as the explained and unexplained components when compared to IPW without controls.

IPW MAR relies on estimating the selection propensity score $\Pr(S = 1|G, X, W)$ to control for the employment decision based on observables, again by logit regression. Figure A4.6 in Appendix 4 presents histograms of estimated selection probabilities for individuals who worked less than 1,000 hours in the past calendar year ($S = 0$) and those who worked 1,000 hours or more ($S = 1$). We note that the selection probability is close to zero for a subset of individuals but clearly larger than zero for most of the sample. 90 (1%) observations are dropped from estimation, once the additional condition that selection propensity scores must not be smaller than 0.01 is added to the previous trimming rule. The total wage gap (0.365 log points) and the explained component (0.219 log points) are considerably larger than under IPW controlling for W (but ignoring selection). In contrast, the magnitude of the unexplained component (0.147 log points) is slightly smaller, resulting in an overall drop of its share in the total wage gap to 40%. All estimates discussed so far are statistically significant at the 1% level.

In addition to controlling for observables, our last estimator, IPW IV, uses the number of children under 15 and under 6 years as instruments to control for selection. It requires the estimation of $p(Q) = \Pr(S = 1|Q)$ (with $Q = (G, X, W, Z)$), $\Pr(G = 1|W, p(Q))$, and $\Pr(G = 1|X, W, p(Q))$. Figures A4.7, A4.8, and A4.9 provide the logit estimates of the respective propensity scores. Common support is by and large satisfactory. The trimming rule discards observations with estimates of $\Pr(G = 1|X = x, W = w, p(Q) = p(q)) < 0.01$, of $\Pr(G = 1|W = w, p(Q) = p(q)) > 0.99$, and of $p(q) < 0.01$, all in all 584 cases (9%). This needs to be kept in mind when interpreting the results, as trimming generally changes the target population for which the parameters are estimated. The total wage gap drops substantially when compared to previous estimates and amounts to 0.141 log points. The

unexplained component is similar in magnitude to the IPW MAR estimate, while the explained part is very close to zero but even negative. However, the IPW IV estimates are far from being statistically significant at any conventional level, pointing to a weak instrument problem.

We conduct several sensitivity checks by gradually reducing the set of post-group characteristics X. Table A4.6 in Appendix 4 presents the estimates obtained when dropping any higher-order and interaction terms of X, such that the functional forms in the outcome and propensity score specifications become less flexible. While the total wage gap estimates remain largely unchanged, the explained components generally decline slightly (by about 0.03 log points), and the unexplained components increase, on average, by the same amount. The exception is the IPW IV decomposition, where both the total gap and its explained component somewhat increase, whereas the size and the share of the unexplained component decline. However, all the IPW IV estimates remain statistically insignificant. All in all, these differences are minor, which suggests that our results are rather robust to the exclusion of higher-order and interaction terms of X.

Our next robustness check excludes not only the higher-order and interaction terms, but also all variables in X that reflect developments or histories like years in marriage, years worked in current occupation, etc. We point out that many of these variables are frequently not included in wage decompositions, even though they appear a priori similarly important as characteristics measured at a particular point in time. For instance, one would suspect that not only the current occupation matters for human capital accumulation and the determination of the current wage, but also employment history and tenure in the current occupation. The exclusion of these additional variables generally decreases the explained component and increases the unexplained component, which accounts for 77% to 96% of the total gap across the first four methods. IPW IV yields different and even more extreme estimates, which are, however, at best marginally significant. Table 4.2 provides the results.

Table 4.2: Robustness check: parsimonious set of X

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)				Trimmed	
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	% tot.	obs.	%
Oaxaca-Bl.	0.299	0.019	0.000	0.067	0.019	0.000	22%	0.231	0.022	0.000	77%	0	0%
IPW no W	0.298	0.019	0.000	0.026	0.023	0.269	9%	0.272	0.026	0.000	91%	1	0%
IPW with W	0.269	0.017	0.000	0.011	0.023	0.648	4%	0.258	0.027	0.000	96%	2	0%
IPW MAR	0.362	0.032	0.000	0.076	0.025	0.002	21%	0.287	0.032	0.000	79%	1	0%
IPW IV	0.124	0.324	0.703	-0.186	0.102	0.067	-151%	0.310	0.328	0.345	251%	850	13%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications. The trimming rule discards observations with propensity scores (specific to each estimator) below 0.01 or above 0.99.

The Oaxaca-Blinder decomposition yields quite stable estimates when compared to the main specification of Table 4.1. The total gap estimate does not change, while the explained component decreases and the unexplained component increases each by about 0.02 log points, or about 5 percentage points of the total gap. For the IPW estimators not accounting for selection, the explained components decline by about 0.1 log point, now constituting only a small share of the total gap and losing their statistical significance. Over 90% of the total wage gap remains unexplained both for IPW with and without controlling for W. Also for the IPW estimators accounting for selection, the explained components decrease considerably, while the explained components increase and the total gap is slightly smaller than before. In the case of IPW MAR, the unexplained part now accounts for nearly 80% of the total wage gap. All the IPW MAR estimates are statistically significant at the 1% level. The IPW IV estimator yields rather implausible results. The large unexplained component of 0.31 log points comprises 251% of the total wage gap, due to a negative estimate of the explained component. However, none of these estimates are statistically significant at the 5%.

As a final robustness check for IPW IV, we add an indicator for whether an individual's mother worked for pay when the individual was 14 years old as an additional instrument for selection into paid work. Table A4.7 in Appendix 4 shows that the estimates remain unchanged compared to the main specification. Overall, our empirical results suggest that estimates of the gender wage decomposition are dependent on the choice of underlying identification assumptions and, to some extent, the definition of the observed characteristics X. Given the variability of estimates across methods and specifications, we advise to be cautious w.r.t. the use of wage decompositions for policy conclusions, for instance about the magnitude of gender discrimination in the labor market.

4.5 Conclusion

We assessed the sensitivity of average gender wage gap decompositions in data from the US National Longitudinal Survey of Youth 1979, comparing several decomposition methods and sets of included variables. We first discussed the identification problem from a causal perspective, namely separating the explained component of the wage effect of gender operating through observed characteristics from the unexplained component. Five decomposition techniques were reviewed. Starting with the linear Oaxaca-Blinder decomposition, we gradually relaxed the identifying assumptions regarding functional

form, exogeneity of observed characteristics and gender, and selection into employment. Specifically, we considered inverse probability weighting (IPW) as a semiparametric analog of the standard Oaxaca-Blinder decomposition. We also included IPW versions controlling for confounders (of observed characteristics, gender, and the wage outcome) or for both confounders and sample selection into employment, the latter either based on observed variables or instruments. When applying all five estimators to the data, we also considered less and more parsimonious definitions of the observed characteristics and instruments included in the analysis.

We found the total wage gap as well as the explained and unexplained components to differ importantly across some of the methods considered. Furthermore, the definition of the observed characteristics related to the explained component mattered: Including only levels of variables rather than both levels and histories generally reduced the explained and increased the unexplained components across the considered estimators. Given our results, the usefulness of wage decompositions that neither account for identification issues like endogeneity and selection into employment nor for histories of observed characteristics appears questionable in terms of policy conclusions, for instance, when aiming at quantifying gender discrimination. Unfortunately, a vast number of empirical applications rely on exactly such kind of decompositions. At the very least, we advise checking the robustness of the results across several decomposition methods and variable specifications to improve upon the status quo of the literature.

Bibliography

Abowd, J., Crepon, B., and Kramarz, F. (2001). Moment estimation with attrition: An application to economic models. *Journal of the American Statistical Association*, 96:1223–1230.

Achuonjei, P., dos Santos, F., and Reyes, Y. (2003). Farmers' perceptions of the effectivness of the Agricultural Rehabilitation Project (ARP) information campaign, East Timor. Proceedings of the 19th Annual Conference of Association for International Agricultural and Extension Education (Raleigh, North Carolina, USA). Available at `https://www.aiaee.org/attachments/article/1236/Achuonjei11.pdf` (accessed August 15, 2017).

Ahn, H. and Powell, J. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58:3–29.

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine*, 27:1282–1304.

Albert, J. M. and Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, 67:1028–1038.

Altbach, P. G. (2016). *Global perspectives on higher education*. John Hopkins University Press, Baltimore.

Altmann, S., Falk, A., Jaeger, S., and Zimmermann, F. (2015). Learning about job search: A field experiment with job seekers in Germany. IZA discussion paper No. 9040.

Altmann, S. and Traxler, C. (2014). Nudges at the dentist. *European Economic Review*, 72:19 – 38.

An, W. and Wan, X. (2016). R: Local average response functions for instrumental variable estimation of treatment effects. `https://cran.r-project.org/web/packages/LARF/` (accessed June 2017).

Angrist, J., Bettinger, E., and Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in colombia. *American Economic Review*, 96:847–862.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Epiricist's Companion*. Princeton University Press.

APM Database (2015). Agricultural policy measures database compiled for The Former Yugoslav Republic of Macedonia under the FAO/SWG project "Streamlining of agriculture and rural development policies of SEE countries for EU accession". Unpublished data.

Arellano, M. and Bonhomme, S. (2010). Quantile selection models. *Unpublished manuscript*.

Armantier, O. and Boly, A. (2011). A controlled field experiment on corruption. *European Economic Review*, 55:1072 – 1082.

Armantier, O. and Boly, A. (2013). Comparing corruption in the laboratory and in the field in Burkina Faso and in Canada. *The Economic Journal*, 123:1168–1187.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113:7353–7360.

Athey, S., Imbens, G., Kong, Y., and Ramachandra, V. (2016). An introduction to recursive partitioning for heterogeneous causal effects estimation using causalTree package. `https://github.com/susanathey/causalTree` (accessed June 2017).

Azmat, G. and Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics*, 30:32 – 40.

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182.

Barr, A. and Serra, D. (2010). Corruption and culture: An experimental analysis. *Journal of Public Economics*, 94:862 – 869.

Barsky, R., Bound, J., Charles, K., and Lupton, J. (2002). Accounting for the black-white wealth gap: A nonparametric approach. *Journal of the American Statistical Association*, 97:663–673.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81:608–650.

Benders, R. M., Kok, R., Moll, H. C., Wiersma, G., and Noorman, K. J. (2006). New approaches for household energy conservation—in search of personal household energy budgets and energy reduction options. *Energy Policy*, 34:3612 – 3622.

Bertrand, M. (2011). New perspectives on gender. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, pages 1543–1590. Elsevier.

Bičakova, A. (2014). Selection into labor force and gender unemployment gaps. *CERGE-EI Working Paper*, 513.

Blau, F. and Kahn, L. (2006). The US gender pay gap in the 1990s: Slowing convergence. *Industrial and Labor Relations Review*, 60:45–66.

Blinder, A. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8:436–455.

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71:655–679.

Brunello, G., Fort, M., Schneeweis, N., and Winter-Ebmer, R. (2016). The causal effect of education on health: What is the role of health behaviors? *Health Economics*, 25:314–336.

Bureau of Labor Statistics, U.S. Department of Labor (2001). National Longitudinal Survey of Youth 1979 cohort, 1979-2000 (rounds 1-19). Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH.

Busso, M., DiNardo, J., and McCrary, J. (2009). New evidence on the finite sample properties of propensity score matching and reweighting estimators. *IZA Discussion Paper No. 3998*.

Carroll, R., Ruppert, D., and Stefanski, L. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2009). Inference on counterfactual distributions. *CeMMAP working paper CWP09/09*.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, 126:1593–1660.

Chetty, R. and Saez, E. (2013). Teaching the tax code: Earnings responses to an experiment with EITC recipients. *American Economic Journal: Applied Economics*, 5:1–31.

Cobb-Clark, D. A. (2016). Biology and gender in the labor market. *IZA DP No. 10386*.

Conti, G., Heckman, J. J., and Pinto, R. (2016). The effects of two influential early childhood interventions on health and healthy behaviour. *The Economic Journal*, 126:F28–F65.

Corbacho, A., Gingerich, D. W., Oliveros, V., and Ruiz-Vega, M. (2016). Corruption as a self-fulfilling prophecy: Evidence from a survey experiment in Costa Rica. *American Journal of Political Science*, 60:1077–1092.

Das, M., Newey, W. K., and Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies*, 70:33–58.

Denisova-Schmidt, E. (2017). The challenges of academic integrity in higher education: Current trends and outlook. *CIHE Perspectives 5*. Boston: Boston College.

Denisova-Schmidt, E. (2019). Corruption in higher education. In Teixeira, Nuno, P., and Shin, J.-C., editors, *Encyclopedia of International Higher Education Systems and Institutions*. Springers.

Denisova-Schmidt, E. and de Wit, H. (2017). The global challenge of corruption in higher education. *IAU HORIZONS*, 22:28–29.

Denisova-Schmidt, E., Huber, M., and Leontyeva, E. (2016). Do anti-corruption educational campaigns reach students? Some evidence from Russia and Ukraine. *Educational Studies Moscow*, 1:61–83.

Denisova-Schmidt, E., Huber, M., and Prytula, Y. (2015). An experimental evaluation of an anti-corruption intervention among Ukrainian university students. *Eurasian Geography and Economics*, 56:713–734.

d'Haultfoeuille, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154:1–15.

Dimant, E. and Tosato, G. (2018). Causes and effects of corruption: What has past decade's research taught us? A survey. *Journal of Economic Surveys*, 32:335–356.

Dimitrievski, D., Kotevska, A., Janeska Stamenkovska, I., Tuna, E., and Nacka, M. (2014). Agriculture and agricultural policy in the Former Yugoslav Republic of Macedonia. In Volk, T., Erjavec, E., and Mortensen, K., editors, *Agricultural Policy and European Integration in Southeastern Europe*. Budapest: Food and Agriculture Organization of the United Nations.

DiNardo, J., Fortin, N., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64:1001–1044.

Dollar, D., Fisman, R., and Gatti, R. (2001). Are women really the "fairer" sex? Corruption and women in government. *Journal of Economic Behavior and Organization*, 46:423–429.

Duflo, E. and Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly Journal of Economics*, 118:815–842.

Duncan, O. D. (1967). Discrimination against negroes. *Annals of the American Academy of Political and Social Science*, 371:85–103.

Dwyer, J., Buckwell, A., Hart, K., Menadue, H., Mantino, F., Erjavec, E., and Ilbery, B. (2012). Study: How to improve the sustainable competitiveness and innovation of the agriculture sector, IP/ B/AGRI/IC/2011-100. Brussels: European Parliament. Available at `www.europarl.europa.eu/studies` (accessed August 15, 2017).

Dwyer, J. and Powell, J. (2016). Rural development programmes and transaction effects: Reflections on maltese and english experience. *Journal of Agricultural Economics*, 67:545–565.

European Commission (2005). Council Regulation (EC) No. 1698/2005 of 20 August 2005 on Support for Rural Development by the European Agricultural Fund for Rural Development (EAFRD). Brussels: Commission of the European Communities.

European Commission (2013). IPA Rural Development Programme (IPARD) for the Former Yugoslav Republic of Macedonia. Fifth Modification. Available at `http://ec.europa.eu/agriculture/enlargement/countries/fyrom/ipard_en.pdf` (accessed August 15, 2017).

Federal State Statistics Service (2016). Chislennost' naselenia Rossiyskoy Federatsiy po municipalnim obrazovaniyam [The population of the Russian Federation by municipalities]. Online bulletin, Federal State Statistics Service. http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/publications/catalog/afc8ea004d56a39ab251f2bafc3a6fce (accessed June 2017).

Feld, S., Frenzen, H., Krafft, M., Peters, K., and Verhoef, P. C. (2013). The effects of mailing design characteristics on direct mail campaign performance. *International Journal of Research in Marketing*, 30:143 – 159.

Ferraro, P. J. and Miranda, J. J. (2013). Heterogeneous treatment effects and mechanisms in information-based environmental policies: Evidence from a large-scale field experiment. *Resource and Energy Economics*, 35:356 – 379.

Findley, M., Nielson, D., and Sharman, J. (2014). *Global shell games*. Cambridge University Press, Cambridge.

Finn, J. D. and Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, 27:557–577.

Finn, J. D., Fulton, D., Zaharias, J., and Nye, B. A. (1989). Carry-over effects of small classes. *Peabody Journal of Education*, 67:75–84.

Firpo, S., Fortin, N. M., and Lemieux, T. (2007). Decomposing wage distributions using recentered influence functions regressions. *Mimeo, University of British Columbia*.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77:953–973.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). An analysis of sample attrition in panel data: The michigan panel study of income dynamics. *Journal of Human Resources*, 33:251–299.

Flores, C. A. and Flores-Lagunes, A. (2009). Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. *IZA DP No. 4237*.

Folger, J. and Breda, C. (1989). Evidence from Project STAR about class size and student achievement. *Peabody Journal of Education*, 67:17–33.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Chapter 1 - Decomposition methods in economics. volume 4, Part A of *Handbook of Labor Economics*, pages 1 – 102. Elsevier.

Frank, B., Lambsdorff, J. G., and Boehm, F. (2011). Gender and corruption: Lessons from laboratory corruption experiments. *The European Journal of Development Research*, 23:59–71.

Frölich, M. (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.

Galadima, M. (2014). Constraints on farmers' access to agricultural information delivery: Survey of rural farmers in Yobe State, Nigeria. *IOSR Journal of Agriculture and Veterinary Science*, 7:18–22.

Garbarino, E., Slonim, R., and Sydnor, J. (2011). Digit ratios (2d:4d) as predictors of risky decision making for both sexes. *Journal of Risk and Uncertainty*, 42:1–26.

García, J., Hernández, P. J., and López-Nicolás, A. (2001). How wide is the gap? An investigation of gender wage differences using quantile regression. *Empirical Economics*, 26:149–167.

Gershenson, S., Jacknowitz, A., and Brannegan, A. (2017). Are student absences worth the worry in U.S. primary schools? *Education Finance and Policy*, 12:137–165.

Goraus, K., Tyrowicz, J., and van der Velde, L. (2015). Which gender wage gap estimates to trust? A comparative analysis. *Review of Income and Wealth*, 63:118–146.

Gottfried, M. A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis*, 31:392–415.

Greiner, D. J. and Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *The Review of Economics and Statistics*, 93:775–785.

Gronau, R. (1974). Wage comparisons - a selectivity bias. *Journal of Political Economy*, 82:1119–1143.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20:25–46.

Hainmueller, J. and Xu, Y. (2013). ebalance: A Stata package for entropy balancing. *Journal of Statistical Software, Articles*, 54:1–18.

Hassan, S., Shaffril, H. A. M., Ali, M. S. S., and Ramli, N. S. (2010). Agriculture agency, mass media and farmers: combination for creating knowledgeable agriculture community. *African Journal of Agricultural Research*, 5:3500–3513.

Hausman, J. and Wise, D. (1979). Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica*, 47:455–473.

Heckman, J. (1976a). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Heckman, J. J. (1976b). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492.

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

Holmes, L. (2015). *Corruption: a very short introduction*. Oxford University Press, Oxford.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. In *Proceedings of the American Statistical Association, Biometrics Section*, page 2401–2415. Alexandria, VA: American Statistical Association.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite population. *Journal of American Statistical Association*, 47:663–685.

Hsu, Y., Huber, M., and Lai, T. (2018a). Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting. *Forthcoming in the Journal of Econometric Methods*.

Hsu, Y., Huber, M., Lee, Y., and Pipoz, L. (2018b). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *SES Working papers 495, University of Fribourg*.

Huber, M. (2012). Identification of average treatment effects in social experiments under alternative forms of attrition. *Journal of Educational and Behavioral Statistics*, 37:443–474.

Huber, M. (2014a). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29:920–943.

Huber, M. (2014b). Treatment evaluation in the presence of sample selection. *Econometric Reviews*, 33:869–905.

Huber, M. (2015). Causal pitfalls in the decomposition of wage gaps. *Journal of Business and Economic Statistics*, 33:179–191.

Huber, M., Denisova-Schmidt, E., Leontyeva, E., and Solovyeva, A. (2017). Combining experimental evidence with machine learning to assess anti-corruption educational campaigns among Russian university students. *SES Working paper 487, University of Fribourg.*

Huber, M., Kotevska, A., Stojcheska, A. M., and Solovyeva, A. (2018). Evaluating an information campaign about rural development policies in FYR Macedonia. *Agricultural and Resource Economics Review*, pages 1–25.

Huber, M., Lechner, M., and Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, 175:1–21.

Huber, M. and Mellace, G. (2014). Testing exclusion restrictions and additive separability in sample selection models. *Empirical Economics*, 47:75–92.

Huber, M. and Solovyeva, A. (2018a). Direct and indirect effects under sample selection and outcome attrition. *SES Working paper 496, University of Fribourg.*

Huber, M. and Solovyeva, A. (2018b). On the sensitivity of wage gap decompositions. *SES Working paper 497, University of Fribourg.*

Imai, K. (2009). Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment. *Journal of the Royal Statistical Society Series C*, 58:83–104.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25:51–71.

Imai, K. and Yamamoto, T. (2011). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Unpulished manuscript.*

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

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77:1481–1512.

IPARD II (2015). EU Instrument for Pre-Accession Rural Development Programme 2014—2020. Final version as adopted by the European Commission on 13.02.2015. Skopje, Republic of Macedonia. Available at `http://ipardpa.gov.mk/Root/mak/_docs/Zakonodavstvo/IPARD%20II%20Programme_ENG.pdf` (accessed November 20, 2017).

Jetter, M. and Walker, J. K. (2015). Good girl, bad boy: Corrupt behavior in professional tennis. Working paper, Center for Research in Economics and Finance (CIEF).

John, L. K., Loewenstein, G., and Rick, S. I. (2014). Cheating more for less: Upward social comparisons motivate the poorly compensated to cheat. *Organizational Behavior and Human Decision Processes*, 123:101 – 109.

Judd, C. M. and Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5:602–619.

Juhn, C., Murphy, K., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of Political Economy*, 101:410–442.

Kasamara, V. and Sorokina, A. (2017). Rebuilt empire or new collapse? Geopolitical visions of Russian students. *Europe-Asia Studies*, 69:262–283.

Klemenčič, M. (2014). Student power in a global perspective and contemporary trends in student organising. *Studies in Higher Education*, 39:396–411.

Korostelev, A., Romenskiy, V., and Sagieva, K. (Hosts) (2017). Progulka rasserzhennyh shkol'nikov: kak pokolenie YouTube vyshlo na ulicu i kak ego nakazhut [Angry pupils' walk: how the YouTube generation went out into the streets and how it will be punished]. In Pushkarev, V., Yapparova, L., Borzunova, M., Alexandrov, A., Zhelvnov, A., Ruzavin, P. et al., editor, *Zdes' i sejchas. Vechernee shou [Here and now. The evening show]*. Dozhd', Moscow, Russia. `https://tvrain.ru/teleshow/vechernee_shou/on_vam_ne_dimon-430761/` (accessed June 2017).

Kotevska, A., Bogdanov, N., Nikolic, A., Dimitrievski, D., Martinovska-Stojcheska, A., Tuna, E., Milic, T., Simonovska, A., Papic, R., Petrovic, L., Uzunovic, M., Becirovic, E., Angjelkovic, B., Gjoshevski, D., and Georgiev, N. (2015). The impact of socio-economic structure of rural population on success of Rural Development Policy — Macedonia, Serbia and Bosnia and Herzegovina. Skopje: Association of Agricultural Economists of Republic of Macedonia.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114:497–532.

Krueger, A. B. and Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, 111:1–28.

Kullback, J. (1959). *Information Theory and Statistics*. Wiley, New York.

Kunze, A. (2008). Gender wage gap studies: consistency and decomposition. *Empirical Economics*, 35:63–76.

Lemieux, T. (1998). Estimating the effects of unions on wage inequality in a panel data model with comparative advantage and nonrandom selection. *Journal of Labor Economics*, 16:261–291.

Liebman, J. B. and Luttmer, E. F. P. (2015). Would people behave differently if they better understood social security? Evidence from a field experiment. *American Economic Journal: Economic Policy*, 7:275 – 299.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121.

Ludwig, J., Mullainathan, S., and Spiess, J. (2017). Machine learning tests for effects on multiple outcomes. *Mimeo, Cornell University*.

Lwoga, E. T., Stilwell, C., and Ngulube, P. (2011). Access and use of agricultural information and knowledge in Tanzania. *Library Review*, 60:383–395.

Maasoumi, E. and Wang, L. (2016). The gender gap between earnings distributions. *Working paper, Emory University*.

Machado, C. (2017). Unobserved selection heterogeneity and the gender wage gap. *Journal of Applied Econometrics*, 32:1348–1366.

Machado, J. and Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20:445–465.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis.* Taylor and Francis, New York.

Manning, J. T., Reimers, S., Baron-Cohen, S., Wheelwright, S., and Fink, B. (2010). Sexually dimorphic traits (digit ratio, body height, systemizing–empathizing scores) and gender segregation between occupations: Evidence from the BBC internet study. *Personality and Individual Differences*, 49:511 – 515.

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human Resources*, 24:343–360.

Melly, B. (2005). Decomposition of differences in distribution using quantile regression. *Labour Economics*, 12:577–590.

Mora, R. (2008). A nonparametric decomposition of the Mexican American average wage gap. *Journal of Applied Econometrics*, 23:463–485.

Morrissey, T. W., Hutchison, L., and Winsler, A. (2014). Family income, school attendance, and academic achievement in elementary school. *Developmental Psychology*, 50:741–753.

Mulligan, C. B. and Rubinstein, Y. (2008). Selection, investment, and women's relative wages over time. *Quarterly Journal of Economics*, 123:1061–1110.

National Bank of the Republic of Macedonia (2017). List of exchange rates. Available at `http://www.nbrm.mk/kursna_lista-en.nspx` (accessed August 15, 2017).

Neuman, S. and Oaxaca, R. L. (2003). Gender versus ethnic wage differentials among professionals: Evidence from Israel. *Annales d'Économie et de Statistique*, 71/72:267–292.

Neuman, S. and Oaxaca, R. L. (2004). Wage decompositions with selectivity-corrected wage equations: A methodological note. *The Journal of Economic Inequality*, 2:3–10.

Newey, W., Powell, J., and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67:565–603.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14:201–206.

Newey, W. K. (2007). Nonparametric continuous/discrete choice models. *International Economic Review*, 48:1429–1439.

Ñopo, H. (2008). Matching as a tool to decompose wage gaps. *Review of Economics and Statistics*, 90:290–299.

Nye, B., Hedges, L. V., and Konstantopoulos, S. (2001). The long-term effects of small classes in early grades: Lasting benefits in mathematics achievement at grade 9. *The Journal of Experimental Education*, 69:245–257.

Nye, J. and Orel, E. (2015). The influence of prenatal hormones on occupational choice: 2d:4d evidence from Moscow. *Personality and Individual Differences*, 78:39 – 42.

Oaxaca, R. (1973). Male-female wage differences in urban labour markets. *International Economic Review*, 14:693–709.

Obrazovanie v Rossiiskoi Federatsii: 2014 [Education in the Russian Federation: 2014] (2014). Statistical compilation, National Research Institute "Higher School of Economics", Moscow, Russia.

Odongo, D. O., Wakhungu, W. J., and Stanley, O. (2017). Causes of variability in prevalence rates of communicable diseases among secondary school students in Kisumu County, Kenya. *Journal of Public Health*, 25:161–166.

Olivetti, C. and Petrongolo, B. (2008). Unequal pay or unequal employment? A cross-country analysis of gender gaps. *Journal of Labor Economics*, 26:621–654.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:669–710 (with discussion).

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, 17:276–284.

Ready, D. D. (2010). Socioeconomic disadvantage, school attendance, and early cognitive development: The differential effects of school exposure. *Sociology of Education*, 83:271–286.

Rivas, M. F. (2013). An experiment on corruption and gender. *Bulletin of Economic Research*, 65:10–42.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In Green, P., Hjort, N., and Richardson, S., editors, *In Highly Structured Stochastic Systems*, pages 70–81, Oxford. Oxford University Press.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155.

Robins, J. M. and Richardson, T. (2010). Alternative graphical causal models and the identification of direct effects. In Shrout, P., Keyes, K., and Omstein, K., editors, *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. Oxford University Press.

Robins, J. M., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 90:846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of American Statistical Association*, 90:106–121.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39:33–38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25:279–292.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31:161–170.

Saez, E. (2009). Details matter: The impact of presentation and information on the take-up of financial incentives for retirement saving. *American Economic Journal: Economic Policy*, 1:204 – 228.

Serra, D. and Wantchekon, L. (2012). *New advances in experimental research on corruption.* Emerald, Bingley, U.K.

Shah, A., Laird, N., and Schoenfeld, D. (1997). A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, 92:775–779.

Sloczynski, T. (2013). Population average gender effects. *IZA Discussion Paper No. 7315.*

Smith, J. and Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125:305–353.

Solovyeva, A. (2018). Student experience with academic dishonesty and corruption in the Khabarovsk Region of Russia. *Higher Education in Russia and Beyond (HERB)*, 3:9–11.

Spindler, M., Chernozhukov, V., and Hansen, C. (2016). R: High-dimensional metrics. `https://cran.r-project.org/web/packages/hdm/` (accessed June 2017).

State Statistical Office of the Republic of Macedonia (2015). Indicators. Available at `http://www.stat.gov.mk/Default_en.aspx` (accessed August 15, 2017).

Swamy, A., Knack, S., Lee, Y., and Azfar, O. (2001). Gender and corruption. *Journal of Development Economics*, 64:25 – 55.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *The Annals of Statistics*, 40:1816–1845.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2014). On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, 25:282–291.

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, 63:926–934.

Transparency International Russia (2015a). Episode 1: Bribe. YouTube video. `https://www.youtube.com/watch?v=zGeworhwEFo` (accessed July 2017).

Transparency International Russia (2015b). Episode 3: Corruption corporate raid. YouTube video. `https://www.youtube.com/watch?v=4aTjUyX67xc` (accessed July 2017).

van der Weele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20:18–26.

Vansteelandt, S., Bekaert, M., and Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, 1:129–158.

Volkov, D. (2017). Effekt ot filma «On vam ne Dimon» pochti proshel [The effect of the film "He is not Dimon to you" has almost passed]. *Gazeta.ru*. `https://www.gazeta.ru/comments/2017/05/25_a_10691315.shtml` (accessed June 2017).

Wooldridge, J. (2002). Inverse probability weigthed M-estimators for sample selection, attrition and stratification. *Portuguese Economic Journal*, 1:141–162.

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

Yamaguchi, K. (2014). Decomposition of gender or racial inequality with endogenous intervening covariates: An extension of the DiNardo-Fortin-Lemieux method. *RIETI Discussion Paper Series 14-E-061*.

Zakon za Zemjodelstvo i Ruralen Razvoj [Law of Agriculture and Rural Development] (2010). Sluzben Vesnik na Republika Makedonija 49: 1 – 84. Available at `http://www.slvesnik.com.mk/Issues/9E6050C54BDFFE46AD6BC1607D67D671.pdf` (accessed August 15, 2017).

Zheng, W. and van der Laan, M. J. (2012). Targeted maximum likelihood estimation of natural direct effects. *The International Journal of Biostatistics*, 8:1–40.

Appendices

Appendix 1

Table A1.1: Propensity score specification $\Pr(T = 1|X)$

Regressors	Coef.	s.e.	z-value	p-value
Age	-0.025	0.016	-1.520	0.129
Male (binary)	0.132	0.203	0.650	0.514
Education: high school (binary)	0.831	0.329	2.530	0.011
Education: college/ university (binary)	0.652	0.391	1.670	0.095
Education missing	1.151	0.449	2.560	0.010
Household head's occupation: agriculture (binary)	0.006	0.196	0.030	0.975
Years in farming	0.014	0.016	0.850	0.393
Household size	0.061	0.077	0.790	0.431
Profitable farm[a]	0.397	0.165	2.410	0.016
Subsidy dependent[b]	-0.066	0.111	-0.590	0.553
Frequency of cooperation[c]	-0.027	0.069	-0.400	0.692
Share of agricultural production sold on a market	-0.009	0.006	-1.450	0.147
Share of income from farming	-0.001	0.005	-0.240	0.810
Capacity: farmed area (ha)	-0.085	0.091	-0.930	0.351
Capacity: total livestock (number of heads)	-0.019	0.032	-0.600	0.552
Household head's occupation: missing (binary)	-0.342	0.660	-0.520	0.604
Constant	-0.153	1.029	-0.150	0.882

Notes: Standard deviations are in parentheses. Robust standard errors are in brackets. [a] *Profitable farm*: 1="very unprofitable"; 2="moderately unprofitable"; 3="break-even"; 4="moderately profitable"; 5="very profitable". [b] *Subsidy dependent*: 1="not dependent"; 2="slightly dependent"; 3="very dependent". [c] *Frequency of cooperation*: 1="never"; 2="rarely"; 3="not sure"; 4="sometimes"; 5="always".

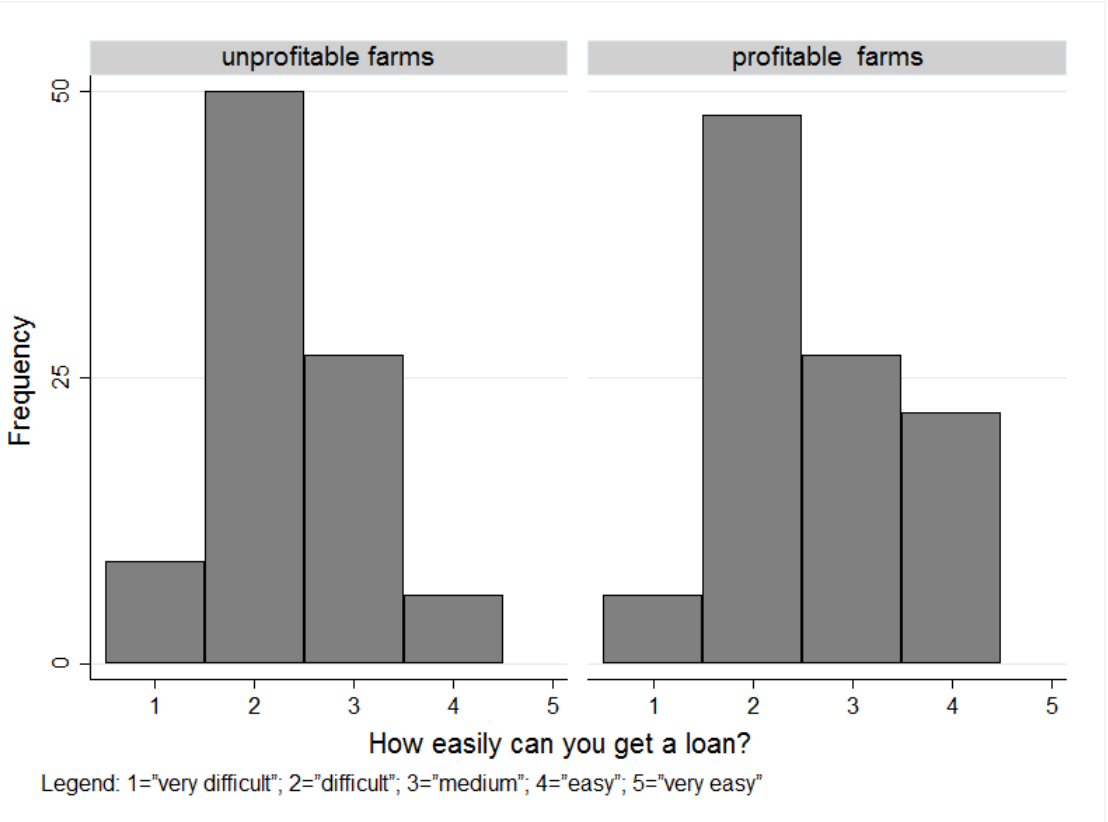

Figure A1.1: Ease of getting a loan by farm profitability

Source: The 2015 survey of farmers conducted by the authors.
Notes: The graph is based on the evaluation sample. Observations with missing information on profitability (1 obs.) and loan accessibility (62 obs.) are excluded, resulting in 195 observations (92 unprofitable farms and 103 profitable farms).

Appendix 2

Table A2.1: F-tests of covariate balance

Covariate variables	F-test	Prob $>F$
University 1	1.13	0.34
University 2	0.45	0.77
University 3	0.52	0.72
University 4	1.41	0.23
University 5	0.94	0.44
University 6	1.17	0.32
University 7	1.26	0.28
Major: humanities	0.27	0.90
Major: social sciences	0.53	0.72
Major: technical sciences	0.87	0.48
Major: natural sciences	1.14	0.34
Current academic year: bachelor	0.37	0.83
Current academic year: master	1.03	0.39
Current academic year: diploma	1.40	0.23
Reason for university education: to obtain good education	1.00	0.41
Reason for university education: hard to find job without education	0.32	0.87
Reason for university education: must have degree	1.35	0.25
Reason for university education: wanted to please parents	1.08	0.37
Reason for university education: everyone does that	0.81	0.52
Reason for university education: to delay army service	0.50	0.74
Academic performance (1=satisfactory... 5=excellent)	0.45	0.78
Presents to teachers at school (1=never... 5=systematically)	0.51	0.73
Paying fees at school (1=never... 5=systematically)	0.16	0.96
You/friends encountered any wrongdoing at USE	0.59	0.67
You/friends encountered any wrongdoing at univ.admission	0.90	0.46
Have you heard of your friends solving problems using connections?	0.34	0.85
Have you heard of your solved problems through bribery?	1.15	0.33
Female	0.43	0.79
University education is state financed	1.41	0.23
Place of residence before university: village or town	0.32	0.86

Continued on next page

Table A2.1 – continued from previous page

Covariate variables	F-test	Prob>F
Place of residence before university: city with population 2–250k	0.36	0.84
Place of residence before university: city with population 250–500k	0.44	0.78
Place of residence before university: city with population >500k	1.95	0.10
Age	0.19	0.95
Family status: both parents	2.52	0.04
Family status: only mother	1.59	0.18
Family status: only father	1.95	0.10
Family status: no parents	2.10	0.08
Number of siblings: 0	0.80	0.53
Number of siblings: 1	0.17	0.95
Number of siblings: 2	0.71	0.58
Number of siblings: 3 and more	0.17	0.95
Order of birth	0.75	0.56
Mother's education: secondary	1.19	0.31
Mother's education: higher	0.87	0.48
Mother's education: academic title	0.64	0.64
Father's education: secondary	0.49	0.74
Father's education: higher	0.97	0.42
Father's education: academic title	1.73	0.14
Mother's occupation: high level manager	1.14	0.34
Mother's occupation: middle level manager	1.69	0.15
Mother's occupation: highly qualified specialist	1.36	0.24
Mother's occupation: clerk	1.79	0.13
Mother's occupation: worker	0.79	0.53
Mother's occupation: entrepreneur	1.09	0.36
Mother's occupation: housewife or retiree	0.72	0.58
Mother's occupation: unemployed	0.61	0.66
Mother's occupation: military personnel	1.16	0.33
Father's occupation: high level manager	1.33	0.26
Father's occupation: middle level manager	0.91	0.46
Father's occupation: highly qualified specialist	0.69	0.60
Father's occupation: clerk	0.16	0.96
Father's occupation: worker	0.45	0.77
Father's occupation: entrepreneur	1.68	0.15
Father's occupation: househusband or retiree	1.99	0.09

Continued on next page

Table A2.1 – continued from previous page

Covariate variables	F-test	Prob>F
Father's occupation: unemployed	1.24	0.29
Father's occupation: military personnel	1.18	0.32
Financial situation (1=can only afford food... 5=can afford everything)	1.41	0.23
Monthly expenditures: <10k rub	0.98	0.42
Monthly expenditures: 10 – 20k rub	1.38	0.24
Monthly expenditures: >20k rub	1.27	0.28
Current accommodation: dormitory	0.77	0.55
Current accommodation: living with parents	1.19	0.31
Current accommodation: rent	0.75	0.56
Current accommodation: own an apartment	0.37	0.83
USE points: <150 points	0.25	0.91
USE points: 150 – 200 points	1.26	0.28
USE points: 200 – 250 points	1.62	0.17
USE points: >250 points	2.30	0.06
Student works	1.08	0.36
Employment related to education	2.11	0.08
Encountered bribery at university (1=never... 5=systematically)	0.02	1.00
How often do you use the following practices? (1=never... 5=systematically)		
Use crib sheets at exams	1.26	0.28
Submit papers downloaded from the internet	0.83	0.51
Buy papers	1.66	0.16
Write papers plagiarizing some chapters from the internet	1.97	0.10
Copy from other students during exams or tests	1.46	0.21
Deceive professors about study problems	1.06	0.38
Ask professors preferential treatment	1.09	0.36

Note: The F-tests test the equality of coefficients across the treatment groups in a regression of each individual characteristic on treatment indicators with heteroskedasticity robust standard errors.

Table A2.2: Estimates based on OLS with LASSO-selected covariates

Outcome	Official brochure			Tailored brochure			Video: bribery			Video: reiderstvo		
	Effect	se	p-v.	Effect	se	p-v.	Effect	se	p-v.	Effect	se	p-v.
How often do you think students use the following practices? (1=never... 5=systematically)												
Use crib sheets at exams	-0.04	0.06	0.51	-0.02	0.05	0.71	0.00	0.06	0.97	-0.04	0.06	0.45
Submit papers downloaded from the internet	-0.02	0.07	0.76	-0.02	0.06	0.75	0.12	0.07	0.08	-0.01	0.07	0.86
Buy papers	0.08	0.07	0.27	0.06	0.07	0.36	0.06	0.08	0.40	0.08	0.08	0.31
Write papers plagiarizing some chapters from the internet	0.06	0.06	0.32	0.15	0.06	0.01	0.05	0.07	0.42	0.02	0.07	0.71
Copy from other students during exams or tests	0.09	0.06	0.16	0.08	0.06	0.17	0.07	0.07	0.27	0.04	0.07	0.60
Deceive professors about study problems	-0.05	0.08	0.52	0.02	0.07	0.84	-0.03	0.08	0.72	0.05	0.08	0.55
Ask professors preferential treatment	-0.02	0.07	0.73	0.08	0.07	0.29	0.06	0.08	0.45	0.04	0.07	0.56
When do you think these practices are acceptable? (1=definitely no... 5=definitely yes)												
When useless course	0.19	0.08	0.03	-0.01	0.08	0.94	-0.06	0.08	0.45	-0.01	0.09	0.92
When students work	0.13	0.08	0.09	-0.01	0.08	0.85	0.02	0.08	0.84	-0.09	0.08	0.30
If hard to learn material	0.14	0.08	0.08	-0.03	0.08	0.69	0.01	0.08	0.92	-0.10	0.08	0.24
Always acceptable	0.04	0.07	0.56	-0.07	0.07	0.27	0.04	0.07	0.59	0.00	0.07	0.98
Never acceptable	-0.15	0.09	0.11	-0.16	0.09	0.07	-0.10	0.09	0.26	-0.02	0.10	0.82
What does corruption mean to you? (1= definitely no... 5= definitely yes)												
Necessity	0.08	0.07	0.25	0.00	0.07	0.98	0.03	0.07	0.72	0.02	0.07	0.77
Means of income	0.18	0.09	0.04	0.04	0.09	0.63	0.13	0.09	0.16	-0.08	0.09	0.36
Crime	0.01	0.07	0.94	-0.04	0.07	0.55	0.08	0.07	0.26	-0.01	0.07	0.86
Means to solve problems	0.05	0.08	0.50	0.04	0.08	0.58	-0.04	0.08	0.63	-0.13	0.09	0.14
Compensation for low salaries	0.13	0.08	0.12	0.09	0.08	0.30	-0.01	0.08	0.91	-0.04	0.09	0.61
Evil	0.00	0.08	0.98	0.01	0.08	0.89	0.00	0.08	0.97	-0.11	0.09	0.25
In your view, how does corruption affect...? (1=strictly negative... 5=fully positive)												
Your career opportunities	0.02	0.07	0.74	-0.04	0.07	0.52	-0.07	0.07	0.35	0.00	0.07	0.97
Your quality of life	0.02	0.07	0.76	-0.06	0.07	0.37	-0.04	0.07	0.61	-0.01	0.07	0.89
Your education	0.00	0.07	0.99	-0.03	0.07	0.64	-0.09	0.07	0.21	-0.04	0.07	0.59
Your health	0.03	0.07	0.69	-0.05	0.07	0.49	0.00	0.07	0.99	-0.09	0.07	0.24
Your safety	-0.04	0.07	0.57	-0.03	0.07	0.61	-0.09	0.07	0.19	-0.09	0.07	0.22
Russian economy	0.04	0.05	0.48	-0.03	0.05	0.54	0.00	0.05	0.95	0.06	0.06	0.27
Russian politics	0.01	0.05	0.83	-0.03	0.05	0.48	-0.02	0.05	0.66	0.02	0.06	0.72
Russian education	0.07	0.05	0.15	0.00	0.05	0.97	-0.01	0.05	0.88	0.03	0.06	0.56
Russian health system	0.12	0.06	0.03	0.00	0.05	0.96	0.03	0.06	0.59	0.05	0.06	0.41
Russian police	0.09	0.06	0.10	-0.02	0.05	0.65	0.01	0.05	0.89	0.06	0.06	0.33
Can corruption be eradicated in Russia? (1=definitely no... 5=definitely yes)	-0.03	0.07	0.65	-0.09	0.07	0.20	-0.10	0.07	0.16	0.02	0.07	0.79
Take part in roundtable? (0=no, 1=yes)	-0.01	0.01	0.39	-0.01	0.01	0.31	-0.02	0.01	0.07	0.00	0.01	0.82
Take part in survey next year?(0=no, 1=yes)	-0.03	0.02	0.09	-0.03	0.02	0.12	-0.03	0.02	0.09	-0.02	0.02	0.36

Notes: 'Effect' represents the estimate from an OLS regression of an outcome variable on a set of regressors selected in the post-double-selection LASSO procedure, 'se' provides asymptotic standard error, and 'p-v.' stands for p-value.

Table A2.3: Multiple outcomes test: Subsample based on gender

Question group	Official brochure	Tailored brochure	Video: bribery	Video: *reiderstvo*
Panel A: Male students				
How often do you think students use the following [corrupt] practices?	0.52	0.93	0.93	0.07
When do you think these [corrupt] practices are acceptable?	0.03	0.63	0.87	0.13
What does corruption mean to you?	0.19	0.28	0.56	0.60
In your view, how does corruption affect aspects of your life?	0.30	0.90	0.95	0.18
In your view, how does corruption affect public spheres in Russia?	0.02	0.34	0.73	0.06
Interest in anti-corruption activities	0.88	0.85	0.90	0.43
Panel B: Female students				
How often do you think students use the following [corrupt] practices?	0.74	0.55	0.13	0.21
When do you think these [corrupt] practices are acceptable?	0.97	0.49	0.16	0.28
What does corruption mean to you?	0.59	0.39	0.13	0.17
In your view, how does corruption affect aspects of your life?	0.89	0.45	0.15	0.25
In your view, how does corruption affect public spheres in Russia?	0.52	0.61	0.17	0.33
Interest in anti-corruption activities	0.53	0.65	0.15	0.39

Note: The p-values of the joint significance tests are presented.

Table A2.4: Effects in the male subsample

Outcome	Control mean	Official brochure			Tailored brochure			Video: bribery			Video: *reiderstvo*		
		Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.
How often do you think students use the following practices? (1=never... 5=systematically)													
Use crib sheets at exams	3.86	-0.01	0.09	0.94	0.04	0.09	0.67	0.03	0.10	0.77	0.08	0.10	0.46
Submit papers downloaded from the internet	3.54	0.03	0.10	0.77	-0.05	0.10	0.63	0.07	0.11	0.52	0.03	0.12	0.80
Buy papers	3.29	0.03	0.11	0.82	-0.02	0.11	0.87	0.15	0.11	0.16	-0.12	0.12	0.35
Write papers plagiarizing some chapters from the internet	3.80	0.04	0.10	0.73	0.14	0.10	0.15	0.12	0.10	0.23	-0.01	0.11	0.94
Copy from other students during exams or tests	3.71	0.18	0.10	0.09	0.13	0.10	0.20	0.13	0.10	0.20	0.05	0.12	0.66
Deceive professors about study problems	3.15	-0.07	0.12	0.55	0.05	0.12	0.68	0.05	0.12	0.71	0.00	0.12	0.98
Ask professors preferential treatment	2.58	-0.01	0.11	0.96	0.10	0.12	0.40	0.16	0.12	0.18	-0.02	0.11	0.85
When do you think these practices are acceptable? (1= definitely no... 5= definitely yes)													
When a course is useless	2.61	0.39	0.13	0.00	0.15	0.12	0.22	-0.02	0.13	0.88	-0.05	0.14	0.70
When students work	3.01	0.24	0.12	0.04	-0.02	0.12	0.88	-0.06	0.13	0.64	-0.13	0.14	0.35
If hard to learn material	2.69	0.18	0.12	0.13	0.00	0.11	1.00	-0.01	0.12	0.93	-0.11	0.13	0.41
Always acceptable	2.19	0.10	0.12	0.42	-0.03	0.11	0.76	0.10	0.11	0.39	-0.08	0.12	0.50
Never acceptable	3.10	-0.07	0.13	0.60	-0.20	0.13	0.11	-0.21	0.13	0.12	-0.06	0.15	0.68
What does corruption mean to you? (1= definitely no... 5= definitely yes)													
Necessity	1.99	0.12	0.11	0.28	-0.02	0.10	0.89	-0.01	0.11	0.96	-0.07	0.12	0.54
Means of income	2.93	0.32	0.14	0.02	0.08	0.13	0.56	0.14	0.13	0.31	-0.32	0.15	0.03
Crime	4.01	0.13	0.11	0.22	0.02	0.11	0.85	0.05	0.12	0.65	0.03	0.11	0.78
Means to solve problems	3.21	0.06	0.12	0.59	-0.05	0.12	0.71	0.02	0.13	0.86	-0.11	0.13	0.39
Compensation for low salaries	2.67	0.14	0.13	0.29	0.13	0.13	0.31	-0.11	0.13	0.41	-0.07	0.14	0.63
Evil	3.83	0.09	0.12	0.43	-0.01	0.12	0.96	0.14	0.12	0.23	-0.23	0.14	0.11
In your view, how does corruption affect...? (1=strictly negative... 5=fully positive)													
Your career opportunities	2.37	0.06	0.11	0.58	-0.08	0.10	0.43	-0.10	0.11	0.39	0.10	0.11	0.39
Your quality of life	2.45	-0.01	0.10	0.95	-0.12	0.10	0.26	-0.09	0.11	0.44	-0.03	0.11	0.79
Your education	2.33	-0.06	0.10	0.54	-0.08	0.10	0.46	-0.23	0.10	0.03	-0.06	0.11	0.59
Your health	2.36	0.01	0.11	0.93	-0.14	0.11	0.20	-0.04	0.11	0.76	-0.10	0.11	0.39
Your safety	2.21	-0.04	0.11	0.71	-0.09	0.11	0.43	-0.13	0.11	0.22	-0.14	0.12	0.25
Russian economy	1.52	0.13	0.08	0.12	0.00	0.07	0.95	0.05	0.08	0.57	0.10	0.09	0.29
Russian politics	1.56	0.11	0.08	0.18	0.02	0.08	0.85	0.03	0.08	0.73	0.12	0.09	0.18
Russian education	1.59	0.14	0.08	0.09	0.02	0.08	0.81	-0.03	0.08	0.69	0.10	0.09	0.26
Russian health system	1.56	0.26	0.10	0.01	0.03	0.08	0.70	0.06	0.09	0.51	0.13	0.10	0.17
Russian police	1.48	0.18	0.09	0.05	-0.02	0.08	0.82	0.02	0.09	0.87	0.16	0.10	0.12
Can corruption be eradicated in Russia? (1=definitely no... 5=definitely yes)	2.47	0.01	0.11	0.95	-0.17	0.11	0.11	-0.05	0.11	0.64	0.12	0.12	0.32
Take part in roundtable? (0=no, 1=yes)	0.06	-0.04	0.02	0.06	-0.03	0.02	0.07	-0.03	0.02	0.08	-0.02	0.02	0.32
Take part in survey next year? (0=no, 1=yes)	0.14	-0.05	0.03	0.10	-0.05	0.03	0.06	-0.06	0.03	0.05	-0.05	0.03	0.16

Notes: 'Effect' represents the difference between the mean outcome value in each treatment group and the control mean, 'se' provides asymptotic standard error robust to heteroskedasticity, and '*p*-v.' stands for *p*-value.

Table A2.5: Effects in the female subsample

Outcome	Control mean	Official brochure			Tailored brochure			Video: bribery			Video: *reiderstvo*		
		Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.	Effect	se	*p*-v.
How often do you think students use the following practices? (1=never... 5=systematically)													
Use crib sheets at exams	3.99	-0.04	0.08	0.67	-0.07	0.08	0.34	-0.01	0.09	0.92	-0.10	0.08	0.22
Submit papers downloaded from the internet	3.45	0.01	0.10	0.90	0.03	0.09	0.74	0.17	0.10	0.08	-0.01	0.10	0.95
Buy papers	3.14	0.18	0.10	0.08	0.14	0.10	0.17	0.02	0.11	0.85	0.24	0.11	0.02
Write papers plagiarizing some chapters from the internet	3.72	0.13	0.09	0.14	0.21	0.08	0.01	0.06	0.09	0.54	0.16	0.09	0.09
Copy from other students during exams or tests	3.77	0.07	0.09	0.42	0.06	0.09	0.51	0.08	0.09	0.38	0.05	0.10	0.60
Deceive professors about study problems	3.06	-0.01	0.11	0.95	0.07	0.11	0.53	-0.05	0.11	0.65	0.12	0.11	0.28
Ask professors preferential treatment	2.39	-0.07	0.10	0.51	0.13	0.10	0.22	-0.01	0.11	0.89	0.10	0.11	0.37
When do you think these practices are acceptable? (1= definitely no... 5= definitely yes)													
When a course is useless	2.64	0.18	0.12	0.12	-0.08	0.11	0.48	-0.01	0.11	0.95	0.08	0.12	0.52
When students work	2.95	0.09	0.11	0.40	0.03	0.11	0.76	0.09	0.11	0.43	-0.05	0.11	0.65
If it is hard to learn material	2.73	0.16	0.11	0.14	-0.02	0.11	0.84	0.09	0.11	0.41	-0.08	0.11	0.45
Always acceptable	2.04	0.09	0.09	0.34	-0.01	0.09	0.91	0.08	0.10	0.42	0.07	0.10	0.47
Never acceptable	2.92	-0.17	0.12	0.16	-0.12	0.12	0.31	0.02	0.13	0.90	0.04	0.13	0.76
What does corruption mean to you? (1= definitely no... 5= definitely yes)													
Necessity	1.87	0.08	0.10	0.40	0.06	0.10	0.54	0.08	0.10	0.43	0.09	0.10	0.35
Means of income	2.78	0.10	0.12	0.42	0.01	0.12	0.92	0.11	0.13	0.36	0.14	0.12	0.28
Crime	4.14	-0.14	0.10	0.17	-0.13	0.10	0.17	0.06	0.10	0.53	-0.05	0.10	0.65
Means to solve problems	2.96	0.09	0.10	0.37	0.13	0.11	0.23	-0.07	0.12	0.53	-0.08	0.12	0.53
Compensation for low salaries	2.52	0.14	0.12	0.25	0.05	0.11	0.64	0.07	0.12	0.59	0.01	0.12	0.91
Evil	3.83	-0.08	0.12	0.53	0.02	0.11	0.87	-0.14	0.12	0.27	-0.02	0.13	0.86
In your view, how does corruption affect...? (1=strictly negative... 5=fully positive)													
Your career opportunities	2.30	0.02	0.10	0.85	-0.02	0.09	0.84	-0.02	0.10	0.86	-0.04	0.10	0.68
Your quality of life	2.33	0.06	0.10	0.55	-0.04	0.09	0.69	0.00	0.10	0.96	0.02	0.10	0.89
Your education	2.13	0.05	0.09	0.57	0.03	0.09	0.77	0.03	0.10	0.79	0.01	0.10	0.92
Your health	2.21	0.04	0.10	0.70	0.02	0.09	0.83	0.00	0.10	0.99	-0.09	0.10	0.37
Your safety	1.98	-0.03	0.09	0.74	0.01	0.09	0.90	-0.05	0.09	0.56	-0.05	0.09	0.59
Russian economy	1.51	0.00	0.07	0.95	-0.02	0.06	0.77	-0.02	0.07	0.81	0.05	0.08	0.54
Russian politics	1.58	-0.04	0.07	0.60	-0.03	0.06	0.64	-0.04	0.07	0.54	-0.05	0.08	0.55
Russian education	1.52	0.03	0.07	0.64	0.01	0.06	0.82	0.01	0.07	0.90	0.00	0.07	0.98
Russian health system	1.52	0.04	0.07	0.62	-0.01	0.07	0.94	0.01	0.07	0.90	-0.01	0.07	0.88
Russian police	1.41	0.04	0.07	0.61	-0.01	0.06	0.93	0.01	0.07	0.86	0.00	0.07	0.98
Can corruption be eradicated in Russia? (1=definitely no... 5=definitely yes)	2.56	-0.12	0.09	0.21	-0.04	0.09	0.67	-0.17	0.10	0.09	-0.10	0.10	0.31
Take part in roundtable? (0=no, 1=yes)	0.03	0.00	0.02	0.86	0.00	0.02	0.96	-0.02	0.01	0.25	0.01	0.02	0.47
Take part in survey next year? (0=no, 1=yes)	0.11	-0.02	0.03	0.40	-0.01	0.03	0.67	-0.02	0.03	0.53	0.00	0.03	0.91

Notes: 'Effect' represents the difference between the mean outcome value in each treatment group and the control mean, 'se' provides asymptotic standard error robust to heteroskedasticity, and '*p*-v.' stands for *p*-value.

Appendix 3

3.0.1 Proof of Theorem 1

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|M,X)\cdot \Pr(S=1|D,M,X)} \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right] && (3.1)\\
= & \underset{X}{E}\left[\underset{M|X=x}{E}\left[E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|M,X)\cdot \Pr(S=1|D,M,X)}\middle| M=m, X=x\right] \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right]\right] \\
= & \underset{X}{E}\left[\underset{M|X=x}{E}\left[E\left[\frac{Y \cdot S}{\Pr(S=1|D,M,X)}\middle| D=d, M=m, X=x\right] \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right]\right] \\
= & \underset{X}{E}\left[\underset{M|X=x}{E}\left[E\left[Y|D=d, M=m, X=x, S=1\right] \cdot \frac{\Pr(D=1-d|M,X)}{\Pr(D=1-d|X)}\right]\right] \\
= & \underset{X}{E}\left[\underset{M|D=1-d,X=x}{E}\left[E\left[Y|D=d, M=m, X=x, S=1\right]\right]\right] \\
= & \underset{X}{E}\left[\underset{M|D=1-d,X=x}{E}\left[E\left[Y|D=d, M=m, X=x\right]\right]\right] \\
= & \underset{X}{E}\left[\underset{M|D=1-d,X=x}{E}\left[E\left[Y(d,m)|D=d, M=m, X=x\right]\right]\right] \\
= & \underset{X}{E}\left[\underset{M|D=1-d,X=x}{E}\left[E\left[Y(d,m)|D=d, X=x\right]\right]\right] \\
= & \underset{X}{E}\left[\underset{M(1-d)|X=x}{E}\left[E\left[Y(d,m)|D=1-d, X=x\right]\right]\right] \\
= & \underset{X}{E}\left[\underset{M(1-d)|X=x}{E}\left[E\left[Y(d,m)|D=1-d, M(1-d)=m, X=x\right]\right]\right] \\
= & \underset{X}{E}\left[\underset{M(1-d)|X=x}{E}\left[E\left[Y(d,m)|M(1-d)=m, X=x\right]\right]\right] \\
= & \underset{X}{E}\left[\underset{M(1-d)|X=x}{E}\left[E\left[Y(d,M(1-d))|X=x\right]\right]\right] = E[Y(d,M(1-d))].
\end{aligned}
$$

Note that $\underset{A|B=b}{E}[C]$ denotes the expectation of C taken over the distribution of A conditional on $B=b$. The first equality follows from the law of iterated expectations, the second and third from basic probability theory, the fourth from Bayes' theorem, the fifth from Assumption 3, the sixth from the observational rule (implying for instance that Y given $D=d$ and $M=m$ is $Y(d,m)$), the seventh from Assumption 2, the eighth from Assumption 1, the ninth from Assumption 2, the tenth from Assumption 1, which implies that $Y(d,m) \perp D | M(1-d)=m, X=x$, and the last

from the law of iterated expectations.

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|X) \cdot \Pr(S=1|D,M,X)}\right] \\
= & \underset{X}{E}\left[E\left[\frac{Y \cdot I\{D=d\} \cdot S}{\Pr(D=d|X) \cdot \Pr(S=1|D,M,X)}\middle| X=x\right]\right] \\
= & \underset{X}{E}\left[E\left[\frac{Y \cdot S}{\Pr(S=1|D,M,X)}\middle| D=d, X=x\right]\right] \\
= & \underset{X}{E}\left[\underset{M|D=d,X=x}{E}\left[\frac{E[Y \cdot S|D=d,M=m,X=x]}{\Pr(S=1|D,M,X)}\middle| D=d, X=x\right]\right] \\
= & \underset{X}{E}\left[\underset{M|D=d,X=x}{E}[E[Y|D=d,M=m,X=x,S=1]|D=d,X=x]\right] \\
= & \underset{X}{E}\left[\underset{M|D=d,X=x}{E}[E[Y|D=d,M=m,X=x]|D=d,X=x]\right] \\
= & \underset{X}{E}[E[Y|D=d,X=x]] \\
= & \underset{X}{E}[E[Y(d,M(d))|D=d,X=x]] \\
= & \underset{X}{E}[E[Y(d,M(d))|X=x]] = E[Y(d,M(d))].
\end{aligned}
\tag{3.2}
$$

The first, third, sixth, and ninth equalities follow from the law of iterated expectations, the second and fourth from basic probability theory, the fifth from Assumption 3, the seventh from the observational rule, and the eighth from Assumption 1.

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\} \cdot S}{\Pr(D=d|X) \cdot \Pr(M=m|D,X) \cdot \Pr(S=1|D,M,X)}\right] \\
= & \underset{X}{E}\left[E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\} \cdot S}{\Pr(D=d|X) \cdot \Pr(M=m|D,X) \cdot \Pr(S=1|D,M,X)}\middle| X=x\right]\right] \\
= & \underset{X}{E}[E[Y|D=d,M=m,X=x,S=1]] \\
= & \underset{X}{E}[E[Y|D=d,M=m,X=x]] \\
= & \underset{X}{E}[E[Y(d,m)|D=d,M=m,X=x]] \\
= & \underset{X}{E}[E[Y(d,m)|D=d,X=x]] \\
= & \underset{X}{E}[E[Y(d,m)|X=x]] = E[Y(d,m)]
\end{aligned}
\tag{3.3}
$$

The first and seventh equalities follow from the law of iterated expectations, the second from basic probability theory, the third from Assumption 3, the fourth from the observational rule, the fifth from Assumption 2, and the sixth from Assumption 1.

3.0.2 Proof of Theorem 2

$$
\begin{aligned}
& E\left[\frac{Y\cdot I\{D=d\}}{\Pr(D=d|M,X,p(W))}\cdot\frac{\Pr(D=1-d|M,X,p(W))}{\Pr(D=1-d|X,p(W))}\middle|S=1\right] \qquad (3.4)\\
= & \underset{X,p(W)|S=1}{E}\left[\underset{M|X=x,p(W)=p(w),S=1}{E}\left[E\left[\frac{Y\cdot I\{D=d\}}{\Pr(D=d|M,X,p(W))}\middle|M=m,X=x,p(W)=p(w),S=1\right]\right.\right.\\
\times & \left.\left.\frac{\Pr(D=1-d|M,X,p(W)}{\Pr(D=1-d|X,p(W))}\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[\underset{M|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y|D=d,M=m,X=x,p(W)=p(w),S=1\right]\right.\right.\\
\times & \left.\left.\frac{\Pr(D=1-d|M,X,p(W))}{\Pr(D=1-d|X,p(W))}\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[\underset{M|D=1-d,X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|D=d,M=m,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[\underset{M|D=1-d,X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|D=d,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[\underset{M(1-d)|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|D=1-d,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[\underset{M(1-d)|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|D=1-d,M(1-d)=m,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[\underset{M(1-d)|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,m)|M(1-d)=m,X=x,p(W)=p(w),S=1\right]\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[\underset{M(1-d)|X=x,p(W)=p(w),S=1}{E}\left[E\left[Y(d,M(1-d))|X=x,p(W)=p(w),S=1\right]\right]\right]\\
= & E[Y(d,M(1-d))|S=1].
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the second from basic probability theory, the third from Bayes' theorem and the observational rule, the fourth from Assumptions 2 and 5 (which imply $Y(d,m)\perp M|D=d',X=x,p(W)=p(w),S=1$), the fifth from Assumptions 1 and 5 (which imply $\{Y(d,m),M(1-d)\}\perp D|X=x,p(W)=p(w),S=1$), the sixth from Assumptions 2 and 5, the seventh from Assumptions 1 and 5 (which imply $Y(d,m)\perp D|M(1-d)=m,X=x,p(W)=p(w),S=1$), and the last from the law of iterated expectations.

$$
\begin{aligned}
& E\left[\frac{Y\cdot I\{D=d\}}{\Pr(D=d|X,p(W))}\middle|S=1\right] \qquad (3.5)\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[\frac{Y\cdot I\{D=d\}}{\Pr(D=d|X,p(W))}\middle|X=x,p(W)=p(w),S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y|D=d,X=x,p(W)=p(w),S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,M(d))|D=d,X=x,p(W)=p(w),S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,M(d))|X=x,p(W)=p(w),S=1\right]\right]=E[Y(d,M(d))|S=1].
\end{aligned}
$$

The first and last equalities follow from the law of iterated expectations, the second from basic probability theory, the third from the observational rule, and the fourth from Assumptions 1 and 5 (which imply

$Y(d,m) \perp D | X = x, p(W) = p(w), S = 1)$.

$$
\begin{aligned}
& E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\}}{\Pr(D=d|X,p(W)) \cdot \Pr(M=m|D,X,p(W))}\middle| S=1\right] \qquad (3.6)\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\}}{\Pr(D=d|X,p(W)) \cdot \Pr(M=m|D,X,p(W))}\middle| X=x, p(W)=p(w), S=1\right]\middle| S=1\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y|D=d, M=m, X=x, p(W)=p(w), S=1\right]\middle| S=1\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,m)|D=d, M=m, X=x, p(W)=p(w), S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,m)|D=d, X=x, p(W)=p(w), S=1\right]\right]\\
= & \underset{X,p(W)|S=1}{E}\left[E\left[Y(d,m)|X=x, p(W)=p(w), S=1\right]\right] = E[Y(d,m)|S=1]
\end{aligned}
$$

The first and sixth equalities follow from the law of iterated expectations, the second from basic probability theory, the third from the observational rule, the fourth from Assumptions 2 and 5 (which imply $Y(d,m) \perp M | D = d, X = x, p(W) = p(w), S = 1$), and the fifth from Assumptions 1 and 5 (which imply $Y(d,m) \perp D | X = x, p(W) = p(w), S = 1$).

3.0.3 Proof of Theorem 3

$$
\begin{aligned}
& E\left[\left(\frac{Y\cdot D}{\Pr(D=1|M,X,p(W))}-\frac{Y\cdot(1-D)}{1-\Pr(D=1|M,X,p(W))}\right)\cdot\frac{\Pr(D=d|M,X,p(W))\cdot S}{\Pr(D=d|X,p(W))\cdot p(W)}\right] \qquad (3.7)\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}\left[E\left[\frac{Y\cdot D\cdot S}{\Pr(D=1|M,X,p(W))\cdot p(W)}\right.\right.\right.\\
- & \left.\left.\left.\frac{Y\cdot(1-D)\cdot S}{1-\Pr(D=1|M,X,p(W))\cdot p(W)}\right|M=m,X=x,p(W)=p(w)\right]\cdot\frac{\Pr(D=d|M,X,p(W))}{\Pr(D=d|X,p(W))}\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}\left[E\left[\frac{Y\cdot S}{p(W)}\middle|D=1,M=m,X=x,p(W)=p(w)\right]\right.\right.\\
- & \left.\left.E\left[\frac{Y\cdot S}{p(W)}\middle|D=0,M=m,X=x,p(W)=p(w)\right]\cdot\frac{\Pr(D=d|M,X,p(W))}{\Pr(D=d|X,p(W))}\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}[E[Y|D=1,M=m,X=x,p(W)=p(w),S=1]\right.\\
- & \left.E[Y|D=0,M=m,X=x,p(W)=p(w),S=1]\cdot\frac{\Pr(D=d|M,X,p(W))}{\Pr(D=d|X,p(W))}\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|D=d,X=x,p(W)=p(w)}{E}[E[Y(1,m)|D=1,M=m,X=x,p(W)=p(w),S=1]\right.\\
- & \left.E[Y(0,m)|D=0,M=m,X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|D=d,X=x,p(W)=p(w)}{E}[E[Y(1,m)|D=1,X=x,p(W)=p(w),S=1]\right.\\
- & \left.E[Y(0,m)|D=0,X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M(d)|X=x,p(W)=p(w)}{E}[E[Y(1,m)-Y(0,m)|X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M(d)|X=x,p(W)=p(w)}{E}[E[Y(1,m)-Y(0,m)|X=x,p(W)=p(w)]]\right]=\theta(d)
\end{aligned}
$$

The first and last equalities follow from the law of iterated expectations, the second from basic probability theory, the third from basic probability theory and the fact that $\Pr(S=1|D,M,X,p(W))=\Pr(S=1|D,M,X,Z)=p(W)$ (as $p(W)$ is a deterministic function of Z conditional on D,M,X), the fourth from Bayes' theorem and the observational rule, the fifth from Assumptions 2 and 5 (which imply $Y(d,m)\perp M|D=d',X=x,p(W)=p(w),S=1$), the sixth from Assumptions 1 and 5 (which imply $\{Y(d,m),M(d')\}\perp D|X=x,p(W)=p(w),S=1$), and the seventh from Assumption 7 by acknowledging that $p(W)=F_V$.

$$
\begin{aligned}
& E\left[\frac{Y\cdot I\{D=d\}\cdot S}{\Pr(D=d|M,X,p(W))\cdot p(W)}\cdot\left(\frac{\Pr(D=1|M,X,p(W))}{\Pr(D=1|X,p(W))}-\frac{1-\Pr(D=1|M,X,p(W))}{1-\Pr(D=1|X,p(W))}\right)\right] \qquad (3.8)\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}\left[E\left[\frac{Y\cdot I\{D=d\}\cdot S}{\Pr(D=d|M,X,p(W))\cdot p(W)}\middle| M=m,X=x,p(W)=p(w)\right]\right.\right.\\
\times & \left.\left.\left(\frac{\Pr(D=1|M,X,p(W))}{\Pr(D=1|X,p(W))}-\frac{1-\Pr(D=1|M,X,p(W))}{1-\Pr(D=1|X,p(W))}\right)\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|X=x,p(W)=p(w)}{E}\left[E[Y|D=d,M=m,X=x,p(W)=p(w),S=1]\cdot\left(\frac{\Pr(D=1|M,X,p(W))}{\Pr(D=1|X,p(W))}-\frac{1-\Pr(D=1|M,X,p(W))}{1-\Pr(D=1|X,p(W))}\right)\right]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|D=1,X=x,p(W)=p(w)}{E}[E[Y(d,m)|D=d,M=m,X=x,p(W)=p(w),S=1]]\right.\\
- & \left.\underset{M|D=0,X=x,p(W)=p(w)}{E}[E[Y(d,m)|D=d,M=m,X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M|D=1,X=x,p(W)=p(w)}{E}[E[Y(d,m)|D=d,X=x,p(W)=p(w),S=1]]\right.\\
- & \left.\underset{M|D=0,X=x,p(W)=p(w)}{E}[E[Y(d,m)|D=d,X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}\left[\underset{M(1)|X=x,p(W)=p(w)}{E}[E[Y(d,m)|X=x,p(W)=p(w),S=1]]-\underset{M(0)|X=x,p(W)=p(w)}{E}[E[Y(d,m)|X=x,p(W)=p(w),S=1]]\right]\\
= & \underset{X,p(W)}{E}[E[Y(d,M(1))-Y(d,M(0))|X=x,p(W)=p(w)]]=\delta(d)
\end{aligned}
$$

The first and last equalities follow from the law of iterated expectations, the second from basic probability theory and the fact that $\Pr(S=1|D,M,X,p(W))=\Pr(S=1|D,M,X,Z)=p(W)$, the third from Bayes' theorem and the observational rule, the fourth from Assumptions 2 and 5 (which imply $Y(d,m)\perp M|D=d',X=x,p(W)=p(w),S=1$), the fifth from Assumptions 1 and 5 (which imply $\{Y(d,m),M(d')\}\perp D|X=x,p(W)=p(w),S=1$), and the sixth from Assumption 7 by acknowledging that $p(W)=F_V$.

$$
\begin{aligned}
& E\left[\left(\frac{Y\cdot D}{\Pr(D=1|X,p(W))}-\frac{Y\cdot(1-D)}{1-\Pr(D=1|X,p(W))}\right)\cdot\frac{I\{M=m\}\cdot S}{\Pr(M=m|D,X,p(W))\cdot p(W)}\right] \qquad (3.9)\\
= & \underset{X,p(W)}{E}\left[E\left[\left(\frac{Y\cdot D}{\Pr(D=1|X,p(W))}-\frac{Y\cdot(1-D)}{1-\Pr(D=1|X,p(W))}\right)\cdot\frac{I\{M=m\}\cdot S}{\Pr(M=m|D,X,p(W))\cdot p(W)}\middle|X=x,p(W)=p(w)\right]\right]\\
= & \underset{X,p(W)}{E}\left[E[Y|D=1,M=m,X=x,p(W)=p(w),S=1]-E[Y|D=0,M=m,X=x,p(W)=p(w),S=1]\right]\\
= & \underset{X,p(W)}{E}\left[E[Y(1,m)|D=1,M=m,X=x,p(W)=p(w),S=1]-E[Y(0,m)|D=0,M=m,X=x,p(W)=p(w),S=1]\right]\\
= & \underset{X,p(W)}{E}\left[E[Y(1,m)|D=1,X=x,p(W)=p(w),S=1]-E[Y(0,m)|D=0,X=x,p(W)=p(w),S=1]\right]\\
= & \underset{X,p(W)}{E}\left[E[Y(1,m)-Y(0,m)|X=x,p(W)=p(w),S=1]\right]\\
= & \underset{X,p(W)}{E}\left[E[Y(1,m)-Y(0,m)|X=x,p(W)=p(w)]\right]=\gamma(m)
\end{aligned}
$$

The first and last equalities follow from the law of iterated expectations, the second from basic probability theory and the fact that $\Pr(S=1|D,M,X,p(W))=\Pr(S=1|D,M,X,Z)=p(W)$, the third from the observational rule, the fourth from Assumptions 2 and 5 (which imply $Y(d,m)\perp M|D=d,X=x,p(W)=p(w),S=1$), the fifth from Assumptions 1 and 5 (which imply $Y(d,m)\perp D|X=x,p(W)=p(w),S=1$), and the sixth from Assumption 7 by acknowledging that $p(W)=F_V$.

Appendix 4

Table A4.1: Summary statistics and mean differences by gender

Variables	Male($G=1$)	Female($G=0$)	Difference	p-value
Outcome Y (non-logged, refers to selected population with $S=1$)				
Hourly wage	19.370	14.164	5.206	0.000
Mediators X (refer to 1998 unless otherwise is stated)				
Married	0.566	0.568	-0.002	0.882
Years married total since 1979	6.430	7.537	-1.107	0.000
Northeastern region	0.153	0.155	-0.002	0.857
North Central region	0.242	0.237	0.005	0.602
West region	0.206	0.195	0.011	0.244
South region (ref.)	0.399	0.414	-0.015	0.205
Years lived in current region since 1979	14.839	15.246	-0.407	0.000
Resides in SMSA	0.811	0.816	-0.005	0.584
Years lived in SMSA since 1979	13.488	14.201	-0.713	0.000
Less than high school (ref.)	0.129	0.101	0.028	0.000
High school graduate	0.459	0.416	0.043	0.000
Some college	0.208	0.271	-0.063	0.000
College or more	0.204	0.213	-0.009	0.413
First job before 1975	0.065	0.046	0.019	0.001
First job in 1976–79	0.115	0.128	-0.013	0.083
First job after 1979 (ref.)	0.821	0.825	-0.004	0.623
Numer of jobs ever had	10.555	9.239	1.316	0.000
Tenure with current employer (wks.)	276.056	212.662	63.394	0.000
Industry: Primary sector	0.227	0.078	0.149	0.000
Industry: Manufacturing (ref.)	0.140	0.053	0.087	0.000
Industry: Transport	0.115	0.048	0.067	0.000
Industry: Trade	0.134	0.142	-0.008	0.322
Industry: Finance	0.040	0.064	-0.024	0.000
Industry: Services	0.121	0.124	-0.003	0.768
Industry: Professional services	0.113	0.297	-0.184	0.000
Industry: Public administration	0.054	0.052	0.002	0.751
Years worked in current industry since 1982	3.555	2.622	0.933	0.000

Continued on next page

Table A4.1 – continued from previous page

Variables	Male($G=1$)	Female($G=0$)	Difference	p-value
Manager	0.234	0.258	-0.024	0.022
Technical occupation (ref.)	0.039	0.038	0.001	0.907
Occupation in sales	0.067	0.082	-0.015	0.021
Clerical occupation	0.056	0.212	-0.156	0.000
Occupation in service	0.102	0.163	-0.061	0.000
Farmer or laborer	0.276	0.042	0.234	0.000
Operator (machines, transport)	0.170	0.063	0.107	0.000
Years worked in this occupation since 1982	2.180	1.727	0.453	0.000
Employment status: employed	0.877	0.748	0.129	0.000
Number of years employed status since 1979	13.204	11.271	1.933	0.000
Employed full time	0.846	0.599	0.247	0.000
Share of full-time employment 1994-98	0.896	0.658	0.238	0.000
Total number of weeks worked since 1979	661.794	560.408	101.386	0.000
Total number of weeks unempl. since 1979	62.343	49.744	12.599	0.000
Total number of weeks out of LF since 1979	146.118	265.276	-119.158	0.000
Bad health prevents from working	0.045	0.055	-0.010	0.071
Years not working due to bad health s. 1979	0.326	0.557	-0.231	0.000
Pre-treatment covariates W				
Hispanic (ref.)	0.193	0.186	0.007	0.488
Black	0.287	0.297	-0.010	0.413
White	0.520	0.517	0.003	0.840
Born in the U.S.	0.935	0.939	-0.004	0.544
No religion	0.045	0.034	0.011	0.031
Protestant	0.501	0.500	0.001	0.957
Catholic (ref.)	0.352	0.352	0.000	0.967
Other religion	0.096	0.112	-0.016	0.036
Mother born in U.S.	0.884	0.896	-0.012	0.102
Mother's educ. <high school (ref.)	0.376	0.421	-0.045	0.000
Mother's educ. high school graduate	0.393	0.369	0.024	0.048
Mother's educ. some college	0.094	0.091	0.003	0.616
Mother's educ. college/more	0.076	0.071	0.005	0.411
Father born in US	0.878	0.884	-0.006	0.410
Father's educ. <high school (ref.)	0.351	0.366	-0.015	0.201
Father's educ. high school graduate	0.291	0.297	-0.006	0.560

Continued on next page

Table A4.1 – continued from previous page

Variables	Male($G=1$)	Female($G=0$)	Difference	p-value
Father's educ. some college	0.087	0.076	0.011	0.105
Father's educ. college/more	0.131	0.117	0.014	0.085
Order of birth	3.195	3.259	-0.064	0.256
Age in 1979	17.501	17.611	-0.110	0.047
Selection indicator S				
Worked 1,000 hrs or more past year	0.867	0.696	0.171	0.000
Instrumental variables Z				
Number of children under 15	1.286	1.209	0.077	0.008
Number of children under 6	0.353	0.295	0.058	0.000
Mother worked at 14	0.543	0.539	0.004	0.718
N of obs.	3,162	3,496	.	.

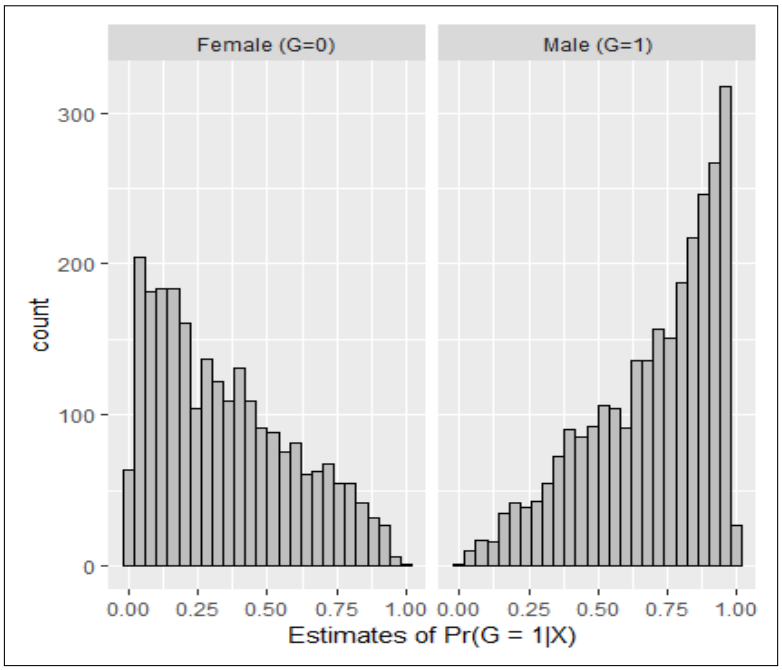

Figure A4.1: Distribution of the estimated $\Pr(G=1|X)$ by treatment states in seleted population

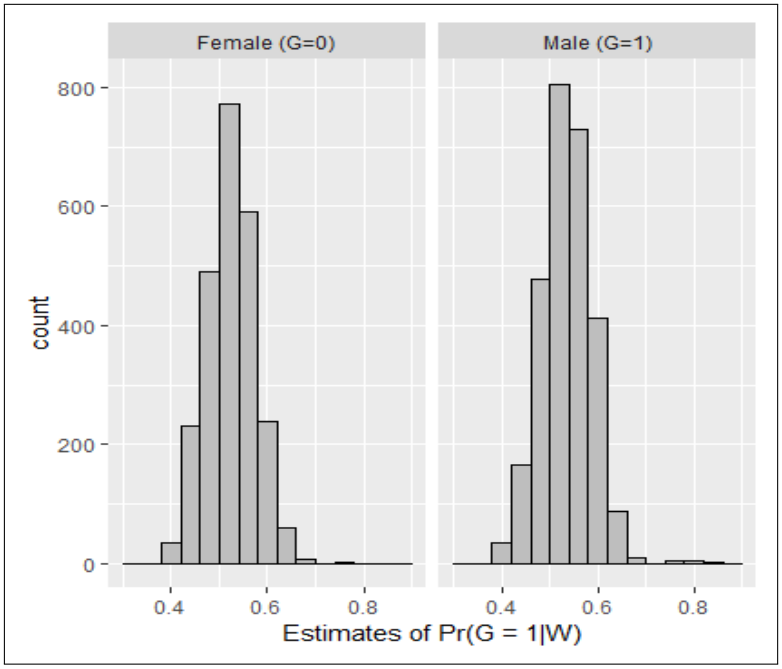

Figure A4.2: Distribution of the estimated $\Pr(G = 1|W)$ by treatment states in seleted population

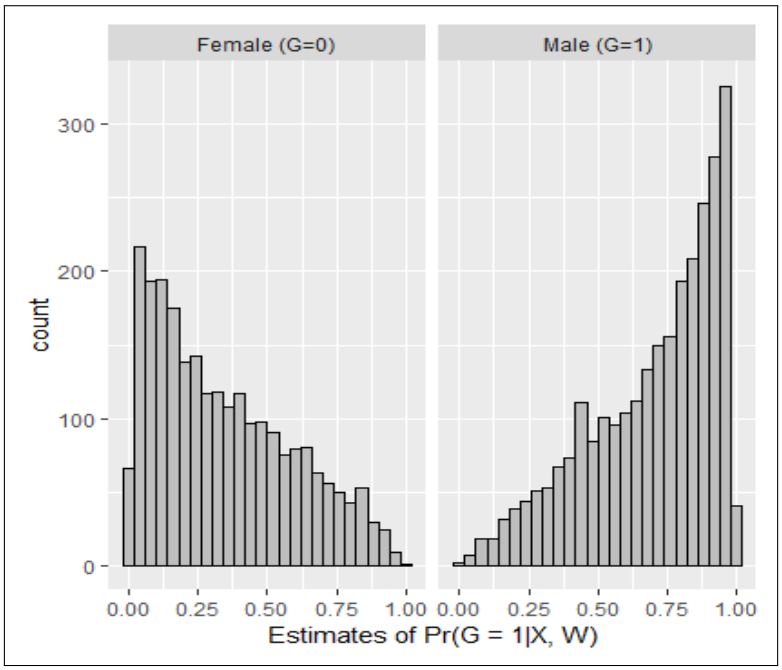

Figure A4.3: Distribution of the estimated $\Pr(G = 1|X, W)$ by treatment states in seleted population

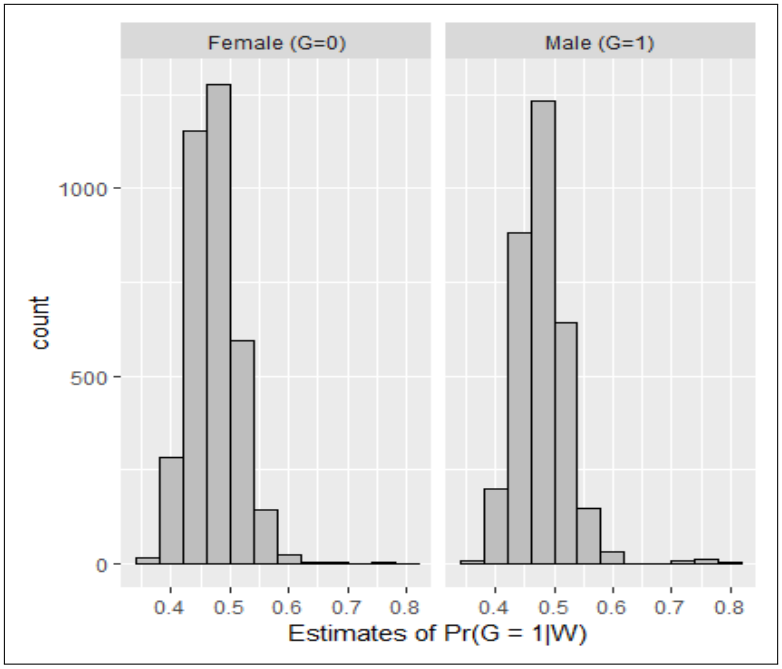

Figure A4.4: Distribution of the estimated $\Pr(G = 1|W)$ by treatment states in total population

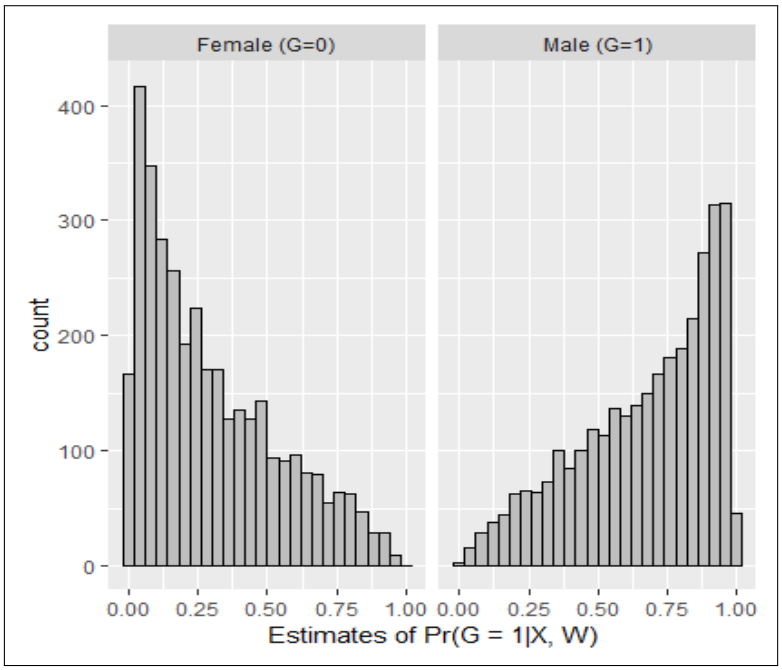

Figure A4.5: Distribution of the estimated $\Pr(G = 1|X, W)$ by treatment states in total population

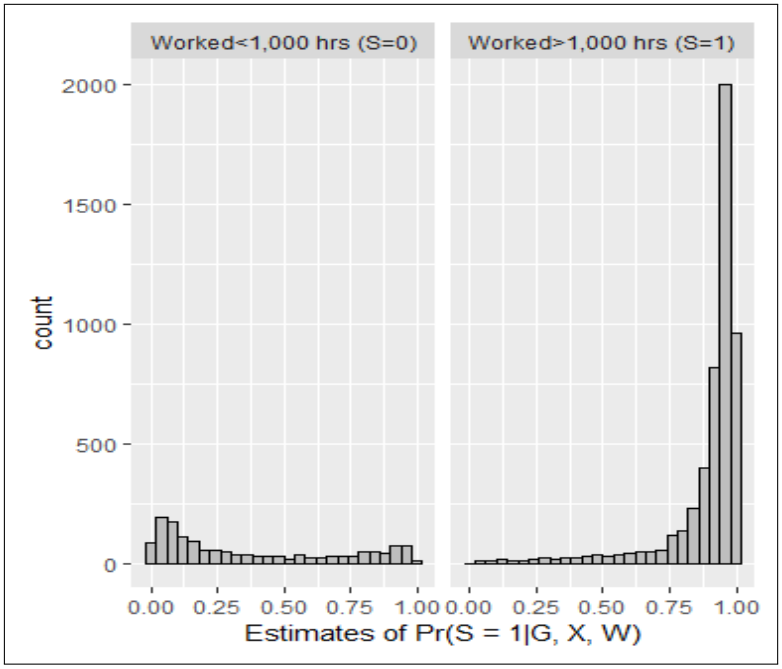

Figure A4.6: Distribution of the estimated $\Pr(S = 1|G, X, W)$ by selection states

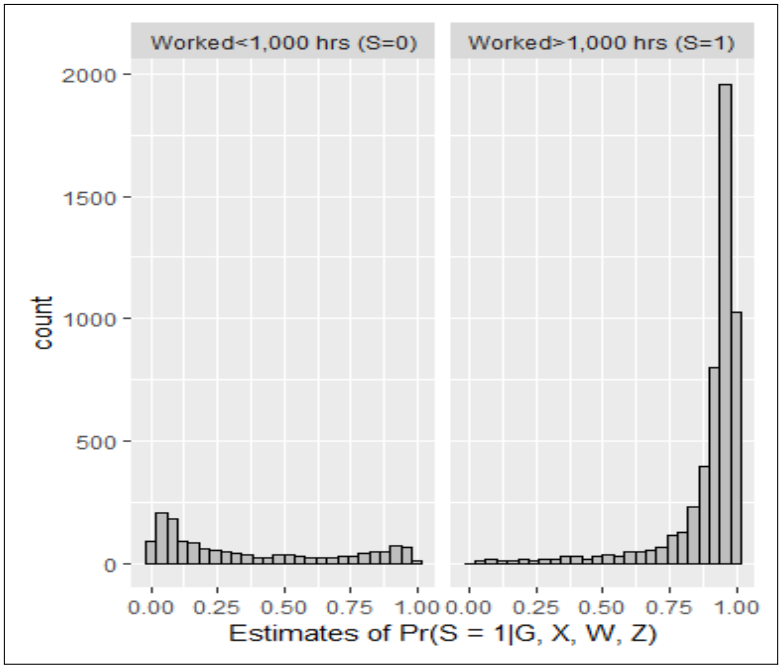

Figure A4.7: Distribution of the estimated $p(Q) = \Pr(S = 1|G, X, W, Z)$ by selection states

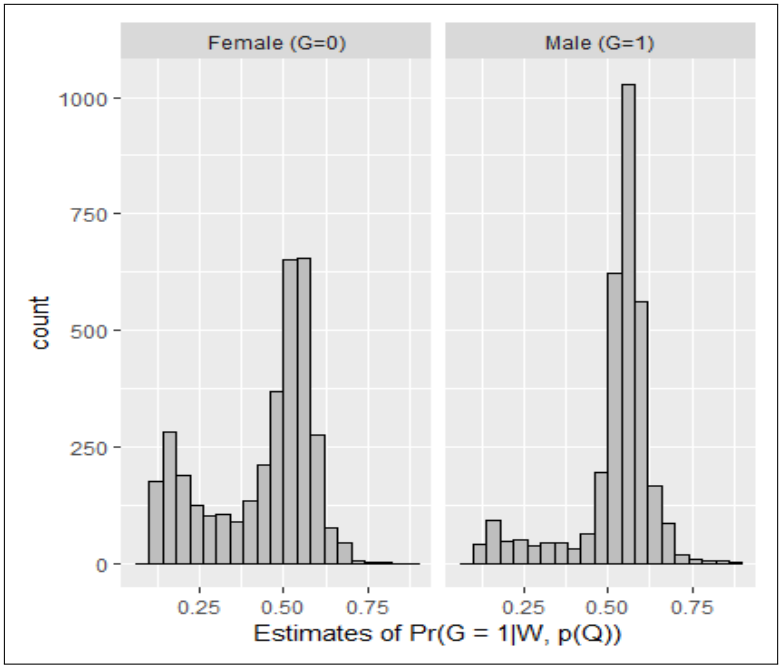

Figure A4.8: Distribution of the estimated $\Pr(G = 1|W, p(Q))$ by treatment states in total population

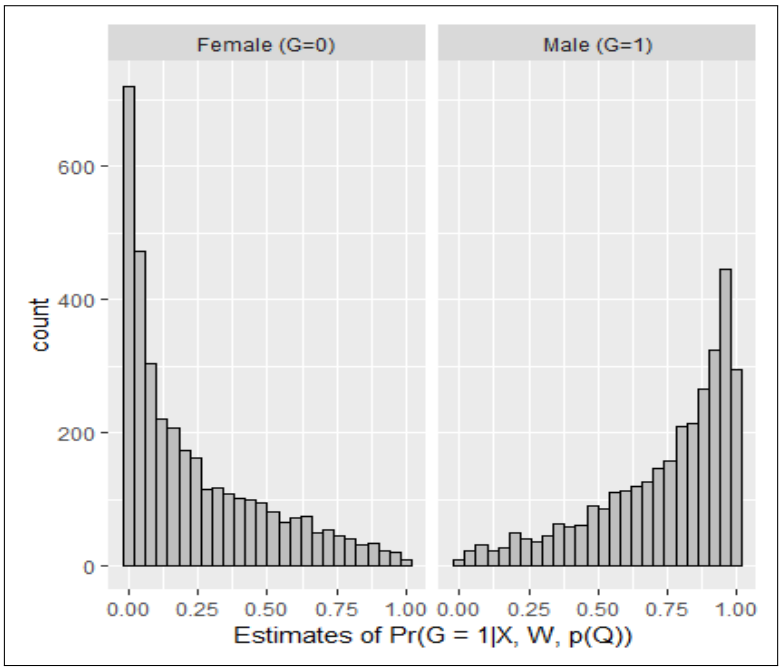

Figure A4.9: Distribution of the estimated $\Pr(G = 1|X, W, p(Q))$ by treatment states in total population

Table A4.2: Summary of the estimated treatment propensity scores in selected population

	Min	Mean	Max	Min	Mean	Max
	Female (G=0)			Male (G=1)		
$\Pr(G=1\|X)$	0.00166	0.34454	0.9819	0.01835	0.6943	0.99047
$\Pr(G=1\|W)$	0.30751	0.52389	0.8023	0.39349	0.53517	0.87171
$\Pr(G=1\|X,W)$	0.00133	0.34042	0.9816	0.01574	0.69795	0.99287

Table A4.3: Summary of the estimated treatment propensity scores in total population

	Min	Mean	Max	Min	Mean	Max
	Female (G=0)			Male (G=1)		
$\Pr(G=1\|W)$	0.36159	0.47140	0.76295	0.37095	0.47881	0.80260
$\Pr(G=1\|X,W)$	0.00081	0.29707	0.97202	0.01322	0.67155	0.99619
$\Pr(G=1\|W,p(Q))$	0.10313	0.43167	0.80670	0.09923	0.52273	0.89403
$\Pr(G=1\|X,W,p(Q))$	3.22×10^{-7}	0.23804	0.99983	0.00065	0.73682	0.99999

Table A4.4: Summary of the estimated selection propensity scores in total population

	Min	Mean	Max	Min	Mean	Max
	Did not work (S=0)			Worked (S=1)		
$\Pr(S=1\|G,X,W)$	0.00327	0.36952	0.99272	0.02076	0.89392	0.99911
$\Pr(S=1\|G,X,W,Z)$	0.00315	0.36669	0.99386	0.02150	0.89473	0.99909

Table A4.5: Number of trimmed observations for each propensity score

Trimming condition	obs.	% tot.
Treatment propensity scores in selected population		
$\Pr(G=1\|X)<0.01$	28	0.5
$\Pr(G=1\|W)<0.01$	0	0.0
$\Pr(G=1\|W)>0.99$	0	0.0
$\Pr(G=1\|X,W)<0.01$	28	0.5
Treatment and selection propensity scores in total population		
$\Pr(G=1\|W)<0.01$	0	0.0
$\Pr(G=1\|W)>0.99$	0	0.0
$\Pr(G=1\|X,W)<0.01$	61	0.9
$\Pr(S=1\|G,X,W)<0.01$	29	0.4
$\Pr(S=1\|G,X,W,Z)<0.01$	30	0.4
$\Pr(G=1\|W,p(Q))<0.01$	0	0.0
$\Pr(G=1\|W,p(Q))>0.99$	0	0.0
$\Pr(G=1\|X,W,p(Q))<0.01$	554	8.3

Table A4.6: Robustness check: no interactions in X

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)				Trimmed	
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	% tot.	obs.	%
Oaxaca-Bl.	0.299	0.019	0.000	0.084	0.020	0.000	28%	0.215	0.023	0.000	72%	0	0%
IPW no W	0.295	0.019	0.000	0.093	0.029	0.001	32%	0.201	0.030	0.000	68%	21	0%
IPW with W	0.265	0.017	0.000	0.074	0.028	0.009	28%	0.192	0.030	0.000	72%	22	0%
IPW MAR	0.375	0.034	0.000	0.175	0.033	0.000	46%	0.201	0.033	0.000	53%	44	1%
IPW IV	0.148	0.324	0.649	0.031	0.102	0.758	21%	0.116	0.328	0.723	79%	673	10%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications. The trimming rule discards observations with propensity scores (specific to each estimator) below 0.01 or above 0.99.

Table A4.7: Mother worked at 14 as an additional IV, full set of X

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)				Trimmed	
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	% tot.	obs.	%
IPW IV	0.140	0.156	0.369	-0.005	0.080	0.948	-4%	0.145	0.175	0.408	104%	583	9%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications. The trimming rule discards observations with $\Pr(G = 1|X = x, W = w, p(Q) = p(q)) < 0.01$, $\Pr(G = 1|W = w, p(Q) = p(q)) > 0.99$, and $p(q) < 0.01$.