Empirical Essays on Preference Formation and Economic Decision Making

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

Svitlana Tyahlo

from Ukraine

Accepted by the Faculty of Economics and Social Sciences
on 21 September 2020 at the proposal of

Prof. Dr. Reiner Eichenberger (first advisor)
Prof. Dr. Holger Herz (second advisor)

Fribourg, 2020

To my family

Contents

Summary **1**

1 How war affects political attitudes: Evidence from eastern Ukraine **3**

1.1 Introduction . . . 3
1.2 Background of the conflict in Donbas . . . 7
1.3 Data and treatment definition . . . 9
1.4 Econometric methods and identification . . . 14
1.5 Main results . . . 20
1.6 Robustness checks . . . 23
1.7 Conclusion . . . 28
1.A Additional figures and tables . . . 30

2 Education, fertility and labor force attachment: A mediation analysis of older women **40**

2.1 Introduction . . . 40
2.2 Literature . . . 42
2.2.1 Education and labor market outcomes . . . 42
2.2.2 Education and fertility . . . 43
2.2.3 Fertility and female labor supply . . . 44
2.2.4 Our contribution . . . 45
2.3 Data . . . 45
2.4 Methodology . . . 49
2.5 Results . . . 51
2.5.1 Selective mortality . . . 53
2.5.2 Heterogeneous effects . . . 55
2.5.3 Endogenous schooling and fertility . . . 57
2.6 Conclusion . . . 58
2.A Additional tables . . . 60

3 Linguistics and time preferences: The role of language in shaping intertemporal choices **69**

3.1 Introduction . . . 69
3.2 Institutional background . . . 73

3.3 Theory, hypotheses and data . . . 74
3.3.1 Time discounting and language . . . 74
3.3.2 Preference measures . . . 76
3.3.3 Procedures . . . 78
3.4 Results . . . 79
3.4.1 Descriptive statistics . . . 79
3.4.2 Analysis of switch points . . . 81
3.4.3 Analysis of individual preference parameters . . . 84
3.4.4 Robustness checks . . . 87
3.5 Conclusion . . . 90
3.A Additional figures and tables . . . 93

4 Intertemporal choice under social comparison: A real-effort experiment 99
4.1 Introduction . . . 99
4.2 Work environment . . . 103
4.3 Experiment . . . 104
4.3.1 Intertemporal choice task . . . 104
4.3.2 Treatments . . . 106
4.3.3 Procedures and data . . . 107
4.4 Conceptual framework . . . 108
4.4.1 Behavioral predictions . . . 108
4.4.2 Behavioral measures . . . 111
4.5 Results . . . 112
4.5.1 Within-group analysis . . . 112
4.5.2 Analysis of treatment effects . . . 115
4.5.3 Robustness checks . . . 119
4.6 Discussion . . . 122
4.6.1 Treatment effects . . . 122
4.6.2 Time in/consistency . . . 124
4.6.3 Further insights . . . 125
4.7 Conclusion . . . 126
4.A Additional figures and tables . . . 127

References 134

Summary

This dissertation consists of four independent papers on economically relevant topics that captured my attention because of personal background and professional experiences, such as coming from Ukraine, being an educated woman, living in a bilingual Swiss region and observing people delaying their work. Throughout my papers, I empirically analyze the determinants of individual attitudes and behavior that entail important economic consequences. In particular, the first two papers are observational studies related to political economy and labor market policy, respectively, while the last two papers involve incentivized intertemporal choice experiments. Analytical methods are carefully chosen to address specific research questions given the quantity and quality of data. Since the primary goal is to establish causal relationships, the results and thus the identifying assumptions are challenged with transparent robustness checks and critical discussions. The papers are organized in chapters as following.

Chapter 1 evaluates the impact of the ongoing war in eastern Ukraine on the political attitudes towards Ukraine and Russia among the population in the war-affected area controlled by the Ukrainian government. This topic is important because political sentiments can shape political powers, voting outcomes and thus prospective institutions, which will govern approaches to conflict resolution and post-war economic recovery in the region. The war geography allows defining two groups of citizens with different war exposure. Based on unique survey data from 2013 (prior to the war) and 2015 (after the outbreak of the war), a before-after analysis and a difference-in-difference approach are applied to examine the effect of war exposure. The results indicate that one year of conflict negatively affected attitudes towards Russia, while no statistically robust differences are found for sentiments towards Ukraine. These findings are in line with the outcome of the 2014 Ukrainian parliamentary elections, when the pro-Russian vote substantially dropped in eastern Ukraine compared to the previous elections.

In Chapter 2, I investigate long-run effects of education on female labor market attachment mediated by realized fertility. In the context of population ageing, educated women are an important resource of labor supply. Investments into female education, however, raise opportunity costs of motherhood in terms of forgone earnings, leading to a trade-off between labor force participation and fertility decisions. It is therefore important to understand whether education increases female working lives mainly through reduced fertility. Using mediation analysis supported by the rich data from the Survey of Health, Ageing and Retirement in Europe, I find that schooling is positively associated with employment status, working hours and labor survival age of older women. Education-induced fertility modestly contributes to the total schooling effects, which is stimulating news for policy makers.

Chapter 3 sheds light on the relationship between language and time preferences. While recent evidence suggests that the grammatical association of the present and the future in a language positively correlates with patience across language groups, the underlying mechanisms remain unclear. Our study compares time preferences of two language groups that differ in their encoding of time. More precisely, we conduct incentivized choice experiments among French and German speakers in a bilingual region in Switzerland, with shared institutions and very similar socioeconomic conditions between the two language groups. We find that German speakers are significantly more patient than French speakers, and differences are particularly pronounced when immediate payments are involved. The estimated preference parameters of a quasi-hyperbolic discounting model suggest that German speakers display lower discounting and are significantly less prone to present bias than French speakers.

Chapter 4 also examines intertemporal choice behavior. The context however differs. In a randomized real-effort experiment with Amazon Mechanical Turk workers, I study the impact of social comparison on intertemporal choice. Social comparison entails varying information about effort reallocations of previous participants who completed a similar task. I find that social comparison affects effort allocations of men and women differentially. Observing that 48% of past participants behaved time-consistently, men are significantly more likely to make dynamically consistent choices, while women are significantly less likely to do so. On average, men also exhibit significantly smaller estimates of the present-bias parameter of a quasi-hyperbolic discounting model. I further find that observing peer procrastination significantly increases the propensity of women to behave time-consistently but does not affect men. These findings suggest that gender-specific social comparison based on situational similarity can be an effective solution to dynamic inconsistency in effort even when individual time preferences are not known beforehand.

1 How war affects political attitudes: Evidence from eastern Ukraine[1]

Overview

This study empirically evaluates the impact of the ongoing war in eastern Ukraine on the political attitudes towards Ukraine and Russia among the population living close to the war zone on the territory controlled by the Ukrainian government. Exploiting unique survey data from 2013 (prior to the war) and 2015, we employ a before-after analysis and a difference-in-differences approach to infer how the war has affected two different groups defined by exposure to the war zone. We consider both linear and semiparametric estimation based on inverse probability weighting. Our results suggest that one year of conflict negatively affected attitudes towards Russia, while no statistically robust differences were found for sentiments towards Ukraine.

1.1 Introduction

How does a military conflict affect the political capital of neighboring countries with historical connections among the war-exposed population? We address this question empirically by evaluating the impact of the ongoing military conflict in eastern Ukraine on the political attitudes held by the population in the war-affected area towards Ukraine and Russia.

The pro-Russian unrest in the east of Ukraine, which started shortly after the "Euromaidan" movement and Russia's annexation of Crimea, escalated to a violent war in April 2014, where pro-Ukrainian and pro-Russian views clashed. Given its geopolitical context and its implications for lasting stability and security in Europe, this military conflict echoed far beyond in multilateral political and economic relations.[2] Exploiting unique survey data from a repeated cross section, we investigate how preferences about the political status of Ukraine and sentiments towards Ukraine and Russia evolved among individuals living close to the war zone on the Ukrainian-controlled

[1]This chapter is co-authored with Martin Huber. The earlier version of this paper is available as SES Working Paper No 472 (2016). The collection of the data used in this study was funded by the Grant CR11I1L_135348 "Region, Nation and Beyond. A Transcultural and Interdisciplinary Reconceptualization of Ukraine" of the Swiss National Science Foundation (2013–2015) and the Wolodymyr George Danyliw Foundation (2015).

[2]In response to Russia's supposed role in the conflict in Ukraine, the European Union (EU) and other countries (e.g. the United States, Canada, Norway, Switzerland, Japan, and Australia) introduced a range of diplomatic and economic sanctions against a list of individuals and companies from Russia and Ukraine. Resorting to countermeasures, Russia banned food imports from a number of countries. Christen et al. (2015) assess the potential economic consequences of export sanctions between the EU plus Switzerland and Russia.

territory between early 2013, i.e. 13 months before the war, and 2015, i.e. 11 months after the outbreak of the war. Our analysis therefore focuses on the eastern part of Ukraine as far as controlled by government forces, namely Donetsk, Luhansk, Kharkiv, Dnipropetrovsk, and Zaporizhia oblasts (administrative units), which before the war used to share strong cultural and socioeconomic ties. Specifically, these regions represent the most extensively Russian-speaking part of Ukraine that had the longest common history with Russia (see Barrington and Herron, 2004).

We use this setting to answer two key questions: How does exposure to a military conflict affect the opposing — pro-Ukrainian vs. pro-Russian — political views? Does the degree of exposure to the conflict matter? These are important questions to address because political sentiments can shape political powers, voting outcomes and thus prospective institutions, which will govern approaches to conflict resolution and post-war economic recovery in the region. Hence, this paper sheds light on the political costs of the ongoing war to the involved parties.[3] As to the first question, the effect of war on political attitudes is not a priori obvious but depends on people's perceptions of who is an aggressor and who is a defender, which is context specific. For the second question, we anticipate a stronger effect in the population with higher exposure to war because of greater inconveniences these people experience.

To infer the causal effect of the conflict, our empirical strategy relies on variability in exposure to war across the country. In fact, the violent fighting only took place in certain parts of Donetsk and Luhansk oblasts, the so-called Donbas region. However, due to the conflict, the access to important services provided in the cities of Donetsk and Luhansk — the administrative centers that fell under the control of pro-Russian forces — got disrupted, affecting all people in Donbas. Furthermore, the entire country was affected by the conflict, for instance, through the economic and social consequences, the recruitment of troops, the discussions in politics, the media, and the civil society. This allows splitting the sample into two treatment groups: individuals in Donetsk and Luhansk oblasts (the Donbas region) represent the high exposure group, while individuals in other eastern oblasts belong to the low exposure group.

We employ two econometric approaches. First, we apply a before-after analysis to examine intra-group changes in attitudes over time. This only yields unbiased effect estimates if time trends in attitudes can be ruled out, at least conditional on observed characteristics. Our second strategy is based on a difference-in-differences (DiD) approach to investigate inter-group divergence over time. Under particular assumptions, namely when the impact in the high exposure group weakly dominates that in the low exposure group in absolute terms, a lower bound for the absolute effect of the war on political attitudes is obtained (see Fricke, 2017). We control for a range of observed socioeconomic characteristics and consider both parametric and semiparametric estimation. The

[3] According to the preliminary findings of the International Criminal Court, there is an international armed conflict between Russia and Ukraine in eastern Ukraine. See `https://www.icc-cpi.int/itemsDocuments/181205-rep-otp-PE-ENG.pdf`, retrieved 7 December 2019.

latter is based on inverse probability weighting by the propensity score to belong to the high exposure group.

Our findings suggest that the political attitudes towards Russia have deteriorated as a consequence of the war. In either group, the before-after differences in supporting a union or one state with Russia are substantially negative and highly significant in the majority of our specifications. Furthermore, for the view that the Ukrainian and Russian cultures are the same, the DiD approach yields a sizable and significant negative effect in the main specification, which is mostly driven by before-after differences in Donbas. The estimated effects are relatively robust to the choice of control variables and the definition of the treatment groups. In contrast, no statistically robust effects on attitudes and sentiments towards Ukraine and Ukrainians, i.e. sympathy for Ukraine and self-association with other Ukrainians, are found.

This study relates to the growing literature investigating the political impacts of exposure to war and violence. For instance, Bellows and Miguel (2009), Blattman (2009), and Voors et al. (2012) find that individuals who personally experienced wartime violence and trauma during civil wars in Africa increase their political and civic engagements and community leadership. Also Bateson (2012) provides cross-country evidence that individuals who report recent crime victimization are more politically active than their non-victimized peers. Applying a regression discontinuity-type design, Garcia-Ponce and Pasquale (2015) document that individuals indirectly exposed to political repression in Zimbabwe self-report higher levels of trust into the state and its institutions. Erikson and Stoker (2011) examine how the 1969 Vietnam draft lottery influenced political attitudes of males in the United States. Men with vulnerable draft numbers are found to be more antiwar and liberal in their voting behavior than those with safe draft numbers. Employing a DiD and synthetic control approach, Montalvo (2011) shows that the terrorist attacks in Madrid shifted voters' choices in the 2004 congressional election. Though Balcells and Torrats-Espinosa (2018) confirm that terrorist attacks in Spain enhance individuals' intent to participate in a future election, they find no evidence that the attacks change support for the incumbent party.

More closely related to our paper, Rohner et al. (2013), who employ an instrumental variable method to study the effect of the civil conflict in Uganda, find that intense violence strengthens within-ethnic group ties but significantly weakens trust towards other Ugandans. Using endorsement experiments, Lyall et al. (2013) document the asymmetric effects of wartime violence on attitudes towards out-group vs. in-group combatants in Afghanistan: while violence by the out-group leads to reduced support for that group and increased support for the in-group, in-group violence does not lead to a transfer of support to the out-group. As one potential mechanism of war, DellaVigna et al. (2014) investigate how exposure to nationalistic cross-border Serbian public radio affects the voting behavior and anti-Serbian sentiment in the post-conflict region of Croatia at the border with Serbia.

In the case of Ukraine, Rozenas et al. (2017) analyze long-term political effects of Soviet state violence in western Ukraine. Using an instrumental variable approach and a fuzzy regression discontinuity design, they find that communities subjected to a greater intensity of deportation in the 1940s are now less likely to vote for pro-Russian parties. Most relevant for our work is the study by Coupé and Obrizan (2016b), which investigates the effects of violence on political outcomes in two cities of the Donetsk region that were temporarily controlled by the pro-Russian militants. Relying on cross sectional survey data, their results suggest that physical damage is negatively associated with the turnout probability and the likelihood to know local political representatives. Property damage, on the other hand, increases self-reported votes for pro-Western parties and reduces support for Donbas remaining a part of Ukraine or for any compromise with the pro-Russian forces.[4]

In contrast to the majority of studies that focus on post-violence outcomes, our analysis — like the one by Coupé and Obrizan (2016b) — concerns a point in time when the war was still ongoing. However, our study differs from that of Coupé and Obrizan (2016b) in three dimensions. First, the outcome variables considered mostly differ: while the present work has a stronger focus on political attitudes towards Ukraine and Russia, Coupé and Obrizan (2016b) predominantly (but not exclusively) examine election behavior. We believe that political attitudes — though they may not entirely be covered by available political choices — positively correlate with individual voting behavior, which in turn shapes the resulting institutions. Second, Coupé and Obrizan (2016b) rely on observations within an area of intense fighting, while we exploit variation across areas with higher and lower exposure to war. Third, while Coupé and Obrizan (2016b) use a single cross section, our repeated cross section allows observing the outcome variables already prior to the conflict and applying a DiD approach in order to tackle confounding related to time-constant unobservables. In addition to an OLS-based implementation, we base DiD estimation on a semiparametric weighting approach, which is more flexible in terms of functional form assumptions than the estimators conventionally used in the empirical literature.

Our findings confirm our hypotheses and the general result from previous studies that exposure to war and violence has consequences with respect to political attitudes, albeit the context differs from much of the literature. Specifically, our estimates suggest that the attitude towards Russia has deteriorated as a consequence of the war, which is probably driven by Russia's perceived role as a major proponent of the separatist objectives through its politics, media, and likely military support.[5] From a political perspective, this suggests that the conflict did not make eastern Ukrainians more sympathetic towards Russia, at least in the government-controlled areas,

[4] Another study by Coupé and Obrizan (2016a) identifies, based on a DiD approach, a significant drop in happiness in the war-affected Donbas compared to the rest of Ukraine. A further empirical study concerned with the conflict in eastern Ukraine is Zhukov (2016), who, however, does not investigate any effects, but rather the triggers of "rebel" activity. His results suggest that pre-conflict economic ties with Russia are a better predictor for "rebel" activity than the ethnolinguistic composition of municipalities.

[5] At the annual press conference on 17 December 2015, Vladimir Putin admitted that there were military intelligence officers operating in the east of Ukraine. See `www.theguardian.com/world/live/2015/dec/17/vladimir-putins-annual-press-conference-live`, retrieved 23 April 2016.

as some might have speculated in the light of the close linguistic and cultural ties with Russia. Quite the contrary, Russia appears to have lost political capital in the most Russian-speaking part of Ukraine, while no statistically significant negative effects on attitudes and sentiments towards Ukraine were found. These results are somewhat in line with the finding of Coupé and Obrizan (2016b) that the experience of property damage decreases the support for the view that the Ukrainian government should compromise with Russia. Our data also confirm the outcome of the 2014 Ukrainian parliamentary elections, when the vote share for the pro-Russian parties substantially dropped in eastern Ukraine, particularly in the unoccupied area of Donbas, compared to the previous elections in 2012.

The remainder of this paper is organized as follows. Section 1.2 describes the conflict background. Section 1.3 presents the data and the treatment groups. Section 1.4 outlines our econometric approaches based on parametric and semiparametric before-after and DiD estimation and discusses identification issues, such as migration patterns in the region. Sections 1.5 and 1.6 report the main results and robustness checks, respectively. Section 1.7 concludes.

1.2 Background of the conflict in Donbas

Since independence, the territory of Ukraine had been free from military confrontations until the pro-Russian unrest in Donbas (Donetsk and Luhansk oblasts)[6] escalated to the status of war in April 2014 following a series of relevant events. The underlying context is important to understand the impact of the conflict on political sentiments analyzed in this paper.

The signing of the EU-Ukraine Association Agreement, a core element of which was closer economic integration through the Deep and Comprehensive Free Trade Area (DCFTA) between the EU and Ukraine, was intended for late November 2013.[7] In response, Russia — dissatisfied with the potential agreement — imposed import restrictions on certain Ukrainian products and warned of tighter sanctions if the agreement got signed. In September 2013, the chief economic adviser of the Russian president explicitly voiced the possibility of separatist movements in the Russian-speaking regions of Ukraine and suggested that Russia would consider the bilateral treaty defining the countries' border to be void.[8]

Reportedly concerned about the industrial production decline and relations with the members

[6]The territory of Donbas is historically associated with the Ukrainian Cossacks, the so-called Zaporizka Sich (16th–18th centuries). The region was controlled by the Russian Empire from the late 18th century and then by the Soviet Union. Despite a large immigration of Russians into Donbas after World War II, ethnic Ukrainians were still in the majority. There is, however, no agreement between the Ukrainian and Russian versions of the Donbas history (see Wilson, 1995).

[7]The Guide to the EU-Ukraine Association Agreement is available at `eeas.europa.eu/images/top_stories/140912_eu-ukraine-associatin-agreement-quick_guide.pdf`, retrieved 23 April 2016.

[8]See `www.theguardian.com/world/2013/sep/22/ukraine-european-union-trade-russia`, retrieved 4 November 2018.

of the Commonwealth of Independent States (CIS), the Ukrainian government decided to temporarily suspend the preparation for signing the Association Agreement with the EU.[9] Consequently, the "Euromaidan" movement — a wave of public demonstrations — spread from Kyiv to major cities of Ukraine. The initial demand for closer European integration quickly expanded to requests for political change in the country. The culmination came in February 2014, when violent fights with fatalities in the center of Kyiv led to the ouster of the then president who fled the country. The Ukrainian interim government signed the political provisions of the Association Agreement with the EU in late March 2014, and the newly elected president signed the economic part in June 2014.

Mid-March 2014, the "referendum" on the status of Crimea — with no option to vote for the status quo[10] — took place,[11] after the peninsula had already been penetrated by armed forces without insignia most likely belonging to the Russian army.[12] The "referendum" led to Russia's annexation of Crimea. Consequently, the Ukrainian government de facto lost control over the peninsula.

In April 2014, the unrest moved to Donbas, where pro-Russian forces occupied governmental buildings in Donetsk and Luhansk and self-declared the "Donetsk People's Republic" followed by the "Luhansk People's Republic".[13] Ukraine's interim president in turn launched an "anti-terrorist operation" against the pro-Russian fighters. Despite attempts to de-escalate the conflict in Donbas,[14] "referendums" took place on the occupied territories to legitimize the self-declared Donetsk and Luhansk People's Republics.[15] Avoiding formal diplomatic recognition, Russia expressed its respect for the outcomes of the "referendums" and its hope for the "civilized implementation" thereof.[16] Consequently, neither annexed Crimea nor the occupied areas of Donbas participated in the subsequent Ukrainian presidential or parliamentary elections in 2014.

Since the beginning of the war in Donbas, several waves of army mobilization followed across the territories governed by the Ukrainian authorities. In addition, the military tax was introduced

[9] See `en.interfax.com.ua/news/general/176144.html`, retrieved 23 April 2016.

[10] See `www.nytimes.com/2014/03/15/world/europe/crimea-vote-does-not-offer-choice-of-status-quo.html?_r=0`, retrieved 23 April 2016.

[11] The Organization for Security and Co-operation in Europe (OSCE) considered the Crimean "referendum" illegal. The OSCE press release is available at `www.osce.org/cio/116313`, retrieved 23 April 2016.

[12] See, for instance, the report of "Suomen Sotilas" (Soldier of Finland) at `web.archive.org/web/20150330124704/http://www.suomensotilas.fi/en/artikkelit/crimea-invaded-high-readiness-forces-russian-federation`, retrieved 23 April 2016.

[13] See `www.hrw.org/news/2014/09/11/eastern-ukraine-questions-and-answers-about-laws-war`, retrieved 23 April 2016.

[14] See `www.theguardian.com/world/2014/apr/18/ukraine-separatists-occupation-geneva-agreement`, retrieved 23 April 2016.

[15] See `www.theguardian.com/world/2014/may/10/donetsk-referendum-ukraine-civil-war`, retrieved 23 April 2016.

[16] See `www.reuters.com/article/us-ukraine-crisis-russia-kremlin-idUSBREA4B04020140512`, retrieved 23 April 2016.

in August 2014. Furthermore, the Russian state-controlled media — a source of anti-Ukrainian narratives — were banned in the government-held part of Ukraine. Hence, the entire country was affected by the war in Donbas.

As the ceasefire agreement of 5 September 2014[17] had failed to resolve the conflict in Donbas, the intense consultations of the trilateral contact group — consisting of Ukraine, Russia and the Organization for Security and Co-operation in Europe (OSCE) — continued.[18] In January 2015, Ukraine declared Russia an aggressor state supporting terrorism.[19] In March 2015, the Ukrainian parliament approved a law on the "special status" of certain parts of Donbas with the aim of holding local elections there.[20] After months of violations, Ukraine and the pro-Russian fighters ultimately decided to respect the ceasefire from 1 September 2015 on. That entailed a considerable reduction in fighting, yet clashes and casualties continued and lasting peace was still out of reach.[21]

The conflict in Donbas and the preceding events challenged the status quo of Ukraine as an independent and neutral state. Though Ukraine's relationship with Russia had always been a sensitive topic, the decision to join the DCFA with the EU rather than the Customs Union with Russia led to the loss of Ukraine's control over part of its territory and population with substantial economic and human consequences.[22] The split of the country and the annexation of Crimea by Russia became a reality.

1.3 Data and treatment definition

Our data come from a representative population survey in Ukraine conducted in February-March 2013 by the sociological institute "Rating" and repeated in March-April 2015 by the company "Socioinform" on behalf of an interdisciplinary research project on regionalism in Ukraine.[23] The sample consists of a repeated cross section with 6000 individual observations per survey year.[24] The first wave includes all 24 oblasts (administrative units) plus the Autonomous Republic of Crimea and the city of Sevastopol, while the second wave only covers the territory controlled by

[17] The document is available in Russian at `www.osce.org/home/123257`, retrieved 23 April 2016.

[18] The Minsk II protocol of 12 February 2015 is available in Russian at `www.osce.org/cio/140156`, retrieved 23 April 2016.

[19] See `www.unian.info/politics/1036816-ukrainian-parliament-declares-russia-aggressor-state.html`, retrieved 23 April 2016.

[20] See `www.reuters.com/article/us-ukraine-crisis-status-idUSKBN0MD1ZK20150317`, retrieved 23 April 2016.

[21] See, for instance, the UN report at `www.un.org/press/en/2015/sc12154.doc.htm`, retrieved 23 April 2016.

[22] In 2019, the International Court of Justice has agreed it has jurisdiction to hear claims by Ukraine against Russia related to the conflict in eastern Ukraine. See `https://news.un.org/en/story/2019/11/1051001?fbclid=IwAR3e_2GRvrKvydI451_3FXiUwNwUSEgZDVx8LerAz_RiUri76wHmOOrfI4A`, retrieved 23 November 2019.

[23] The research project involved historians, sociologists, anthropologists, economists, literary critics and linguist from Austria, Canada, Germany, Poland, Russia, Switzerland, Ukraine and the USA. Further details are available at `https://gce.unisg.ch/de/ua-regio`, retrieved 3 December 2019.

[24] Since interviews were conducted by different survey agencies, there is no information whether any respondent or household participated in both survey waves.

the Ukrainian authorities at that time. For this reason, Crimea, which was annexed by Russia in March 2014, is excluded, as well as those parts of Donetsk and Luhansk oblasts controlled by the pro-Russian fighters. The sampling was based on stratification by gender, age and municipality type.

The face-to-face interviews were conducted either in Ukrainian or Russian, as preferred by a respondent. Therefore, no misunderstandings related to language issues are to be expected. Interviewers came from the main cities of the oblasts in which they conducted interviews, i.e. interviewers travelled within one oblast only, which reduced the likelihood of an interviewer to be perceived as stranger by respondents and consequently mistrusted.[25] Furthermore, interviewers were instructed to emphasize that the survey was carried out for research purposes only and that analyses would be performed anonymously on the aggregate level, in order to minimize mistrust and reluctance to answer politically sensitive questions. Indeed, response rates are rather high as outlined further below. It is worth mentioning that respondents did not receive any material incentives for their survey participation which could otherwise potentially influence the observed response rates or expressed opinions.

This study focuses on questions about preferences for the political future of Ukraine, attitudes about Ukrainian–Russian relations and sentiments towards Ukraine and Ukrainians.[26] Respondents were asked, inter alia, to state their most preferred option out of several mutually exclusive scenarios: Ukraine remains an independent and neutral state, Ukraine enters a large union including Russia, Ukraine splits into separate states.[27] In our analysis each option is a binary variable which is coded as one if it is picked as preference and zero otherwise. Furthermore, interviewees were asked to which extent they agreed or disagreed with the following statements: the Ukrainian and Russian cultures are exactly the same, Ukraine and Russia should form one state, they love Ukraine, and they speak about Ukrainians as “we” and not as “they”. The former two questions are evaluated on a point scale from 1 (“fully disagree”) to 7 (“fully agree”), and the latter two are assessed on a scale from 1 (“definitely no”) to 5 (“definitely yes”). Besides the outcomes of interest, the data also include a range of socioeconomic information about the respondents, such as gender, age, education, native language, religion, occupation, marital status, household size, residence and self-assessed material conditions.

To evaluate how the war in eastern Ukraine affected political attitudes and sentiments of the local population, we focus our analysis on the territories controlled by Ukraine sufficiently close to the war zone when the second survey wave took place. Besides the government-controlled parts of the war-affected Donetsk and Luhansk oblasts (the Donbas region), these areas include three

[25] In 2015, interviewers in Donetsk oblast came from Donetsk and in Luhansk oblast those came from Siverodonetsk.

[26] The original questionnaires are available on request in Ukrainian or Russian.

[27] The question included two another scenarios: Ukraine joins the EU, Ukraine joins a large union including Central and Eastern European (CEE) states. These two options are excluded from our analysis as they are not at the center of this paper’s focus on attitudes towards Ukraine and Russia. The answers of respondents who chose one of these scenarios are coded as zeros for the other options.

further oblasts: Kharkiv, Dnipropetrovsk, and Zaporizhia (hereinafter referred to as *the remainder of the east*). Taken together, these five oblasts form the east of Ukraine, which is characterized by a high concentration of native Russian speakers and heavy industry as well as the geographic proximity and historic connections to Russia. Figure 1.1 illustrates the front line in Donbas end of March 2015 to show which parts of Donetsk and Luhansk oblasts were under government control at the time of the second survey.

Figure 1.1: Front line in Donbas as of 31 March 2015

Source: Information Analysis Center of the National Security and Defense Council of Ukraine, `www.mediarnbo.org`, retrieved 24 July 2015.

Given that the interviewers could reach only areas controlled by the Ukrainian authorities in early 2015 and that the sociological agencies conducting the surveys were different in the two periods, the included cities and villages in eastern Ukraine only partially coincide across the survey waves. To maximize the comparability of observations in 2013 and 2015, and thus to control for municipality-specific unobservables, our main analysis and most robustness checks only include cities that were observed in both periods.[28] The evaluation sample consists of 920 observations in 2013 and 1153 observations in 2015 (see Appendix Table 1.A1), and also includes 8 municipalities with 380 observations over both periods in Donbas and 19 municipalities with 1693 observations in the remainder of the east. However, our several robustness checks (see Tables 1.7 and 1.8 below) use all of the identifiable municipalities, no matter whether they were sampled in both waves or one wave only.

To assess whether the evaluation sample is representative for the east of Ukraine prior to the war, we test for differences in means of both outcome and control variables across the included and excluded observations, separately for Donbas and the remainder of the east in the first wave. Appendix Tables 1.A2 and 1.A3 document the corresponding descriptive statistics and t-test results. We do not find any significant differences in outcome variables between the included and excluded Donbas cities. In the remainder of the east, the included cities were on average less pro-Ukrainian and more likely to support the country split than the excluded cities. Concerning socioeconomic characteristics, we observe some significant differences across the groups. On average, respondents in the excluded Donbas cities lived in larger households, reported better material conditions, and were more likely to have a spouse or a partner than respondents in the included cities. The comparison of the included and excluded cities in the rest of the east suggests that the former were on average significantly larger, with a higher share of university-educated respondents, native Russian speakers, and followers of the Kyiv Orthodox Church. We therefore bear in mind that in terms of pre-war outcome and control variables, our evaluation sample appears to be much more representative for Donbas than for the remainder of the east. However, the latter seems to be only a minor caveat in the context of our quantitative analysis, which focuses on the effect in the war-affected Donbas.

Our DiD approach outlined in Section 1.4 exploits variability in exposure to war across the country. Since the onset of the conflict, only some parts of Donetsk and Luhansk oblasts have experienced hostilities. Though the front line has moved, the regional capitals of Donetsk and Luhansk have permanently been in the conflict area,[29] which has led to the disruption of important services affecting the entire population in the region. The other government-held territories got exposed to the conflict through the inflow of people from the war area, the recruitment of troops, the discussions in politics, the media, and the civil society. This allows us to define two treatment groups: individuals in Donetsk and Luhansk oblasts (Donbas) form the high treatment group while

[28] The data do not contain any villages in Donbas (Donetsk and Luhansk oblasts) that appear in both waves, while only one village in Dnipropetrovsk oblast is present in both waves and accounts for 21 observations in total. To ensure comparability in municipality size across Donbas and the rest of the east, we therefore also drop villages in the remainder of the east from our evaluation sample.

[29] The maps of the front line for different dates of the conflict are provided by the Information Analysis Center of the National Security and Defense Council of Ukraine at `www.mediarnbo.org`.

individuals residing in the other eastern oblasts belong to the low treatment group.

Table 1.1 reports descriptive statistics and t-test results for political attitudes across the treatment groups in the two periods. Compared to the rest of the east, Donbas on average shows significantly lower preference for independent Ukraine, stronger support for one state with Russia, and weaker affinity with Ukraine in the first wave. In the second wave, Donbas appears significantly less sympathetic towards Ukraine and Ukrainians and more likely to opt for the country split than the rest of the east.

Table 1.1: Mean outcome values for Donbas and the remainder of the east

	2013				2015			
	Donbas	East	Difference	p-value	Donbas	East	Difference	p-value
Independent, neutral state (binary)	0.227	0.353	-0.127	0.036	0.429	0.449	-0.020	0.873
	(0.053)	(0.034)	(0.060)		(0.107)	(0.076)	(0.127)	
Union with Russia (binary)	0.473	0.407	0.066	0.500	0.117	0.101	0.016	0.784
	(0.094)	(0.040)	(0.098)		(0.060)	(0.011)	(0.058)	
Split into separate states (binary)	0.020	0.012	0.008	0.286	0.136	0.034	0.103	0.024
	(0.007)	(0.005)	(0.008)		(0.046)	(0.013)	(0.045)	
Fully the same cultures	4.805	4.007	0.798	0.138	3.634	4.120	-0.486	0.145
(1: fully disagree, ..., 7: fully agree)	(0.459)	(0.314)	(0.537)		(0.219)	(0.261)	(0.333)	
One state with Russia	4.804	3.665	1.139	0.052	2.448	2.458	-0.010	0.976
(1: fully disagree, ..., 7: fully agree)	(0.591)	(0.155)	(0.584)		(0.284)	(0.209)	(0.341)	
I love Ukraine	3.838	4.248	-0.410	0.001	3.836	4.423	-0.587	0.000
(1: definitely no, ..., 5: definitely yes)	(0.125)	(0.039)	(0.125)		(0.104)	(0.052)	(0.112)	
"We" for Ukrainians	3.905	3.956	-0.050	0.724	3.751	4.176	-0.425	0.044
(1: definitely no, ..., 5: definitely yes)	(0.128)	(0.075)	(0.143)		(0.194)	(0.103)	(0.211)	

Note: Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), while the rest of the east forms the low treatment group ('East'). Each variable is averaged over non-missing values. Standard errors are clustered at the city level and reported in parentheses.

In Table 1.2, we present descriptive statistics and t-test results for control variables. It is mostly municipalities with more than 50,000 inhabitants that contribute to our data. Females compose more than 50% of the sample and the average respondents' age is 45–47 years. The majority have at least secondary education, follow one of the Orthodox Churches, work, and are married or in partnership. The data suggest that respondents have on average resided at least 35 years in their actual municipalities. Concerning native language, Russian is more common than Ukrainian in either treatment group.[30] In both waves, we observe that the share of native Ukrainian speakers is significantly lower and the average age of respondents is significantly higher in Donbas than in the rest of the east. In 2013, Donbas also significantly differs by a higher retirement rate, a smaller average household size, and worse off material conditions. In 2015, the share of followers of the Kyiv Orthodox Church, the employment rate and residents' loyalty are significantly lower in Donbas compared to the rest of the east.

[30] Many respondents are bilingual, both native Ukrainian and Russian speakers. In 2013, the share of bilingual respondents is 0.373 in Donbas and 0.336 in the rest of the east ($p = 0.660$, two-sided t-test). In 2015, the corresponding shares are 0.507 and 0.435 ($p = 0.341$, two-sided t-test).

Table 1.2: Mean covariate values for Donbas and the remainder of the east

	2013				2015			
	Donbas	East	Difference	p-value	Donbas	East	Difference	p-value
City size: <50,000 citizens (binary)	0.235	0.147	0.088	0.579	0.199	0.099	0.100	0.493
	(0.148)	(0.071)	(0.158)		(0.143)	(0.053)	(0.146)	
Female (binary)	0.559	0.557	0.002	0.883	0.562	0.564	-0.002	0.946
	(0.013)	(0.007)	(0.014)		(0.028)	(0.008)	(0.028)	
Age	46.011	45.015	0.996	0.080	46.896	45.931	0.965	0.092
	(0.499)	(0.314)	(0.569)		(0.536)	(0.259)	(0.572)	
Secondary specialized education (binary)	0.419	0.376	0.043	0.354	0.413	0.349	0.064	0.120
	(0.044)	(0.020)	(0.046)		(0.038)	(0.020)	(0.041)	
University degree (binary)	0.346	0.426	-0.080	0.230	0.413	0.384	0.028	0.579
	(0.063)	(0.028)	(0.066)		(0.030)	(0.043)	(0.051)	
Native Ukrainian speaker (binary)	0.130	0.268	-0.138	0.025	0.060	0.241	-0.182	0.001
	(0.046)	(0.043)	(0.061)		(0.018)	(0.050)	(0.052)	
Native Russian speaker (binary)	0.395	0.371	0.025	0.775	0.383	0.261	0.122	0.181
	(0.079)	(0.042)	(0.086)		(0.073)	(0.059)	(0.091)	
Moscow Orthodox Church (binary)	0.262	0.169	0.094	0.200	0.290	0.171	0.119	0.355
	(0.068)	(0.034)	(0.073)		(0.133)	(0.022)	(0.129)	
Kyiv Orthodox Church (binary)	0.119	0.192	-0.073	0.215	0.073	0.205	-0.133	0.000
	(0.053)	(0.030)	(0.059)		(0.023)	(0.014)	(0.026)	
Orthodox Church (binary)	0.162	0.249	-0.086	0.198	0.378	0.337	0.042	0.715
	(0.054)	(0.043)	(0.067)		(0.115)	(0.030)	(0.114)	
Working (binary)	0.531	0.575	-0.045	0.173	0.483	0.536	-0.053	0.073
	(0.027)	(0.020)	(0.033)		(0.027)	(0.015)	(0.030)	
Retired (binary)	0.279	0.243	0.037	0.101	0.303	0.269	0.035	0.110
	(0.018)	(0.014)	(0.022)		(0.021)	(0.009)	(0.022)	
Single (binary)	0.162	0.169	-0.007	0.803	0.169	0.133	0.036	0.225
	(0.026)	(0.016)	(0.030)		(0.024)	(0.018)	(0.029)	
Married or in partnership (binary)	0.609	0.636	-0.027	0.604	0.542	0.633	-0.091	0.220
	(0.048)	(0.023)	(0.051)		(0.076)	(0.014)	(0.074)	
Household size	2.556	2.800	-0.244	0.034	2.642	2.744	-0.102	0.216
	(0.094)	(0.072)	(0.115)		(0.065)	(0.054)	(0.082)	
Material conditions	4.648	4.159	0.489	0.014	4.851	5.146	-0.295	0.249
(1: very good, ..., 7: terrible)	(0.192)	(0.078)	(0.199)		(0.251)	(0.093)	(0.256)	
Years in the current municipality	38.685	36.416	2.269	0.292	35.169	37.713	-2.544	0.065
	(2.171)	(0.591)	(2.152)		(1.220)	(0.742)	(1.377)	

Note: Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), while the rest of the east forms the low treatment group ('East'). Each variable is averaged over non-missing values. Standard errors are clustered at the city level and reported in parentheses.

1.4 Econometric methods and identification

We use two econometric approaches to infer the effect of the war on political preferences and sentiments of the two treatment groups with high and low exposure to the conflict in eastern Ukraine as outlined in Section 1.3. First, we apply a before-after analysis to examine changes in attitudes for each of the treatment groups over time conditional on a set of controls, i.e. we compare the outcome variables after the outbreak of the war to those prior to the war. This approach relies on the assumption that there are no time trends in unobservables affecting the outcomes of interest between the survey waves 2013 and 2015, at least after controlling for the socioeconomic variables discussed in the previous section. Considering a linear model, the equation to be estimated takes the following form:

$$Y = \beta_0 + 1_{\{year2015\}}\beta_1 + \mathbf{X}'\beta_2 + U, \tag{1.1}$$

where Y is the outcome, $1_{\{year2015\}}$ is an indicator for the year 2015, $\mathbf{X}$ is a vector of the control variables defined in Section 1.3, and U is the error term. Coefficient β_1 yields the war effect on the outcome of interest. However, if the assumption of no time trends in unobservables (e.g. industry-specific economic relations with Russia) is violated, we cannot separate the war effect from the trend using the before-after comparison, because $1_{\{year2015\}}$ is correlated with U even conditional on $\mathbf{X}$. Albeit we suspect time trends in political preferences and sentiments to be rather negligible given our time window of just two years, we cannot rule them out entirely.

Second, we use a difference-in-differences (DiD) approach to compare the change in attitudes for the high treatment group relative to the low treatment group before and after the outbreak of the war. DiD estimation conventionally assumes that (i) both a treatment and a control group, with the latter not being exposed to the treatment at all, are available, and that (ii) the (hypothetical) average outcomes of the treated and control groups, if neither group had actually received the treatment, would follow a common time trend while their levels may differ across groups. As discussed in Lechner (2011), this is, for instance, satisfied if the effects of unobservables on the outcome, which differ across treatment groups, are constant over time. In our empirical context, however, the assumption of a proper control group with zero treatment intensity does not seem to hold because the entire country is affected by the conflict in Donbas through the economic and social consequences, the recruitment of troops, the discussions in politics, the media, and the civil society.

Adding to the conventional common trend assumption, Fricke (2017) suggests further restrictions which allow comparing high and low (rather than zero) treatment groups over time. His approach identifies a lower bound on the absolute magnitude of effect of a high treatment vs. no treatment in the high treatment group even if the no treatment case is not observed in the data. The identifying assumptions imply that (i) the treatment affects the high and low treatment groups in the same direction compared to no treatment, and that (ii) the effect of the high treatment in the high treatment group is in absolute terms stronger than the effect of the low treatment in the low treatment group.[31] Assuming linearity, DiD estimation is based on the following regression model:

$$Y = \beta_0 + 1_{\{year2015\}}\beta_1 + 1_{\{Donbas\}}\beta_2 + 1_{\{Donbas*year2015\}}\beta_3 + \mathbf{X}'\beta_4 + U, \tag{1.2}$$

where Y is the outcome of interest, $1_{\{year2015\}}$ is an indicator for the year 2015, $1_{\{Donbas\}}$ is an indicator for residing in either Donetsk or Luhansk oblast, $1_{\{Donbas*year2015\}}$ is the interaction of the indicators, $\mathbf{X}$ is a vector of the control variables, and U is the error term. Under the stated conditions, coefficient β_3 corresponds to the lower bound of the war effect.

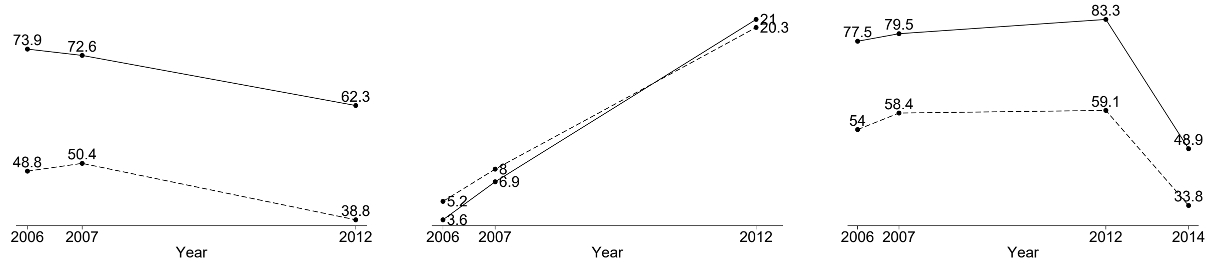
The interpretation of the DiD results presented in Section 1.5 crucially depends on whether the

[31] Fricke (2017) also demonstrates that the conventional common trend assumption defined upon treatment vs. no treatment does not even allow point identifying the effect of a high vs. a low treatment, unless effect homogeneity across treatment groups is assumed.

mentioned identifying assumptions are satisfied in our empirical context. Given the comparably strong cultural coherence of the region considered (see, for example, Arel, 2006), assuming common trends in political attitudes toward Russia and Ukraine in the absence of war appears quite plausible, at least after controlling for important socioeconomic factors defined in Section 1.3.[32] To gauge the plausibility of the common trend assumption, Figure 1.2 compares the vote shares for two pro-Russian parties — the Party of Regions (PR) and the Communist Party of Ukraine (CPU) — in the parliamentary elections in Donbas and the rest of the east before and after the conflict outbreak. While we observe nearly parallel trends in the vote shares in the two areas prior to the conflict, there is a marked divergence in trends of 9.1 percentage points afterwards.[33]

Figure 1.2: Dynamics of regional vote shares (%) in the parliamentary elections 2006–2014

(a) Party of Regions (b) Communist Party of Ukraine (c) Pro-Russian parties jointly



Note: The solid and dashed lines present Donbas and the rest of the east, respectively. In 2014, the pro-Russian parties include the Opposition Bloc and the Communist Party of Ukraine. Vote shares are calculated using the official data of the Central Election Commission, `www.cvk.gov.ua`, retrieved 18 June 2019.

In Section 1.6, we further examine the credibility of the common trend assumption using a placebo test, which is an alternative to testing parallel trends in previous periods. Besides, we also believe that the additional restrictions of Fricke (2017) are satisfied. First, treatment intensity can likely be ordered as a function of exposure to war. Second, the effect of war on political attitudes towards Russia and Ukraine should, at least on average, have the same sign across subregions of eastern Ukraine.[34]

[32]Nevitte et al. (2009) document that, for instance, age, education, and household income positively correlate with voting behavior. We control for these and even more factors in our regression models.

[33]Both parties originated from the Donbas region (Arel, 2006). Though the PR did not participate in the 2014 Ukrainian parliamentary elections, its successor — the Opposition Bloc (OB) — won 29 out of 450 seats, predominantly in eastern Ukraine. Despite electoral support in eastern Ukraine, the CPU did not secure 5% votes nationwide to win seats in the parliament. Compared to 2012, the joint vote share of the pro-Russian parties dropped by 34.4 percentage points in the unoccupied part of Donbas and by 25.3 percentage points in the rest of the east. This suggests that political capital of Russia has declined in eastern Ukraine, particularly in Donbas, after to the outbreak of the war.

[34]Inspecting the before-after estimates in the different specifications presented in Section 1.5 reveals that in most cases the sign of the estimates is the same for the high and low treatment groups, apart from a few cases with insignificant results in at least one group. At the same time, the absolute magnitudes of the estimates are mostly

We estimate equations (1.1) and (1.2) based on OLS, which linearly controls for differences in observed characteristics. As the linearity assumptions may, however, be violated in reality, we also consider a semiparametric approach. The latter is based on inverse probability weighting by the propensity score, i.e. the probability of receiving the treatment conditional on the observed covariates (see Horvitz and Thompson, 1952; Hirano et al., 2003; Abadie, 2005, in the context of DiD estimation). In the first step, we estimate the treatment propensity score by logistic regression. In the second step, we reweight observations (i) in the high treatment group before the war, (ii) in the low treatment group before the war, and (iii) in the low treatment group after the outbreak of the war to match the distribution of covariates in the high treatment group after the outbreak of the war. Finally, we take differences in the mean differences of the reweighted outcomes within treatment groups over time. Concerning inference, we use a block bootstrap procedure (with 999 replications) that accounts for clustering at the municipal level and estimates the standard errors to be used in the t-statistics based on the bootstrap distributions of the resampled parametric and semiparametric effect estimates.

At this point, it is important to acknowledge that several of our control variables (e.g. employment status, material conditions, household size, and residence) are likely to be affected by the conflict and thus endogenous. For this reason, the next section also presents before-after and DiD estimates without controlling for $\mathbf{X}$. Despite the obvious trade-off between dropping covariates at the risk of omitted variable bias and including covariates at the risk of endogeneity (or selection) bias, the obtained estimates are often not significantly affected by the choice of $\mathbf{X}$. The same applies to using only a subset of the observables.[35]

A final concern for our econometric approaches is sample selection due to migration movements between the two waves that may importantly affect the composition of the population in the area. That is, individuals with particular characteristics might have been more likely to migrate. For instance, Coupé and Obrizan (2016b) report that in their sample coming from two cities in Donetsk, those who stayed and those who temporarily left the cities during military confrontations differed by age, education level, religiousness and ability to speak both Ukrainian and Russian. In our models, we control for these and even more characteristics, but sample selection bias in particular related to unobservables might nevertheless be an issue. For instance, pro-Russian individuals could have left the government-held territory for Russia or the occupied areas, while pro-Ukrainian individuals might have done vice versa or migrated further away from the front line within the government-held areas. A specific threat is that the war has induced mass migration out of Donbas such that the residents in 2015 are not representative of those in 2013 anymore, which would jeopardize both our before-after and DiD analyses.

To judge the relevance of such issues, Table 1.3 reports the net internal migration in five oblasts

larger in the high treatment group. If time trends were absent such that before-after estimation was unbiased, this would provide empirical evidence in favor of the additional restrictions in Fricke (2017).

[35] These results are not reported but available on request.

of eastern Ukraine over 2012–2015, as provided by the Statistics Department of Ukraine based on administrative data on the change of permanent residence.[36] While net migration in Donbas — Donetsk and Luhansk oblasts — has already been negative in 2012 and 2013, net outmigration roughly doubled in the war year of 2014. Quite the contrary, net migration to the remainder of the east considerably increased in that year, most likely due to incoming migrants from Donbas. In 2015, all oblasts — except for Kharkiv oblast — experienced net outmigration. Even though we do see a noticeable change in net migration patterns during the war, the movements out of Donbas, for instance, appear moderate compared to the entire population of the area (roughly 6.6 million in 2013, including both government-help and occupied territories).[37]

Table 1.3: Net internal migration (persons) in eastern Ukraine over 2012–2015

Oblasts	2012	2013	2014	2015
Dnipropetrovsk	-1,564	-2,169	431	-1,351
Donetsk	-4,449	-4,516	-10,677	-9,239
Kharkiv	1,984	1,741	8,261	4,981
Luhansk	-4,034	-4,365	-8,120	-5,634
Zaporizhia	-1,361	-1,916	-847	-797

Source: State Statistics Service of Ukraine, `www.ukrstat.gov.ua`, retrieved 12 November 2018.

As an important caveat, however, the statistics do not cover migrants or refugees that did not register in their destination municipality. For this reason, we also consider information from the Interagency Headquarters, a governmental agency that provides estimates for the number of the internally displaced persons (IDPs) as a consequence of the events in Crimea and Donbas. As of 31 March 2015, it is claimed that there were 810,060 IDPs in total, including 789,670 from Donbas,[38] which is a much higher figure than the administrative records indicate. As shown in Figure 1.3, about 69% of IDPs were accommodated in eastern Ukraine: Kharkiv oblast hosted the highest number of IDPs, and nearly 32% moved to the government-held parts of Donbas. Hence, many IDPs relocated within the Donbas region. In addition to internal migration, many Ukrainians also fled to Russia, starting from 2014. Table 1.4 illustrates a dramatic increase in the number of Ukrainians registered with the legal status of refugee or temporary asylum in Russia which reached 311,407 at the end of 2015.

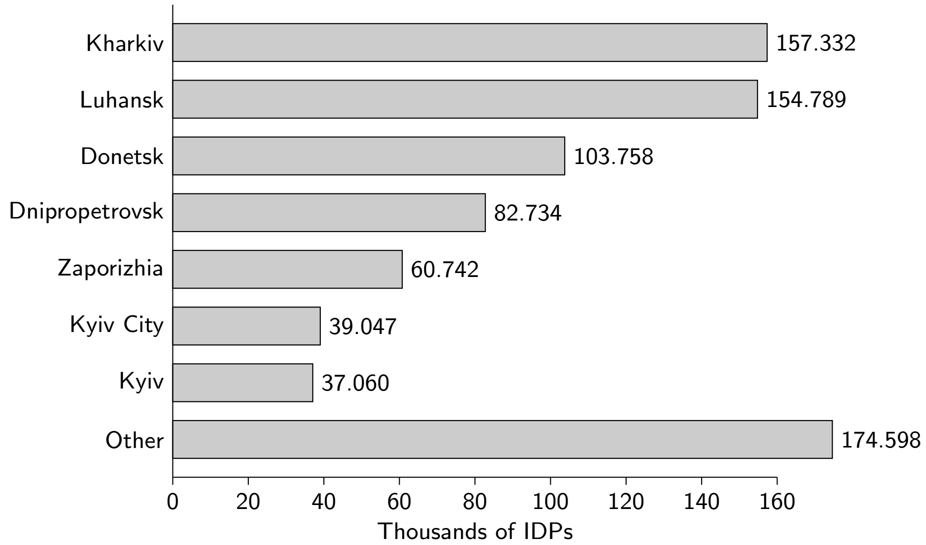
In the light of these migration flows, one could on the one hand argue that migration might create attenuation bias in the main findings of our analysis, namely the negative DiD effects on

[36]The regional Departments of Statistics provided us with the internal migration data at the city level but for the cities in Donetsk oblast the 2013 archive was left in Donetsk and thus currently inaccessible.

[37]The average population in Donetsk and Luhansk oblasts is provided by the State Statistics Service of Ukraine, `www.ukrstat.gov.ua`, retrieved 23 April 2016.

[38]This information is published at the governmental portal, `https://www.kmu.gov.ua/ua/news/248052970`, retrieved 12 November 2018. The "Ukraine Migration Profile 2010–2014" with the numbers of IDPs for all oblasts as of 31 December 2014 is available at: `https://dmsu.gov.ua/assets/files/mig_profil/profile_2015_en.pdf`, retrieved 12 November 2018.

Figure 1.3: Internally displaced persons by destination region as of 31 March 2015



Source: Governmental Portal, `www.kmu.gov.ua`, retrieved 12 November 2018.

Table 1.4: Ukrainians registered with refugee or temporary asylum status in Russia over 2012–2015

Status	2012	2013	2014	2015
Refugee	5	5	229	273
Temporary asylum	0	0	231,558	311,134
Total	5	5	231,787	311,407

Source: General Administration for Migration Issues of the Interior Ministry of Russia, `http://xn--b1ab2a0a.xn--b1aew.xn--p1ai/about/activity/stats/Statistics/Predostavlenie_ubezhishha_v_Rossijskoj`, retrieved 4 May 2016.

political attitudes towards Russia. If pro-Ukrainian individuals left the war-affected areas to resettle somewhere else in eastern Ukraine, the drop in the support of Russia would be overestimated in the low treatment group (either through the migrants themselves or through interaction effects with locals) and underestimated in the high treatment group. This would squeeze the absolute magnitude of the negative DiD estimates. On the other hand, this could be countervailed by pro-Russian individuals leaving the war-affected area for territories outside the control of the Ukrainian government, in particular Russia. In this context, it is worth noting that the figures suggest that migration to Russia was lower than to other Ukrainian regions, such that we suspect attenuation bias to be the more relevant threat. Similarly to our identifying assumptions outlined above, this suggests that our main DiD effects on attitudes towards Russia are in absolute magnitude lower bounds for the true effects. Likewise, the resulting attenuation bias could mask positive effects on the pro-Ukrainian sentiments. Again, we notice that our estimates are not very sensitive to the choice of the socioeconomic characteristics, which one would suspect to correlate with migration decisions. Consequently, our data do not provide evidence that the observed changes in the population composition, potentially due to the war-induced migration, systematically influence our results.

1.5 Main results

Table 1.5 provides the estimates for our main specification, namely when we distinguish between Donbas and the remainder of the east as the high and low treatment groups, and when our evaluation sample consists of the cities observed in both waves. We present the results for before-after and DiD estimations based on (i) unconditional mean differences (columns 1–3) obtained by excluding controls in equations (1.1) and (1.2), (ii) OLS controlling for the socioeconomic factors described in Section 1.3 (columns 4–6), and (iii) propensity score weighting (columns 7–9). This entails altogether nine different estimators which produce the before-after estimates for the high and low treatment groups separately (columns 'Donbas' and 'East', respectively) as well as the DiD estimates (columns 'DiD').

OLS and weighting deliver in general quite similar results in terms of effect directions and magnitudes. We mostly find no statistically significant shift in the sympathy towards Ukraine and self-association with Ukrainians, except for the positive and statistically significant effects based on OLS and estimation without controls in the rest of the east. In contrast, several statistically significant negative associations are found between exposure to war and political attitudes towards Russia, which is somewhat in line with the finding of Coupé and Obrizan (2016b) that the experience of property damage decreases the support for the view that the Ukrainian government should compromise with Russia. For instance, Donbas' perception of the Ukrainian and Russian cultures as fully the same significantly weakened over time, while no significant change is observed in the rest of the east. Hence, the corresponding DiD effects based on the inter-group changes over time are significantly negative.

Furthermore, the preference for a union with Russia significantly declined by more than 30 percentage points in either treatment group. Though the before-after estimates are larger for Donbas, none of the corresponding DiD estimates is statistically significant. We observe a very similar pattern with respect to the preference for one state with Russia. The share of supporters of independent and neutral Ukraine increased over time, yet the estimated effects are mostly insignificant across the treatment groups. In line with Coupé and Obrizan (2016b), the preference for the country split into separate states appears to have increased in Donbas, but only the unconditional estimation yields a significant DiD estimate. Note that in Donbas the preference for the country split is strongly and highly significantly correlated with the view that the conflict can be resolved by ceding the occupied territories ($\rho = 0.365$), which might be driven by the respondents' desire to achieve peace.[39]

[39] Respondents were asked in 2015 which solution to the situation in Donetsk and Luhansk oblasts corresponded to their position. In Donbas, about 5% opted for conducting anti-terrorist operations until the complete liquidation

Table 1.5: Estimates for Donbas and the remainder of the east

	No controls			OLS			Propensity score weighting		
	Donbas	East	DiD	Donbas	East	DiD	Donbas	East	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.202	0.095	0.106	0.243	0.049	0.195	0.235	0.181	0.054
st. error	0.160	0.077	0.178	0.136	0.078	0.156	0.154	0.095	0.171
p-value	0.207	0.218	0.549	0.073	0.535	0.213	0.127	0.058	0.753
observations	304	1399	1703	280	1317	1597	280	1317	1597
Union with Russia	-0.356	-0.306	-0.050	-0.379	-0.310	-0.069	-0.433	-0.335	-0.098
st. error	0.090	0.042	0.098	0.098	0.039	0.103	0.115	0.088	0.141
p-value	0.000	0.000	0.608	0.000	0.000	0.500	0.000	0.000	0.487
observations	304	1399	1703	280	1317	1597	280	1317	1597
Split into separate states	0.116	0.022	0.094	0.097	0.015	0.082	0.093	0.017	0.075
st. error	0.048	0.013	0.049	0.151	0.010	0.151	0.057	0.010	0.057
p-value	0.015	0.083	0.056	0.521	0.134	0.589	0.102	0.094	0.186
observations	304	1399	1703	280	1317	1597	280	1317	1597
Fully the same cultures	-1.171	0.113	-1.284	-1.301	0.043	-1.344	-1.437	0.211	-1.648
st. error	0.511	0.367	0.633	0.469	0.378	0.602	0.582	0.499	0.867
p-value	0.022	0.758	0.043	0.006	0.909	0.026	0.014	0.672	0.057
observations	363	1619	1982	336	1506	1842	336	1506	1842
One state with Russia	-2.356	-1.206	-1.149	-2.586	-1.166	-1.420	-2.810	-0.999	-1.811
st. error	0.715	0.264	0.748	0.724	0.277	0.780	0.928	0.533	1.150
p-value	0.001	0.000	0.124	0.000	0.000	0.069	0.002	0.061	0.115
observations	321	1520	1841	293	1417	1710	293	1417	1710
I love Ukraine	-0.002	0.175	-0.178	0.108	0.205	-0.097	0.162	0.141	0.021
st. error	0.225	0.061	0.233	0.260	0.064	0.270	0.294	0.119	0.331
p-value	0.992	0.004	0.447	0.678	0.001	0.720	0.583	0.237	0.950
observations	380	1694	2074	350	1571	1921	350	1571	1921
"We" for Ukrainians	-0.154	0.221	-0.375	0.017	0.208	-0.192	0.005	0.161	-0.156
st. error	0.285	0.085	0.296	0.235	0.106	0.255	0.273	0.182	0.309
p-value	0.589	0.009	0.206	0.944	0.050	0.452	0.985	0.375	0.614
observations	380	1694	2074	350	1571	1921	350	1571	1921

Note: Sample includes the cities observed in both waves. Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), the rest of the east forms the low treatment group ('East'). Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

To sum up, both the before-after and DiD estimates suggest that the political attitudes towards Russia have deteriorated as a consequence of the war, while this is not the case for sentiments towards Ukraine at least in Donbas. Compared to the rest of the east, the distinction between the Ukrainian and Russian cultures significantly increased in Donbas. The frequency distributions of the non-binary outcomes for Donbas over time also reveal a negative shift in sentiments towards Russia but no pronounced change with respect to Ukraine (see Appendix Figure 1.A1).

These results echo a sharp decline in the pro-Russian vote of 34.4 percentage points in Donbas and 25.3 percentage points in the rest of the east between the parliamentary elections 2012 and 2014 (see Figure 1.2). A possible reason is that people may believe that Russia has contributed to the conflict escalation in Donbas. Indeed, opinions that Russia is involved in the war and that the ongoing conflict is a war with Russia negatively correlate with attitudes towards Russia but

of armed groups, 50% for negotiations and compromise, 10.5% for federalization of the country, and 13.4% for giving up fighting for these areas — can live as they want.

positively correlate with sentiments towards Ukraine (see Appendix Table 1.A4). A possible explanation is that respondents updated their perceptions of Russia during the conflict, for instance, through personal experience, communication, or a change in exposure to information and/or disinformation. Information received from mass media, social networks and/or personal experiences (trauma) could have a learning or persuasion effect on respondents. Note that the presentation of events in Donbas (and Crimea) by Ukrainian and Russian media went opposite directions, and certain Russian media were banned in the unoccupied territory of Ukraine shortly after the conflict outbreak. There is in fact empirical evidence suggesting that media exposure can affect political sentiments expressed in voting behavior (e.g. DellaVigna and Kaplan, 2007; Gerber et al., 2009; Enikolopov et al., 2011; DellaVigna et al., 2014). Hence, media are a plausible mediator of changing attitudes towards Russia in eastern Ukraine, at least in the government-controlled areas.

Since the majority of respondents in our evaluation sample report either Russian or both Ukrainian and Russian as their native language(s), we extend our main analysis to investigate effect heterogeneity for these two language groups. Appendix Tables 1.A5 and 1.A6 report the results obtained in the corresponding subsamples. The directional patterns of the estimated effects are generally comparable to the main results. In the bilingual subsample, however, we observe more pronounced positive effects on individual preferences for independent Ukraine and negative effects on sentiments towards Russia. The findings also suggest that bilingual respondents have weakened their support for one state with Russia significantly more in Donbas than in the rest of the east. In contrast, we generally find less significant negative effects on the attitudes towards Russia among native Russian speakers. Comparing the two language groups, conditional OLS yields significant effect heterogeneity in the before-after estimates for the union with Russia in the rest of the east ($p < 0.01$, two-sided test). As regards the country split, the share of its Russian-speaking supporters has significantly increased in Donbas compared to the rest of the east; effect heterogeneity is statistically significant for the before-after estimates in Donbas ($p < 0.05$, two-sided tests) and for the DiD results ($p < 0.1$, two-sided tests), except for conditional OLS. We find no evidence that native Russian speakers significantly changed their sentiments towards Ukraine or Ukrainians as a consequence of the war, and the DiD estimates remain insignificant in either subsample; the corresponding estimates are not significantly different across language groups ($p > 0.1$, two-sided tests).

We further examine effect heterogeneity across younger and older respondents by splitting our evaluation sample at the median age of 45 into two subsamples. Since the younger tend to be more mobile than the older, exposure to war could have affected the two age groups differentially. The results displayed in Appendix Tables 1.A7 and 1.A8 are generally in line with those in the total sample. For instance, we observe similar patterns in terms of sentiments towards Ukraine and Ukrainians, i.e. no significant DiD estimates in either age group. However, the decline in the pro-Russian support is less significant in the younger subsample; the unconditional before-after estimates for the union with Russia yield significant effect heterogeneity in either treatment group

($p < 0.1$, two-sided tests), and the conditional OLS estimates for the same scenario in the rest of the east ($p < 0.05$, two-sided test). In addition, there is no evidence of any significant changes in the preference for the country split among the older respondents; we find a weakly significant effect heterogeneity in the unconditional before-after estimates for the rest of the east ($p < 0.1$, two-sided test).

We also observe some attitude measures with missing values in our sample. Since, due to the war, respondents in Donbas could have become less likely to answer politically sensitive questions than respondents in the rest of the east, we examine whether item non-response in the outcome variables is selective in our sample, i.e. varies systematically over time and across treatment groups. For this purpose, we create missing dummies for our outcomes of interest and use them as dependent variables in our estimators. We generally find no evidence for selective item non-response within treatment groups over time or across time trends of treatment groups (see Appendix Table 1.A9).

1.6 Robustness checks

We run several sensitivity checks to investigate the robustness of our results with respect to the definition of the treatment groups.

One concern regarding our analysis is that we only consider the remainder of the east as the low treatment group. Though ethnically less Russian than the east, the south of Ukraine is often also perceived as pro-Russian in view of its historical ties with Russia, proportion of Russian-speaking population, and support for pro-Russian parties in elections. To address this concern, we extend our evaluation sample to three southern oblasts: Kherson, Mykolaiv and Odessa. Along with the rest of the east, the south now belongs to the low treatment group, while Donbas remains the high treatment group as in the main specification. The results in Table 1.6 indicate that the before-after estimates for the low treatment group are only marginally affected. We again observe positive tendencies in the low treatment group with respect to the support for independent Ukraine and connectedness with Ukraine and Ukrainians, but the DiD estimates stay insignificant. The results also suggest that the support for the country split has significantly increased in either treatment group, but no strong evidence is found for inter-group changes over time. The negative DiD estimates on the similarity of the Ukrainian and Russian cultures are significant at the 5% level, and the negative before-after estimates for a union or one state with Russia are significant at the 1% level.

Because of incomplete regional overlap between the waves, we have so far focused on the cities observed in both waves to control for municipality-specific unobservables. We now additionally check whether the dropped observations have an impact on the estimates when including them into the evaluation sample.[40] Table 1.7 documents a significant increase in the support for independent

[40] Whenever the evaluation sample includes cities and villages, we additionally control for a village dummy in

Table 1.6: Estimates using the extended sample with the south

	No controls			OLS			Propensity score weighting		
	Donbas	East & South	DiD	Donbas	East & South	DiD	Donbas	East & South	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.202	0.125	0.077	0.243	0.073	0.170	0.235	0.182	0.053
st. error	0.158	0.062	0.168	0.126	0.057	0.140	0.155	0.071	0.165
p-value	0.202	0.043	0.649	0.053	0.201	0.224	0.129	0.010	0.748
observations	304	2035	2339	280	1930	2210	280	1930	2210
Union with Russia	-0.356	-0.315	-0.042	-0.379	-0.313	-0.066	-0.433	-0.333	-0.099
st. error	0.092	0.032	0.097	0.099	0.034	0.106	0.120	0.058	0.128
p-value	0.000	0.000	0.670	0.000	0.000	0.533	0.000	0.000	0.438
observations	304	2035	2339	280	1930	2210	280	1930	2210
Split into separate states	0.116	0.033	0.083	0.097	0.031	0.066	0.093	0.042	0.051
st. error	0.046	0.012	0.048	0.047	0.012	0.049	0.051	0.024	0.059
p-value	0.012	0.005	0.083	0.041	0.009	0.174	0.071	0.082	0.388
observations	304	2035	2339	280	1930	2210	280	1930	2210
Fully the same cultures	-1.171	0.240	-1.411	-1.301	0.088	-1.389	-1.437	0.297	-1.734
st. error	0.496	0.268	0.561	0.517	0.276	0.572	0.588	0.401	0.816
p-value	0.018	0.369	0.012	0.012	0.749	0.015	0.015	0.458	0.033
observations	363	2328	2691	336	2182	2518	336	2182	2518
One state with Russia	-2.356	-1.209	-1.146	-2.586	-1.182	-1.404	-2.810	-1.071	-1.739
st. error	0.753	0.196	0.782	0.913	0.220	0.943	0.930	0.373	1.085
p-value	0.002	0.000	0.143	0.005	0.000	0.136	0.003	0.004	0.109
observations	321	2168	2489	293	2033	2326	293	2033	2326
I love Ukraine	-0.002	0.149	-0.151	0.108	0.131	-0.023	0.162	0.052	0.109
st. error	0.222	0.061	0.228	0.270	0.071	0.281	0.293	0.109	0.293
p-value	0.992	0.015	0.509	0.689	0.064	0.935	0.582	0.630	0.710
observations	380	2469	2849	350	2309	2659	350	2309	2659
"We" for Ukrainians	-0.154	0.049	-0.203	0.017	0.019	-0.002	0.005	-0.093	0.098
st. error	0.288	0.081	0.297	0.241	0.090	0.253	0.294	0.161	0.309
p-value	0.593	0.543	0.495	0.945	0.836	0.994	0.986	0.564	0.751
observations	380	2469	2849	350	2309	2659	350	2309	2659

Note: Sample includes the cities observed in both waves. Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), the rest of the east and the south form the low treatment group ('East & South'). Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

and neutral Ukraine in Donbas over time. We now find that positive sentiments towards Ukraine have significantly increased in the rest of the east compared to Donbas, where the corresponding before-after estimates remain insignificant. Furthermore, the significant growth in the preference for the country split in Donbas translates into the highly significant DiD estimates. We again confirm that either treatment group has significantly diminished its support for a union or one state with Russia. This time, no significant effects are found for the opinion that the Ukrainian and Russian cultures are the same. Considering the subsample of cities only, we find some evidence that the distinction between the two cultures has significantly increased in Donbas over time and that the preference for a union with Russia has deteriorated significantly more in Donbas (see Appendix Table 1.A10).

We then further extend our analysis to the whole Ukraine to examine whether the main results hold in the nationwide sample that includes all identifiable municipalities in 24 oblasts.[41] The high

conditional regression analyses.

[41]Four villages (41 observations) are dropped in 2013 because they cannot be uniquely identified by name only and further information is not available. Crimea, including the city of Sevastopol, is only observed in 2013 and

Table 1.7: Estimates using all observations in eastern Ukraine

	No controls			OLS			Propensity score weighting		
	Donbas	East	DiD	Donbas	East	DiD	Donbas	East	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.199	0.125	0.074	0.193	0.101	0.091	0.217	0.190	0.027
st. error	0.088	0.064	0.106	0.090	0.065	0.108	0.096	0.081	0.118
p-value	0.024	0.051	0.485	0.033	0.116	0.399	0.024	0.019	0.817
observations	945	1908	2853	870	1799	2669	870	1799	2669
Union with Russia	-0.402	-0.316	-0.085	-0.403	-0.317	-0.086	-0.432	-0.325	-0.107
st. error	0.059	0.033	0.067	0.043	0.032	0.054	0.050	0.060	0.081
p-value	0.000	0.000	0.199	0.000	0.000	0.110	0.000	0.000	0.187
observations	945	1908	2853	870	1799	2669	870	1799	2669
Split into separate states	0.126	0.014	0.112	0.126	0.007	0.119	0.116	0.015	0.101
st. error	0.036	0.012	0.038	0.035	0.009	0.037	0.033	0.012	0.037
p-value	0.000	0.241	0.003	0.000	0.450	0.001	0.000	0.216	0.006
observations	945	1908	2853	870	1799	2669	870	1799	2669
Fully the same cultures	-0.477	-0.059	-0.418	-0.462	-0.100	-0.362	-0.581	0.107	-0.689
st. error	0.299	0.282	0.406	0.320	0.290	0.425	0.362	0.407	0.581
p-value	0.110	0.835	0.303	0.149	0.731	0.395	0.108	0.792	0.236
observations	1120	2204	3324	1032	2056	3088	1032	2056	3088
One state with Russia	-1.451	-1.355	-0.095	-1.614	-1.311	-0.302	-1.799	-1.089	-0.709
st. error	0.383	0.217	0.450	0.418	0.244	0.484	0.467	0.396	0.654
p-value	0.000	0.000	0.833	0.000	0.000	0.532	0.000	0.006	0.278
observations	1041	2074	3115	956	1936	2892	956	1936	2892
I love Ukraine	-0.171	0.172	-0.343	-0.174	0.176	-0.350	-0.181	0.102	-0.283
st. error	0.135	0.054	0.145	0.132	0.054	0.141	0.143	0.089	0.171
p-value	0.205	0.001	0.018	0.187	0.001	0.013	0.205	0.254	0.097
observations	1158	2303	3461	1064	2144	3208	1064	2144	3208
"We" for Ukrainians	-0.035	0.179	-0.215	-0.006	0.130	-0.136	-0.035	0.044	-0.079
st. error	0.196	0.074	0.209	0.179	0.083	0.193	0.197	0.123	0.229
p-value	0.858	0.016	0.304	0.972	0.119	0.481	0.860	0.722	0.732
observations	1158	2303	3461	1064	2144	3208	1064	2144	3208

Note: Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), the rest of the east form the low treatment group ('East'). Propensity scores are estimated by logistic regression. We additionally control for a village dummy in the conditional regression analyses. Standard errors are clustered at the city level using 999 bootstrap replications.

treatment group remains the same as in the previous robustness check, while the low treatment group includes the rest of Ukraine. The findings in Table 1.8 suggest that love towards Ukraine has significantly increased in the latter compared to the former. Furthermore, the DiD estimates in the unconditional and conditional OLS analyses show that the increase in the share of respondents opting for independent Ukraine is significantly greater in Donbas than in the rest of Ukraine. Similar to the earlier results, the preference for the country split has grown significantly more in Donbas. Again, the before-after estimates concerning a union or one state with Russia are significantly negative at the 1% level. For a union with Russia, the negative DiD results are also highly significant. Unlike in the main specification, we find insignificant DiD estimates on the similarity between Ukrainian and Russian cultures as the negative time trends in either treatment

therefore also excluded from the analysis. The inclusion of Crimea does not change the results that are not reported but available on request.

group are not too different.

Table 1.8: Estimates using the nationwide sample

	No controls			OLS			Propensity score weighting		
	Donbas	Other	DiD	Donbas	Other	DiD	Donbas	Other	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.199	0.007	0.192	0.193	-0.009	0.202	0.217	0.168	0.049
st. error	0.088	0.025	0.091	0.089	0.024	0.092	0.095	0.042	0.103
p-value	0.023	0.774	0.036	0.031	0.704	0.029	0.023	0.000	0.634
observations	945	8880	9825	870	8545	9415	870	8545	9415
Union with Russia	-0.402	-0.190	-0.212	-0.403	-0.184	-0.220	-0.432	-0.301	-0.130
st. error	0.053	0.017	0.056	0.043	0.018	0.047	0.049	0.035	0.060
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030
observations	945	8880	9825	870	8545	9415	870	8545	9415
Split into separate states	0.126	0.010	0.117	0.126	0.011	0.115	0.116	0.026	0.090
st. error	0.036	0.006	0.036	0.038	0.006	0.038	0.035	0.010	0.038
p-value	0.000	0.097	0.001	0.001	0.080	0.003	0.001	0.013	0.017
observations	945	8880	9825	870	8545	9415	870	8545	9415
Fully the same cultures	-0.477	-0.406	-0.071	-0.462	-0.403	-0.059	-0.581	-0.057	-0.525
st. error	0.306	0.111	0.323	0.332	0.098	0.346	0.379	0.190	0.451
p-value	0.119	0.000	0.827	0.165	0.000	0.866	0.125	0.766	0.245
observations	1120	9997	11117	1032	9559	10591	1032	9559	10591
One state with Russia	-1.451	-1.062	-0.389	-1.614	-0.998	-0.616	-1.799	-1.158	-0.640
st. error	0.401	0.089	0.415	0.406	0.089	0.415	0.453	0.202	0.541
p-value	0.000	0.000	0.349	0.000	0.000	0.138	0.000	0.000	0.236
observations	1041	9639	10680	956	9221	10177	956	9221	10177
I love Ukraine	-0.171	0.182	-0.353	-0.174	0.183	-0.357	-0.181	0.145	-0.326
st. error	0.134	0.025	0.135	0.131	0.025	0.133	0.144	0.059	0.155
p-value	0.202	0.000	0.009	0.184	0.000	0.007	0.208	0.015	0.036
observations	1158	10495	11653	1064	10014	11078	1064	10014	11078
"We" for Ukrainians	-0.035	0.043	-0.078	-0.006	0.015	-0.022	-0.035	0.014	-0.049
st. error	0.197	0.039	0.200	0.181	0.043	0.187	0.200	0.077	0.214
p-value	0.858	0.268	0.695	0.972	0.721	0.908	0.862	0.855	0.819
observations	1158	10495	11653	1064	10014	11078	1064	10014	11078

Note: Sample excludes Crimea and four villages from other regions in 2013. Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), the rest of Ukraine forms the low treatment group ('Other'). Propensity scores are estimated by logistic regression. We additionally control for a village dummy in the conditional regression analyses. Standard errors are clustered at the city level using 999 bootstrap replications.

According to our online search and examination of the front line maps provided by the Information Analysis Center of the National Security and Defense Council of Ukraine, four Donetsk cities in our main sample — Artemivsk, Kostiantynivka, Kramatorsk and Mariupol — directly experienced military confrontations in 2014. This allows us to distinguish between the high and low treatment groups based on reported hostilities. The estimates in Table 1.9 are mostly in line with the results of the main specification, but generally less significant. In particular, while the negative before-after effects on the cultural similarity are significant in the high treatment group, the corresponding DiD estimates are quite imprecise, albeit similar in magnitude. For the high treatment group, the negative effects on attitudes towards Russia are larger than in the main specification.

We conclude that our robustness checks by and large confirm the findings of the main specification. Namely, the political capital of Russia has decreased in the government-held territories

Table 1.9: Estimates for military confrontations

	No controls			OLS			Propensity score weighting		
	Battle	No battle	DiD	Battle	No battle	DiD	Battle	No battle	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.105	0.117	-0.012	0.173	0.073	0.100	0.222	0.150	0.072
st. error	0.241	0.074	0.253	0.304	0.076	0.315	0.201	0.113	0.239
p-value	0.663	0.115	0.963	0.569	0.337	0.751	0.270	0.184	0.764
observations	207	1496	1703	194	1403	1597	194	1403	1597
Union with Russia	-0.426	-0.300	-0.125	-0.511	-0.307	-0.204	-0.580	-0.337	-0.243
st. error	0.146	0.040	0.151	0.993	0.036	0.993	0.246	0.105	0.279
p-value	0.004	0.000	0.407	0.607	0.000	0.837	0.018	0.001	0.385
observations	207	1496	1703	194	1403	1597	194	1403	1597
Split into separate states	0.167	0.021	0.146	0.180	0.014	0.166	0.124	0.011	0.113
st. error	0.047	0.012	0.048	0.749	0.009	0.748	0.240	0.017	0.237
p-value	0.000	0.090	0.003	0.810	0.129	0.824	0.604	0.528	0.633
observations	207	1496	1703	194	1403	1597	194	1403	1597
Fully the same cultures	-1.299	0.047	-1.346	-1.666	-0.030	-1.636	-1.464	0.070	-1.534
st. error	0.784	0.347	0.862	0.703	0.360	0.789	0.728	0.732	1.223
p-value	0.097	0.892	0.118	0.018	0.934	0.038	0.044	0.924	0.210
observations	256	1726	1982	239	1603	1842	239	1603	1842
One state with Russia	-3.094	-1.188	-1.906	-3.371	-1.185	-2.186	-3.415	-0.878	-2.537
st. error	0.913	0.245	0.933	3.099	0.266	3.119	1.207	0.729	1.599
p-value	0.001	0.000	0.041	0.277	0.000	0.483	0.005	0.228	0.113
observations	217	1624	1841	200	1510	1710	200	1510	1710
I love Ukraine	0.078	0.161	-0.083	0.135	0.205	-0.070	0.244	0.137	0.107
st. error	0.353	0.064	0.357	0.491	0.064	0.494	0.562	0.177	0.531
p-value	0.824	0.012	0.817	0.784	0.001	0.887	0.664	0.439	0.840
observations	270	1804	2074	252	1669	1921	252	1669	1921
"We" for Ukrainians	-0.139	0.199	-0.338	0.053	0.207	-0.154	-0.061	0.032	-0.093
st. error	0.312	0.088	0.325	0.263	0.103	0.281	0.444	0.196	0.490
p-value	0.657	0.024	0.298	0.840	0.045	0.582	0.890	0.871	0.849
observations	270	1804	2074	252	1669	1921	252	1669	1921

Note: Sample includes the cities observed in both waves. Artemivsk, Kostiantynivka, Kramatorsk and Mariupol — the cities with reported military confrontations in 2014 — form the high treatment group ('Battle'), the rest of observations form the low treatment group ('No battle'). Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

in eastern Ukraine during the war, while mostly no statistically significant effects are found with respect to sentiments towards Ukraine. These results are also consistent with our previous analyses based on an approximated distance to the war zone reported in SES Working Paper No 472 (2016).

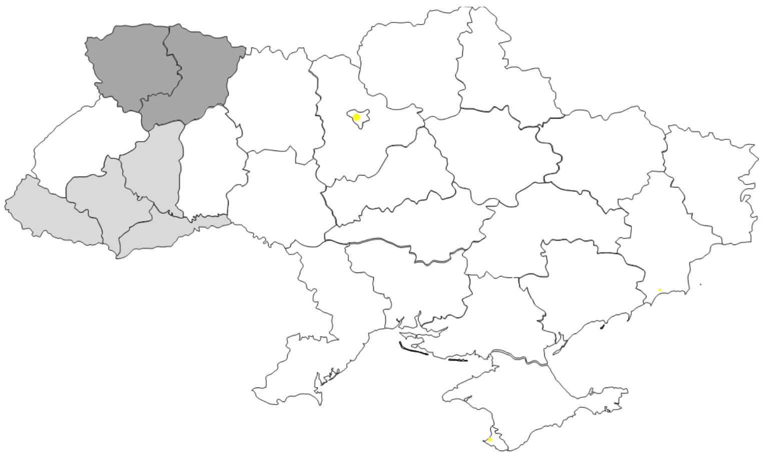
We additionally conduct a placebo test to examine the plausibility of the common trend assumption for our DiD approach. We expect that war exposure is similarly low in six oblasts of western Ukraine with the lowest inflow of IDPs,[42] which makes them least prone to sample selection bias due to the war. These oblasts make up two placebo groups as following: the oblasts of Rivne and Volyn that border with Belarus in the north — the CIS member state with a large Russian-speaking population[43] — fall into the high treatment group, while Chernivtsi,

[42] See the governmental portal, `www.kmu.gov.ua/ua/news/248052970`, retrieved 12 November 2018.

[43] According to the Population census 2009, 41.5% of the Belorussian population are native Russian speakers and

Ivano-Frankivsk, Ternopil and Zakarpattia oblasts form the low treatment group (see Figure 1.4). Our placebo sample includes 15 cities observed in both survey waves. Almost all DiD estimates are insignificant at any conventional level, with the exception of the marginally significant effects on self-association with other Ukrainians (see Appendix Table 1.A11). The placebo study therefore generally supports our DiD approach, which hinges on the assumption that it is the variability in exposure to war which crucially determines the effect of the conflict on political attitudes.

Figure 1.4: Placebo region



Note: The oblasts used in the placebo test are highlighted in dark grey (Volyn and Rivne) and light grey (Ternopil, Ivano-Frankivsk, Zakarpattia and Chernivtsi).

1.7 Conclusion

This paper examines the impact of the military conflict in Donbas on the political attitudes towards Ukraine and Russia among the population living close to the front line in eastern Ukraine. Preferences about the future political status of Ukraine, sentiments towards Ukrainian–Russian relations, as well as self-association with Ukraine and other Ukrainians are at the center of this study. The analysis is based on unique repeated cross sectional data collected from the regions controlled by the Ukrainian authorities a year before and after the outbreak of the war in spring 2014. We apply two econometric strategies to infer the causal effect of war, based on two treatment groups distinguished by their exposure to war. First, we perform before-after comparisons that assume the absence of time trends in the outcome variables in either group. Second, we use a

70% speak Russian at home. See `www.belstat.gov.by/en/perepis-naseleniya/perepis-naseleniya-2009-goda/main-demographic-and-social-characteristics-of-population-of-the-republic-of-belarus/population-classified-by-knowledge-of-the-belarusian-and-russian-languages-by-region-and-minsk-city/`, retrieved 18 June 2019.

difference-in-differences (DiD) approach for multiple treatment intensities that under specific assumptions estimates a lower bound of the causal effect. We consider estimation without controlling for covariates, as well as OLS and semiparametric inverse probability weighting by the propensity score conditional on a range of socioeconomic characteristics.

In line with previous findings in the literature, the results suggest that exposure to war affects political preferences and attitudes. Specifically, the conflict in Donbas appears to have negatively affected attitudes towards Russia among the people living in the government-held parts of Donbas close to the front line. The before-after differences in supporting a political union or one state with Russia are negative and statistically significant at the 1% level in the majority of regressions. Furthermore, for the view that the Ukrainian and Russian cultures are the same, the DiD approach yields sizable negative effects that are significant at the 5–10% level in the main specification. In contrast, we do not find strong evidence that the war has affected sentiments towards Ukraine or self-association with other Ukrainians. A number of robustness checks and the placebo test confirm our main results and do not refute the DiD method.

Our findings have several important implications for Ukraine and, more generally, for warfare without clearly defined sides. First, pro-Russian political powers are unlikely to achieve the pre-war level of electoral votes in eastern Ukraine, at least during the ongoing war. In fact, the 2014 parliamentary elections confirm a substantial drop in the pro-Russian vote to less than 50% in the region. Second, our study shows that, despite denying its involvement in a violent conflict, a country can lose political capital among the war-affected population of another country. Third, our results call for the development of theories that integrate perceptions of civilians about opposing sides in a war to better understand the impact of violence on the political attitudes and behavior of the exposed population.

1.A Additional figures and tables

Table 1.A1: Choice of the evaluation sample for eastern Ukraine (observations)

	2013	2015	Total
Nationwide sample	6000	6000	12000
Eastern Ukraine	1916	1545	3461
Excluded villages	284	264	548
Excluded cities	711	128	839
Evaluation sample	921	1153	2074

Note: The nationwide sample includes all regions of Ukraine, except for Crimea in 2015. Eastern Ukraine covers Dnipropetrovsk, Donetsk, Kharkiv, Luhansk and Zaporizhia oblasts. The evaluation sample includes cities observed in eastern Ukraine in both waves.

Table 1.A2: Mean outcome values in 2013 by city inclusion status

	Donbas				Remainder of the east			
	Included	Excluded	Difference	p-value	Included	Excluded	Difference	p-value
Independent, neutral state (binary)	0.227	0.236	-0.009	0.879	0.353	0.297	0.056	0.448
	(0.053)	(0.030)	(0.058)		(0.034)	(0.069)	(0.074)	
Union with Russia (binary)	0.473	0.569	-0.096	0.313	0.407	0.376	0.031	0.752
	(0.094)	(0.031)	(0.095)		(0.040)	(0.093)	(0.098)	
Split into separate states (binary)	0.020	0.016	0.004	0.705	0.012	0.000	0.012	0.009
	(0.007)	(0.009)	(0.011)		(0.005)	na	(0.004)	
Fully the same cultures	4.805	4.139	0.665	0.156	4.007	4.478	-0.471	0.355
(1: fully disagree, ..., 7: fully agree)	(0.459)	(0.170)	(0.468)		(0.314)	(0.419)	(0.509)	
One state with Russia	4.804	4.118	0.686	0.244	3.665	3.951	-0.287	0.504
(1: fully disagree, ..., 7: fully agree)	(0.591)	(0.177)	(0.588)		(0.155)	(0.416)	(0.429)	
I love Ukraine	3.838	4.030	-0.192	0.157	4.248	4.504	-0.256	0.055
(1: definitely no, ..., 5: definitely yes)	(0.125)	(0.066)	(0.136)		(0.039)	(0.133)	(0.133)	
"We" for Ukrainians	3.905	3.765	0.140	0.339	3.956	4.365	-0.410	0.012
(1: definitely no, ..., 5: definitely yes)	(0.128)	(0.082)	(0.146)		(0.075)	(0.150)	(0.163)	

Note: Each variable is averaged over non-missing values. Standard errors are clustered at the city level and reported in parentheses.

Table 1.A3: Mean covariate values in 2013 by city inclusion status

	Donbas				Remainder of the east			
	Included	Excluded	Difference	p-value	Included	Excluded	Difference	p-value
City size: <50,000 citizens (binary)	0.235	0.359	-0.124	0.487	0.147	1.000	-0.853	0.000
	(0.148)	(0.111)	(0.179)		(0.071)	(0.000)	(0.071)	
Female (binary)	0.559	0.576	-0.017	0.241	0.557	0.539	0.017	0.259
	(0.013)	(0.008)	(0.014)		(0.007)	(0.014)	(0.015)	
Age	46.011	46.446	-0.435	0.494	45.015	46.183	-1.168	0.181
	(0.499)	(0.426)	(0.636)		(0.314)	(0.847)	(0.871)	
Secondary specialized education (binary)	0.419	0.450	-0.031	0.573	0.376	0.348	0.028	0.422
	(0.044)	(0.035)	(0.054)		(0.020)	(0.030)	(0.035)	
University degree (binary)	0.346	0.329	0.018	0.805	0.426	0.296	0.130	0.042
	(0.063)	(0.038)	(0.071)		(0.028)	(0.060)	(0.064)	
Native Ukrainian speaker (binary)	0.130	0.058	0.072	0.117	0.268	0.270	-0.002	0.985
	(0.046)	(0.014)	(0.046)		(0.043)	(0.082)	(0.090)	
Native Russian speaker (binary)	0.395	0.541	-0.145	0.118	0.371	0.183	0.188	0.002
	(0.079)	(0.054)	(0.093)		(0.042)	(0.047)	(0.061)	
Moscow Orthodox Church (binary)	0.262	0.352	-0.089	0.219	0.169	0.182	-0.013	0.874
	(0.068)	(0.032)	(0.073)		(0.034)	(0.079)	(0.083)	
Kyiv Orthodox Church (binary)	0.119	0.109	0.010	0.866	0.192	0.064	0.128	0.006
	(0.053)	(0.032)	(0.060)		(0.030)	(0.036)	(0.046)	
Orthodox Church (binary)	0.162	0.234	-0.071	0.316	0.249	0.300	-0.051	0.676
	(0.054)	(0.050)	(0.071)		(0.043)	(0.120)	(0.123)	
Working (binary)	0.531	0.572	-0.041	0.449	0.575	0.574	0.002	0.973
	(0.027)	(0.048)	(0.055)		(0.020)	(0.044)	(0.046)	
Retired (binary)	0.279	0.282	-0.003	0.941	0.243	0.278	-0.036	0.196
	(0.018)	(0.030)	(0.034)		(0.014)	(0.025)	(0.028)	
Single (binary)	0.162	0.133	0.029	0.320	0.169	0.130	0.039	0.175
	(0.026)	(0.016)	(0.030)		(0.016)	(0.025)	(0.029)	
Married or in partnership (binary)	0.609	0.720	-0.111	0.028	0.636	0.661	-0.025	0.597
	(0.048)	(0.021)	(0.050)		(0.023)	(0.044)	(0.048)	
Household size	2.556	2.775	-0.219	0.030	2.800	2.673	0.127	0.141
	(0.094)	(0.047)	(0.101)		(0.072)	(0.050)	(0.086)	
Material conditions	4.648	4.228	0.420	0.060	4.159	4.496	-0.337	0.031
(1: very good, ..., 7: terrible)	(0.192)	(0.128)	(0.223)		(0.078)	(0.141)	(0.156)	
Years in the current municipality	38.685	35.770	2.915	0.185	36.416	33.513	2.903	0.180
	(2.171)	(0.760)	(2.196)		(0.591)	(2.169)	(2.165)	

Note: Each variable is averaged over non-missing values. Standard errors are clustered at the city level and reported in parentheses.

Figure 1.A1: Distribution of attitudes towards Ukraine and Russia in Donbas in 2013/2015

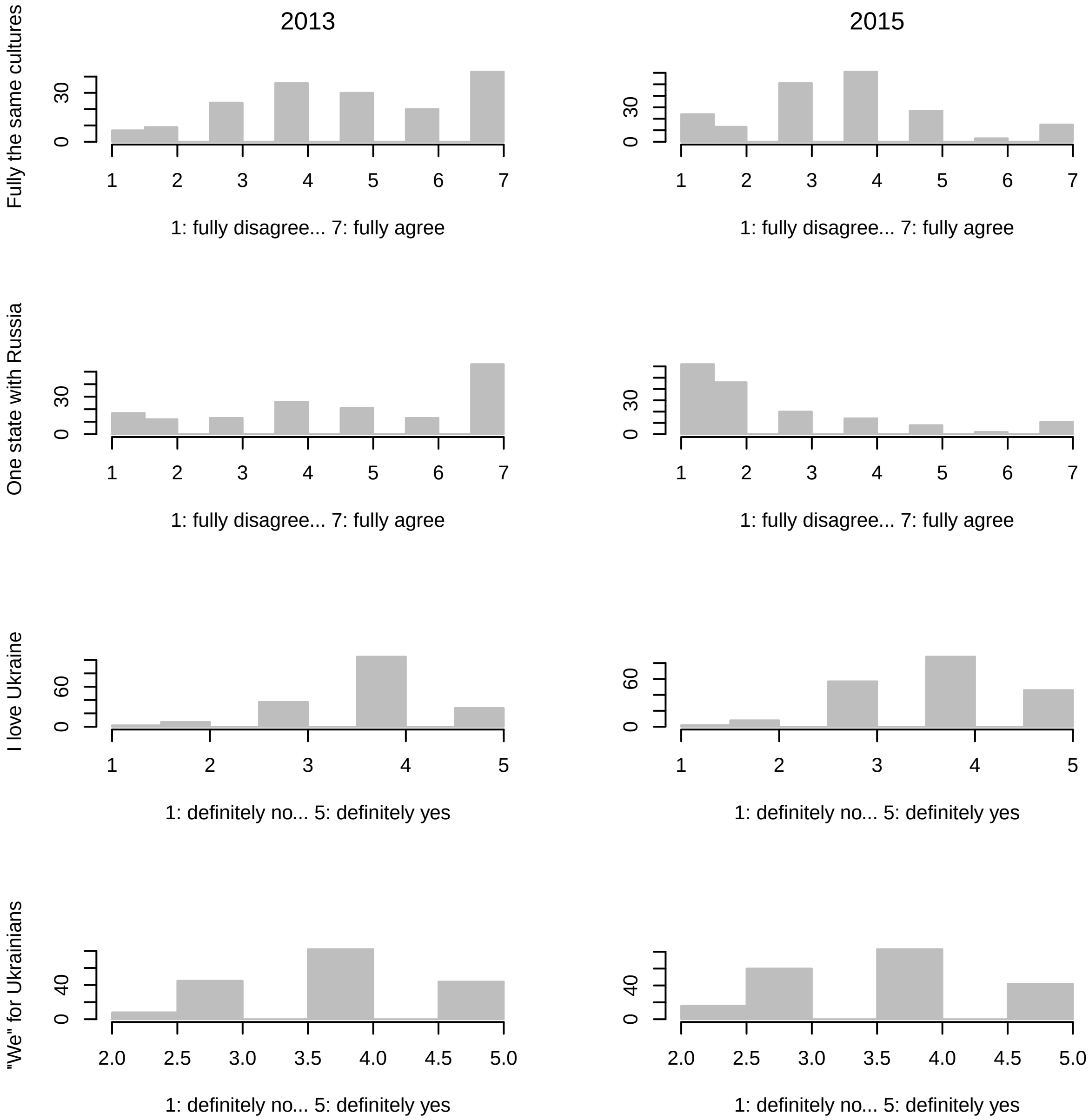


Table 1.A4: Correlations between political attitudes and views about Russia's role in the war

	Donbas		Remainder of the east	
	Russia involved	War with Russia	Russia involved	War with Russia
Independent, neutral state	-0.124	-0.088	-0.151	-0.117
	[0.127]	[0.313]	[0.000]	[0.002]
Union with Russia	-0.189	-0.252	-0.269	-0.243
	[0.019]	[0.003]	[0.000]	[0.000]
Split into separate states	-0.214	-0.134	-0.088	-0.057
	[0.008]	[0.124]	[0.013]	[0.125]
Fully the same cultures	-0.409	-0.283	-0.224	-0.265
	[0.000]	[0.000]	[0.000]	[0.000]
One state with Russia	-0.369	-0.255	-0.310	-0.378
	[0.000]	[0.003]	[0.000]	[0.000]
I love Ukraine	0.184	0.194	0.251	0.193
	[0.009]	[0.012]	[0.000]	[0.000]
"We" for Ukrainians	0.355	0.162	0.271	0.154
	[0.000]	[0.036]	[0.000]	[0.000]

Note: Person's correlation coefficients with p-values in brackets.

Table 1.A5: Estimates for native Russian speakers only

	No controls			OLS			Propensity score weighting		
	Donbas	East	DiD	Donbas	East	DiD	Donbas	East	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.040	0.110	-0.070	0.210	0.062	0.148	0.205	0.293	-0.088
st. error	0.111	0.101	0.157	0.258	0.088	0.271	0.139	0.188	0.243
p-value	0.722	0.278	0.654	0.417	0.485	0.585	0.140	0.120	0.717
observations	113	413	526	104	382	486	104	382	486
Union with Russia	-0.349	-0.285	-0.064	-0.428	-0.255	-0.173	-0.381	-0.267	-0.114
st. error	0.201	0.070	0.211	0.781	0.046	0.781	0.283	0.226	0.335
p-value	0.083	0.000	0.763	0.584	0.000	0.825	0.179	0.238	0.734
observations	113	413	526	104	382	486	104	382	486
Split into separate states	0.311	0.054	0.256	0.244	0.049	0.195	0.306	0.035	0.270
st. error	0.120	0.024	0.122	0.321	0.028	0.322	0.110	0.040	0.124
p-value	0.010	0.023	0.036	0.447	0.085	0.544	0.006	0.373	0.030
observations	113	413	526	104	382	486	104	382	486
Fully the same cultures	-1.182	-0.071	-1.111	-1.240	0.021	-1.262	-1.033	0.186	-1.219
st. error	0.569	0.363	0.689	3.154	0.357	3.182	0.734	0.742	1.188
p-value	0.038	0.845	0.107	0.694	0.952	0.692	0.159	0.802	0.305
observations	142	494	636	131	446	577	131	446	577
One state with Russia	-2.262	-0.890	-1.372	-2.193	-0.563	-1.629	-1.778	-0.431	-1.347
st. error	1.163	0.398	1.225	2.892	0.326	2.924	1.610	1.111	2.066
p-value	0.052	0.025	0.263	0.448	0.084	0.577	0.269	0.698	0.515
observations	108	457	565	97	415	512	97	415	512
I love Ukraine	-0.397	0.044	-0.441	-0.348	0.079	-0.427	-0.451	0.085	-0.536
st. error	0.402	0.092	0.409	0.574	0.097	0.583	0.481	0.275	0.528
p-value	0.323	0.635	0.281	0.544	0.416	0.464	0.348	0.758	0.310
observations	147	522	669	136	470	606	136	470	606
"We" for Ukrainians	-0.457	0.056	-0.513	-0.381	0.107	-0.488	-0.645	0.170	-0.814
st. error	0.374	0.123	0.396	0.995	0.135	1.003	0.444	0.364	0.656
p-value	0.222	0.650	0.196	0.701	0.430	0.626	0.147	0.641	0.215
observations	147	522	669	136	470	606	136	470	606

Note: Sample includes the cities observed in both waves. Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), the rest of the east forms the low treatment group ('East'). Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

Table 1.A6: Estimates for native Ukrainian and Russian speakers (bilingual)

	No controls			OLS			Propensity score weighting		
	Donbas	East	DiD	Donbas	East	DiD	Donbas	East	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.311	0.282	0.029	0.325	0.234	0.091	0.327	0.257	0.070
st. error	0.202	0.059	0.210	0.199	0.075	0.214	0.201	0.131	0.248
p-value	0.124	0.000	0.892	0.103	0.002	0.671	0.104	0.050	0.777
observations	131	538	669	119	512	631	119	512	631
Union with Russia	-0.454	-0.426	-0.028	-0.465	-0.438	-0.027	-0.579	-0.412	-0.167
st. error	0.081	0.048	0.092	0.152	0.047	0.159	0.207	0.138	0.249
p-value	0.000	0.000	0.762	0.002	0.000	0.865	0.005	0.003	0.504
observations	131	538	669	119	512	631	119	512	631
Split into separate states	0.049	0.016	0.034	0.048	0.011	0.036	0.053	0.014	0.039
st. error	0.028	0.013	0.031	0.056	0.010	0.057	0.028	0.011	0.027
p-value	0.081	0.217	0.284	0.394	0.277	0.525	0.064	0.219	0.143
observations	131	538	669	119	512	631	119	512	631
Fully the same cultures	-1.662	0.316	-1.978	-1.929	0.264	-2.193	-2.305	0.663	-2.968
st. error	0.397	0.453	0.607	0.394	0.535	0.658	0.897	0.773	1.254
p-value	0.000	0.484	0.001	0.000	0.621	0.001	0.010	0.391	0.018
observations	158	637	795	144	601	745	144	601	745
One state with Russia	-2.719	-1.304	-1.415	-3.243	-1.418	-1.825	-3.687	-1.046	-2.641
st. error	0.657	0.300	0.711	0.831	0.407	0.938	1.084	0.668	1.338
p-value	0.000	0.000	0.047	0.000	0.000	0.052	0.001	0.118	0.048
observations	151	585	736	136	553	689	136	553	689
I love Ukraine	0.155	0.131	0.024	0.310	0.238	0.071	0.691	0.155	0.536
st. error	0.226	0.065	0.233	0.290	0.071	0.298	0.508	0.158	0.511
p-value	0.493	0.044	0.917	0.285	0.001	0.811	0.174	0.328	0.294
observations	168	661	829	152	623	775	152	623	775
"We" for Ukrainians	0.199	0.139	0.060	0.323	0.160	0.164	0.365	0.104	0.262
st. error	0.300	0.104	0.315	0.335	0.127	0.358	0.455	0.230	0.479
p-value	0.507	0.182	0.849	0.335	0.210	0.647	0.422	0.652	0.585
observations	168	661	829	152	623	775	152	623	775

Note: Sample includes the cities observed in both waves. Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), the rest of the east forms the low treatment group ('East'). Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

Table 1.A7: Estimates for respondents aged up to 45

	No controls			OLS			Propensity score weighting		
	Donbas	East	DiD	Donbas	East	DiD	Donbas	East	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.205	0.014	0.190	0.289	-0.019	0.309	0.269	0.225	0.044
st. error	0.140	0.068	0.156	0.159	0.075	0.176	0.232	0.155	0.287
p-value	0.145	0.831	0.221	0.069	0.796	0.080	0.245	0.147	0.878
observations	152	706	858	141	660	801	141	660	801
Union with Russia	-0.195	-0.238	0.044	-0.332	-0.235	-0.097	-0.444	-0.248	-0.196
st. error	0.147	0.044	0.153	0.181	0.046	0.184	0.212	0.153	0.270
p-value	0.184	0.000	0.776	0.066	0.000	0.597	0.036	0.105	0.467
observations	152	706	858	141	660	801	141	660	801
Split into separate states	0.128	0.042	0.086	0.103	0.023	0.079	0.110	0.025	0.085
st. error	0.046	0.022	0.051	0.098	0.017	0.100	0.064	0.018	0.065
p-value	0.006	0.058	0.092	0.295	0.159	0.426	0.084	0.162	0.187
observations	152	706	858	141	660	801	141	660	801
Fully the same cultures	-1.149	0.066	-1.215	-1.351	0.037	-1.388	-1.520	0.251	-1.771
st. error	0.572	0.356	0.680	0.957	0.412	1.060	0.775	0.647	1.145
p-value	0.045	0.852	0.074	0.158	0.928	0.191	0.050	0.699	0.122
observations	180	808	988	168	739	907	168	739	907
One state with Russia	-2.076	-1.120	-0.956	-2.499	-1.047	-1.452	-2.829	-0.663	-2.166
st. error	0.877	0.257	0.900	0.830	0.313	0.886	0.878	0.866	1.387
p-value	0.018	0.000	0.288	0.003	0.001	0.101	0.001	0.444	0.119
observations	158	767	925	145	703	848	145	703	848
I love Ukraine	0.036	0.257	-0.221	0.150	0.293	-0.143	0.287	0.174	0.114
st. error	0.242	0.078	0.253	0.279	0.075	0.292	0.326	0.220	0.381
p-value	0.880	0.001	0.383	0.591	0.000	0.624	0.379	0.431	0.766
observations	191	848	1039	176	771	947	176	771	947
“We” for Ukrainians	-0.093	0.233	-0.326	0.098	0.238	-0.140	0.148	0.211	-0.063
st. error	0.322	0.108	0.341	0.319	0.140	0.343	0.378	0.326	0.577
p-value	0.772	0.032	0.338	0.758	0.089	0.683	0.695	0.517	0.914
observations	191	848	1039	176	771	947	176	771	947

Note: Sample includes the cities observed in both waves. Donetsk and Luhansk oblasts form the high treatment group (‘Donbas’), the rest of the east forms the low treatment group (‘East’). Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

Table 1.A8: Estimates for respondents aged above 45

	No controls			OLS			Propensity score weighting		
	Donbas	East	DiD	Donbas	East	DiD	Donbas	East	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.197	0.178	0.019	0.188	0.119	0.069	0.170	0.203	-0.033
st. error	0.197	0.097	0.221	0.173	0.091	0.197	0.200	0.117	0.199
p-value	0.317	0.067	0.930	0.278	0.190	0.729	0.394	0.082	0.869
observations	152	693	845	139	657	796	139	657	796
Union with Russia	-0.513	-0.375	-0.138	-0.422	-0.390	-0.031	-0.441	-0.384	-0.056
st. error	0.081	0.058	0.100	0.116	0.046	0.125	0.146	0.107	0.178
p-value	0.000	0.000	0.168	0.000	0.000	0.803	0.002	0.000	0.751
observations	152	693	845	139	657	796	139	657	796
Split into separate states	0.105	0.002	0.104	0.079	0.010	0.069	0.065	0.017	0.048
st. error	0.076	0.010	0.077	0.074	0.010	0.075	0.063	0.017	0.066
p-value	0.163	0.870	0.176	0.288	0.295	0.358	0.301	0.316	0.466
observations	152	693	845	139	657	796	139	657	796
Fully the same cultures	-1.202	0.156	-1.358	-1.282	0.020	-1.303	-1.562	0.314	-1.876
st. error	0.510	0.397	0.648	0.563	0.385	0.676	0.702	0.556	0.981
p-value	0.018	0.693	0.036	0.023	0.958	0.054	0.026	0.572	0.056
observations	183	811	994	168	767	935	168	767	935
One state with Russia	-2.637	-1.312	-1.325	-2.716	-1.307	-1.409	-2.990	-0.932	-2.058
st. error	0.743	0.296	0.786	0.964	0.319	1.020	1.057	0.691	1.309
p-value	0.000	0.000	0.092	0.005	0.000	0.167	0.005	0.177	0.116
observations	163	753	916	148	714	862	148	714	862
I love Ukraine	-0.045	0.090	-0.136	0.034	0.117	-0.083	-0.030	0.089	-0.119
st. error	0.247	0.066	0.257	0.305	0.077	0.314	0.370	0.155	0.388
p-value	0.855	0.172	0.598	0.911	0.130	0.792	0.935	0.567	0.759
observations	189	846	1035	174	800	974	174	800	974
“We” for Ukrainians	-0.214	0.208	-0.422	-0.013	0.181	-0.194	-0.055	0.163	-0.218
st. error	0.286	0.093	0.298	0.236	0.088	0.250	0.288	0.176	0.313
p-value	0.454	0.025	0.156	0.956	0.039	0.437	0.848	0.354	0.487
observations	189	846	1035	174	800	974	174	800	974

Note: Sample includes the cities observed in both waves. Donetsk and Luhansk oblasts form the high treatment group (‘Donbas’), the rest of the east forms the low treatment group (‘East’). Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

Table 1.A9: Effects on item non-response in outcome variables for Donbas and the rest of the east

	No controls			OLS			Propensity score weighting		
	Donbas	East	DiD	Donbas	East	DiD	Donbas	East	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.072	-0.038	0.110	0.087	-0.052	0.139	0.093	-0.078	0.171
st. error	0.105	0.038	0.113	0.128	0.041	0.137	0.147	0.058	0.161
p-value	0.492	0.313	0.330	0.497	0.206	0.309	0.526	0.177	0.289
observations	380	1694	2074	350	1571	1921	350	1571	1921
Fully the same cultures	-0.021	-0.017	-0.004	-0.024	-0.020	-0.004	-0.010	-0.024	0.014
st. error	0.015	0.011	0.019	0.015	0.014	0.021	0.018	0.029	0.034
p-value	0.150	0.125	0.834	0.123	0.147	0.843	0.580	0.407	0.668
observations	380	1694	2074	350	1571	1921	350	1571	1921
One state with Russia	0.072	0.013	0.059	0.048	-0.006	0.053	0.087	-0.001	0.088
st. error	0.038	0.021	0.043	0.048	0.020	0.051	0.070	0.056	0.084
p-value	0.060	0.559	0.173	0.318	0.780	0.298	0.210	0.989	0.292
observations	380	1694	2074	350	1571	1921	350	1571	1921

Note: Sample includes the cities observed in both waves. Donetsk and Luhansk oblasts form the high treatment group (‘Donbas’), the rest of the east forms the low treatment group (‘East’). Dummies for missing observations in outcomes are treated as outcome variables. Missing observations for mutually exclusive scenarios “independent, neutral state”, “union with Russia” and “split into separate states” coincide. There are no missing observations for variables “I love Ukraine” and ““we” for Ukrainians”. Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

Table 1.A10: Estimates using all cities in eastern Ukraine

	No controls			OLS			Propensity score weighting		
	Donbas	East	DiD	Donbas	East	DiD	Donbas	East	DiD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Independent, neutral state	0.222	0.105	0.117	0.210	0.071	0.139	0.232	0.166	0.067
st. error	0.097	0.075	0.124	0.104	0.070	0.125	0.111	0.082	0.132
p-value	0.021	0.158	0.344	0.044	0.312	0.266	0.036	0.043	0.614
observations	855	1557	2412	783	1468	2251	783	1468	2251
Union with Russia	-0.429	-0.302	-0.127	-0.423	-0.306	-0.117	-0.451	-0.312	-0.139
st. error	0.052	0.036	0.064	0.045	0.036	0.058	0.051	0.068	0.086
p-value	0.000	0.000	0.048	0.000	0.000	0.043	0.000	0.000	0.106
observations	855	1557	2412	783	1468	2251	783	1468	2251
Split into separate states	0.130	0.023	0.108	0.130	0.014	0.115	0.119	0.020	0.099
st. error	0.038	0.012	0.040	0.041	0.009	0.042	0.040	0.012	0.042
p-value	0.001	0.068	0.007	0.002	0.094	0.006	0.003	0.105	0.020
observations	855	1557	2412	783	1468	2251	783	1468	2251
Fully the same cultures	-0.633	0.040	-0.673	-0.617	0.009	-0.625	-0.735	0.155	-0.890
st. error	0.313	0.328	0.451	0.357	0.341	0.503	0.419	0.478	0.704
p-value	0.043	0.903	0.136	0.084	0.979	0.214	0.079	0.746	0.206
observations	1009	1793	2802	923	1672	2595	923	1672	2595
One state with Russia	-1.564	-1.263	-0.301	-1.742	-1.217	-0.525	-1.909	-1.045	-0.864
st. error	0.405	0.243	0.487	0.443	0.262	0.513	0.503	0.456	0.728
p-value	0.000	0.000	0.536	0.000	0.000	0.307	0.000	0.022	0.235
observations	940	1683	2623	857	1572	2429	857	1572	2429
I love Ukraine	-0.095	0.143	-0.238	-0.107	0.176	-0.283	-0.109	0.088	-0.197
st. error	0.135	0.061	0.150	0.139	0.064	0.155	0.154	0.098	0.185
p-value	0.483	0.018	0.112	0.440	0.006	0.067	0.478	0.369	0.286
observations	1041	1872	2913	950	1741	2691	950	1741	2691
"We" for Ukrainians	-0.000	0.157	-0.157	0.029	0.141	-0.112	0.009	0.045	-0.036
st. error	0.224	0.086	0.243	0.199	0.095	0.223	0.212	0.133	0.262
p-value	0.998	0.067	0.517	0.886	0.138	0.615	0.967	0.734	0.890
observations	1041	1872	2913	950	1741	2691	950	1741	2691

Note: Donetsk and Luhansk oblasts form the high treatment group ('Donbas'), the rest of the east forms the low treatment group ('East'). Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

Table 1.A11: Placebo DiD estimates

	No controls (1)	OLS (2)	Propensity score weighting (3)
Independent, neutral state	-0.110	-0.137	-0.164
st. error	0.175	0.212	0.274
p-value	0.532	0.518	0.548
observations	513	501	501
Union with Russia	0.009	0.038	0.026
st. error	0.068	0.062	0.072
p-value	0.899	0.540	0.721
observations	513	501	501
Split into separate states	0.002	0.001	0.009
st. error	0.021	0.017	0.070
p-value	0.938	0.967	0.898
observations	513	501	501
Fully the same cultures	-0.472	-0.438	-0.034
st. error	0.479	0.363	0.696
p-value	0.324	0.228	0.962
observations	565	547	547
One state with Russia	-0.326	-0.166	-0.234
st. error	0.512	0.405	0.664
p-value	0.525	0.682	0.724
observations	566	548	548
I love Ukraine	-0.002	0.054	0.131
st. error	0.183	0.202	0.292
p-value	0.991	0.789	0.654
observations	592	573	573
“We” for Ukrainians	-0.783	-0.737	-0.938
st. error	0.422	0.444	0.545
p-value	0.063	0.097	0.085
observations	592	573	573

Note: Sample includes the cities observed in both waves. Rivne and Volyn oblasts form the high treatment group; Chernivtsi, Ivano-Frankivsk, Ternopil and Zakarpattia oblasts form the low treatment group. Propensity scores are estimated by logistic regression. Standard errors are clustered at the city level using 999 bootstrap replications.

2 Education, fertility and labor force attachment: A mediation analysis of older women

Overview

In the era of population ageing, investments into human capital may contribute to longer working lives. Though positive returns to schooling in terms of labor market attachment may seem obvious, little is known about the underlying mechanisms. Given an increasing role of women in labor supply, this paper investigates the long-run effects of education on female labor force attachment mediated by realized fertility. Using mediation analysis supported by the rich data from the Survey of Health, Ageing and Retirement in Europe, we find that schooling is positively associated with employment status, working hours and labor survival age of older women. Education-induced fertility explains a modest share of the total effects (about 3–8%, depending on the outcome measure), with important policy implications.

2.1 Introduction

Populations are ageing rapidly[1] creating serious socioeconomic challenges. As one response, governments are advised to provide older people with better work incentives to increase labor supply (OECD, 2015). In the era of low (declining) fertility rates and population ageing, more working women are seen as an important source of sustainable economic growth (OECD, 2008). The role of women in labor supply has steadily increased over the last decades in many countries,[2] and recent projections suggest that labor participation rates of older female workers will continue to increase (European Commission, 2018). It is therefore essential to gain an in-depth understanding of factors and channels that enhance longer working lives of women.

One factor that can stimulate female labor supply is education because it improves individual earnings capacity. However, more schooling may also lead to lower fertility due to higher oppor-

[1]The number of countries with at least 20% of the population aged 65 and above is projected to grow from 13 in 2018 to 82 in 2050 (Population Reference Bureau, 2018). At present, the developed economies experience the highest share of older persons. In the context of longer life expectancies and low fertility rates (1.58–1.81), the ratio of elderly people (65 and over) to people of working age (15–64) in the EU is predicted to increase from 29.6% in 2016 to 51.2% in 2070 (European Commission, 2018).

[2]Between 1985–2018, the labor force participation rate of 15–64 years old women increased from 55.3% to 64.6% and that of 55–64 years old women from 34.1% to 55.2% in OECD countries. Over the same period, the labor force participation rate of 15–64 years old men decreased from 82.1% to 80.4% and that of 55–64 years old men increased from 66.3% to 73%. Despite the gender gap narrowing, women remain underrepresented in the labor force. See OECD Statistics, `stats.oecd.org`, retrieved 14 October 2019.

tunity costs of motherhood in terms of foregone earnings, which in turn can increase female labor force participation. Since education may directly but also indirectly — through fertility — affect the labor supply of women at an older age, it is important to analyze the composition of the total effect of education on female labor attachment. If the total effect is mainly driven by the indirect effect due to reduced fertility, more education may create risks for social security systems in the long run. While Lusardi and Mitchell (2016) find correlational evidence that older women with more education and fewer children are more likely to extend their working lives, the underlying pathways remain unexplained.

We examine the long-term relationship between education and female labor market attachment, using mediation analysis (see Judd and Kenny, 1981; Baron and Kenny, 1986; Pearl, 2001) to shed light on the underlying mechanisms. Our primary focus lies with testing the hypothesis whether labor market benefits of education (partially) materialize through fertility. To this end, we decompose the total effect of schooling into the natural direct and indirect effects. In doing so, we assume sequential ignorability of education and fertility (see Imai et al., 2010), implying that all their confounders are observed. Our identification strategy is supported with the rich dataset from the Survey of Health, Ageing and Retirement in Europe (SHARE), a representative panel targeting older individuals in several mostly European countries. We also address potential endogeneity issues with multiple robustness checks.

Our results show a significant positive relationship between education and labor market attachment both at the extensive and intensive margins. In particular, more educated women on average display higher employment rates, are less likely to be homemakers, work more hours a week and exit the labor market at an older age. The total effects are partially explained by a fertility decline induced by extra schooling. The indirect effects account for about 3–8% of the total schooling effects, depending on the outcome. We further find that the mediation effects are mainly pronounced at the intensive fertility margin.

This study contributes to a large body of the economic literature on long-term returns to schooling. While assessing the total causal effects of education on different adult outcomes, such as, for instance, crime (e.g. Lochner and Moretti, 2004), earnings (e.g. Brunello et al., 2009), health (e.g. Brunello et al., 2013), cognitive performance (e.g. Banks and Mazzonna, 2012; Schneeweis et al., 2014), mortality (e.g. Gathmann et al., 2015) and fertility (e.g. Fort et al., 2016), the authors often discuss the impact of education on the potential mediating factors but leave the causal mediation effects unexamined. Hence, our paper is among few that provide empirical evidence on the underlying channels that may drive returns to schooling.

There are three other exceptions we know about. Using the mediation framework with instrumental variables, Powdthavee et al. (2013) study whether education affects subjective well-being through income. In another paper, Powdthavee et al. (2015) consider multiple pathways, including the number of children, through which education may shape life satisfaction. Brunello et al. (2016), who also rely on the richness of the SHARE data, investigate the effect of education on health mediated by health behaviors. Though methodologically similar to these studies, our paper investigates the direct and indirect effects of education on different economically important outcomes. To our best knowledge, this is the first attempt to shed light on the interconnections between schooling, fertility and labor supply using mediation analysis.

Our findings suggest that policies must be designed to encourage human capital investments and to facilitate parenthood for females who pursue advanced degrees or professional career in order to foster female labor market participation at an older age and thus to strengthen social security systems.

The remainder of this paper is organized as follows. Section 2.2 reviews the relevant literature and presents the hypotheses. Section 2.3 describes the data. Section 2.4 discusses our empirical strategy and identification issues. Section 2.5 reports the results and robustness checks. Section 2.6 concludes.

2.2 Literature

Economic theory closely links education, fertility and labor supply choices. On the one hand, education increases a trade-off between labor market opportunities and childrearing (Becker, 1965). Given that childcare involves potentially large foregone earnings, the substitution effect predicts fertility decline. On the other hand, higher income increases fertility if parents perceive children as normal goods (ibid.). In addition, a quantity-quality trade-off in children implies that parents with higher income prefer investing into the quality rather than the quantity of their children (Becker and Lewis, 1973). Hence, the effect of education on fertility and subsequent labor supply depends on the magnitudes of the substitution and income effects, which are not a priori obvious without empirical analysis. In the following, we review the relevant empirical literature and define our contribution.

2.2.1 Education and labor market outcomes

Unobserved confounding is a fundamental problem inherent in most observational studies that attempt to estimate the causal effect of education on labor market outcomes. Different methods have been adopted in the literature to account for the potential endogeneity of schooling (see Blundell et al., 2005, for a review). Since Angrist and Krueger (1991), researchers have actively exploited changes in compulsory schooling laws to overcome the need to observe individual ability and family background, which jointly affect (confound) schooling and labor market decisions. Instrumenting education with an increase in the minimum school-leaving age,[3] empirical studies

[3]This identification strategy at least implicitly assumes that minimum school-leaving age reforms affect only the quantity but not the quality of education. Otherwise, returns to schooling may be inconsistently estimated if school quality is an omitted variable. Brunello et al. (2013) have tested whether years of compulsory schooling are a valid instrument for years of schooling. Their results cannot reject the hypothesis that the exclusion restriction holds, i.e.

have reported significant positive returns to schooling in terms of earnings or wages in many countries (e.g. Harmon and Walker, 1995; Oreopoulos, 2006, 2007; Aakvik et al., 2010; Grenet, 2013). On the contrary, Pischke and von Wachter (2008) and Grenet (2013) have found no evidence of significant returns for West Germany or France, respectively.

Some studies reveal heterogeneity in returns by individual characteristics. For instance, Devereux and Hart (2010) have estimated positive returns to schooling for men but insignificant returns for women in Britain. According to Meghir and Palme (2005), additional schooling induced by the Swedish educational reform significantly increased earnings for individuals with low-skilled fathers but had the opposite effect for individuals with high-skilled fathers. Examining the impact of education on the conditional distribution of wages in Europe, Brunello et al. (2009) have found that more schooling reduces wage inequality.[4] This evidence highlights the point that the instrumental variable (IV) estimator identifies only the causal effect for compliers.[5] The results in Ichino and Winter-Ebmer (1999) show that the IV estimates of returns to schooling depend on a choice of IV. Reviewing studies on returns to schooling, Card (2001) concludes that the IV estimates are typically larger but rather imprecise and not significantly different from the corresponding OLS estimates.

A striking feature of the existing literature is its primary focus on men, with a few exceptions (e.g. Meghir and Palme, 2005; Leigh and Ryan, 2008; Devereux and Hart, 2010; Brunello et al., 2009).[6] As to the causal effects of education on labor supply, empirical evidence is also fairly scarce. Oreopoulos (2007) finds that schooling significantly reduces the likelihood of unemployment in Canada and the UK. Using the SHARE data, Brunello et al. (2013) report that more schooling leads to higher employment, and the absolute effect is larger for women than for men. Riddell and Song (2011) have found that education significantly increases re-employment rates of the unemployed in the US.

2.2.2 Education and fertility

To explore the causal effect of education on fertility, educational reforms are again most commonly used as a source of exogenous variation in schooling. The related literature often focuses on single countries. By and large, most studies conclude that education significantly reduces teenage births and postpones motherhood (e.g. Black et al., 2008; Monstad et al., 2008; Silles, 2011; Grönqvist and Hall, 2013; Güneş, 2015). However, empirical evidence on the fertility returns to education

compulsory schooling reforms do not affect school quality.

[4]Their results indicate that compulsory schooling reforms affect mainly individuals at the lower end of the educational distribution.

[5]A subgroup of the population who is induced to change their behavior in line with the instrument (Angrist et al., 1996).

[6]Further examples include studies that use sibling sex composition (Butcher and Case, 1994), tuition and distance to the nearest college (Kane and Rouse, 1995), spouse's and parents' education (Trostel et al., 2002) as instruments for education.

appears rather context-specific, which may be driven by differences in childcare availability.

At the extensive fertility margin, some evidence suggests that more schooling increases the incidence of childlessness in the US and Germany (León, 2004; Cygan-Rehm and Maeder, 2013, respectively) but raises the likelihood of motherhood in Canada (DeCicca and Krashinsky, 2016). On the other hand, Monstad et al. (2008), McCrary and Royer (2011) and Braakmann (2011) have found insignificant effects of education on childlessness in Norway, the UK and the US states (California and Texas), respectively.

At the intensive margin, different studies show that more schooling significantly increases (León, 2004; Cygan-Rehm and Maeder, 2013; DeCicca and Krashinsky, 2016), decreases (Braakmann, 2011) or does not affect (Monstad et al., 2008; McCrary and Royer, 2011; Geruso and Royer, 2014) completed fertility. Fort et al. (2016) have found that an additional year of education leads women to have fewer children in England but positively affects fertility in Continental Europe. Using a removal of mobility restrictions on Israeli Arabs as a natural experiment for access to schooling, Lavy and Zablotsky (2015) have discovered a significant negative effect of education on Arab women's fertility rate. A within-twin study by Amin and Behrman (2014) suggests that more schooling leads to fewer children but does not impact the incidence of childlessness in the US state of Minnesota.

2.2.3 Fertility and female labor supply

The estimation of fertility effects on female labor market outcomes has also received considerable attention in the literature. However, the potential endogeneity of childbearing makes causal analysis challenging. This issue is often tackled with an IV approach. Multiple births (Rosenzweig and Wolpin, 1980) and the gender composition of the first two children (Angrist and Evans, 1998) are commonly assumed to induce exogenous variation in fertility of instrument-specific compliers. Using these instruments, multiple studies have shown that fertility tends to significantly reduce mothers' labor supply (e.g. Bronars and Grogger, 1994; Angrist and Evans, 1998; Jacobsen et al., 1999; Chun and Oh, 2002; Cruces and Galiani, 2007). Contrary, Iacovou (2001) has found insignificant marginal effects of a third child on mothers' labor supply and hours worked in the UK. Analyzing fertility effects over time and with economic development, Aaronson et al. (2017) point to rather insignificant effects on mothers' work activity at low income levels but significant negative effects at higher income levels.[7] Similarly, Agüero and Marks (2008, 2011) report insignificant ef-

[7]Though these findings appear robust to instrument-specific groups of compliers — evidence of external validity according to Angrist et al. (2010) — the internal validity of the twin-birth and same-gender IVs is prone to omitted variable biases. For instance, twin births may occur with higher chances among older, healthier and better-off mothers (Rosenzweig and Wolpin, 2000; Hoekstra et al., 2008; Angrist et al., 2010; Bhalotra and Clarke, 2016); the same-gender instrument may violate the exclusion restriction due to economies of scale associated with raising children of the same gender (Rosenzweig and Wolpin, 2000). The latter concern appears more relevant in developing than developed countries (Bütikofer, 2011). Furthermore, IV strategies based on multiple births or parental preferences

fects of children on mothers' labor force participation in Latin America using infertility shocks. Alternative strategies that exploit fertility treatments (Cristia, 2008; Lundborg et al., 2017) or abortion laws (Angrist and Evans, 1996; Bloom et al., 2009) have established a significant negative effect of children on their mothers' labor supply.

2.2.4 Our contribution

Our paper contributes to the literature in several directions. As mentioned before, authors often exploit the IV approach for causal analysis when detailed data on early-life conditions are not available. Though compelling, the IV assumptions of monotonicity and exclusion restrictions are not testable. Furthermore, the IV estimator can only identify the local average treatment effect for a specific group of compliers that may differ from the population average effect. Our paper stands out in that we have rich data on latent ability, family background and early life conditions to estimate average effects in the population.

Unlike most studies that focus on the total effects of schooling or fertility on labor market outcomes, we examine possible mechanisms governing these relationships by decomposing the total effect of schooling into a direct effect and an indirect effect operating though fertility. Hence, fertility is an intermediate outcome in our empirical analysis. Furthermore, our study provides extra evidence on returns to schooling in terms of female labor market attachment at an older age. We also assess long-term relationship between fertility and female labor outcomes. Finally, our study covers several countries, which enhances the external validity of findings.

Taking into account the economic theory and empirical evidence, we expect to detect a positive relationship between schooling and female labor market attachment in terms of total effects. The direction and magnitude of indirect effects depends on a link between education and realized fertility on the one hand and a connection between fertility and labor supply on the other hand (see Section 2.4). While we anticipate the latter to be rather negative, the direction of the former is less clear due to a likely trade-off between the income and substitution effects of education.

2.3 Data

To assess the role of fertility in mediating the relationship between female education and labor market participation, we use data from the first three waves of the Survey of Health, Ageing and Retirement in Europe (SHARE),[8] a representative panel of individuals aged 50 and above[9] in several European countries — Austria, Belgium, the Czech Republic, Denmark, France, Germany,

for children of particular gender focus on the intensive fertility margin, i.e. do not investigate how motherhood affects labor force participation.

[8]Data release 5.0.0 is used.

[9]The dataset also includes a few individuals aged below 50 because other present household members (e.g. spouses or partners) got interviewed as well. Only 3.5% of females in our evaluation sample are younger than 50, namely 45–49 years old.

Greece, Italy, the Netherlands, Poland, Spain, Sweden, and Switzerland.[10] Panel attrition being an issue in the SHARE, our evaluation sample pools data of all women who participated in either the first wave (2004/2005) or the second wave (2006/2007), dropping those who did not appear in the retrospective third wave (SHARELIFE, 2008/2009). From respondents' first interview, we measure educational attainment, realized fertility, labor market attachment and (other) important sociodemographic characteristics, which are further complemented with the retrospective information on early-life conditions.

We restrict our sample to individuals born in the country of residence, where interview took place, or migrated up to the age of 5 to insure that they are potentially well-integrated into the local labor market at least in terms of exposure to the compulsory education. Only women aged 45 or above enter our sample because they must have already completed their fertility. Then, we exclude females who were younger than 15 or older than 45 at the first childbirth. Our baseline sample includes only observations with non-missing values in the years of education, the number of children and labor force status. Within each country, we drop birth cohorts with fewer than 10 observations. Hence, we end up with a total of 11,698 individuals.

Our outcome variables measure female labor market attachment at the extensive and intensive margins. The former refers to the actual job situation captured by indicator variables for being employed or self-employed (including working for a family business), retired, unemployed or a homemaker. Similar to other studies (e.g. Blundell et al., 2013), we quantify the intensive margin as total hours usually worked per week in the main job.[11] In addition, we elicit labor survival age using actual age for the working females or age when the last job ended for the retired and the others.[12] To sum up, our measures shed light on the long-run aspects of female participation in the labor market.

We measure educational attainment with the number of years spent in education, which is commonly used in the economic literature (e.g. Powdthavee et al., 2015; Brunello et al., 2016; Fort et al., 2016). The second wave of the SHARE provides direct information on the number of years individuals spent in full-time education. In the fist wave, the years of schooling are derived from respondents' highest educational degree. For the purpose of sensitivity analysis, we also construct an indicator variable for having completed at least upper secondary education according to the the ISCED 1997 classification.

As a measure of realized fertility, we use the number of biological children. By design, the SHARE collected child information from one spouse or partner on behalf of the couple. Given the data at hand, we impute missing values on the joint biological children for another spouse or

[10]Israel is not observed in wave 3.

[11]This excludes meal breaks but includes any paid or unpaid overtime. Observations with 168 hours are dropped. We set missing values to zeros for all who did not do any paid work over the last 4 weeks, except for the permanently sick or disabled.

[12]Observations with the elicited labor survival age below 15 are dropped or replaced with the plausible retirement age (wave 2 only).

partner. Note that the records refer only to the children who were still alive at the time of the interview. This raises the question of child mortality induced by parents' education. In the spirit of Fort et al. (2016), we run a robustness check with an indicator variable of whether a person ever had a biological child, which we derive from the SHARELIFE.[13]

In addition to the key variables described above, we exploit a wealth of individual characteristics that are discussed in the literature as potentially important determinants of educational attainment and adult economic outcomes (e.g. Cunha and Heckman, 2007; Case and Paxson, 2009; Currie, 2009; Smith, 2009; Brunello et al., 2016, 2017). These include actual age, the country of residence, a foreign-born indicator, family background and school performance when aged 10, and childhood health conditions until 15 years old. More precisely, the array of childhood controls includes: the number of rooms in the house, the household size, indicators for living with biological mother, father, siblings, grandparents; indicators for hot running water in the house, few books (at most 10), a skilled breadwinner (legislator, senior official, manager, professional, technician or associate professional), poor health, being vaccinated, hospitalized (for at least one month), as well as having parents or guardians smoking, heavily drinking and/or suffering from mental problems.[14]

Table 2.1 reports summary statistics on the key variables of our evaluation sample composed of females born in 1918–1960. On average, individuals are 62 years old, have two children and completed almost 10 years of full-time education. Only 47% have at least upper secondary education. Among all countries, females in southern Europe (Greece, Italy and Spain) appear the least educated. The sample average employment rate is 26%. However, female labor market participation exceeds 40% in Switzerland, Ireland and Scandinavia (Denmark and Sweden). In terms of hours worked a week and labor survival age, Scandinavia is again above the sample averages of 10 and 51, respectively. A large share of individuals (40%) have already retired. The data suggest that Austria and Eastern Europe (the Czech Republic and Poland) experience the highest retirement rates of 60–70%. The unemployment rates vary between 1–6% across the countries. While the share of housewives is 26% in the overall sample, it reaches above 40% in southern Europe and the Netherlands. To sum up, Scandinavian females engage more in paid jobs, whereas less educated south European women tend more to housework.

We present descriptive statistics on the early-life conditions in Appendix Table 2.A1. In short, the majority lived with their biological mother, father and siblings; they got vaccinated and were healthy in childhood. With respect to the housing situation (hot water supply, the number of books and skilled breadwinner), females in Poland and southern Europe faced the least favorable conditions.

[13]Respondents were asked: “Have you [had another/ever had a] biological child — even one who only lived for a short time?”.

[14]To account for item non-response in the early-life characteristics, we create four missingness indicators for household conditions, parental presence, school performance at age 10 and health situation at age 15, respectively.

Table 2.1: Descriptive statistics of main variables by country

Country	Age	Years of education	Upper second. education	N children	Employed	Retired	Unemployed	Homemaker	Hours worked per week	Labor survival age
AT	62.37	9.93	0.60	1.97	0.13	0.63	0.02	0.20	7.26	50.77
BE	63.04	9.94	0.49	2.15	0.23	0.39	0.05	0.28	7.42	47.46
CZ	62.47	11.35	0.46	1.97	0.24	0.70	0.03	0.00	11.09	54.57
CH	61.06	11.17	0.64	1.98	0.45	0.34	0.01	0.18	14.13	52.56
DE	60.79	12.97	0.81	1.97	0.34	0.36	0.06	0.20	12.54	51.18
DK	62.38	12.46	0.72	2.10	0.41	0.47	0.03	0.03	15.36	55.24
ES	63.89	6.25	0.12	2.58	0.16	0.17	0.04	0.59	6.41	49.24
FR	63.14	8.99	0.49	2.15	0.31	0.46	0.03	0.16	10.54	51.66
GR	60.87	8.46	0.37	1.90	0.21	0.27	0.02	0.48	8.01	51.63
IE	59.10	12.58	0.69	3.13	0.45	0.12	0.03	0.34	15.22	50.69
IT	62.02	6.96	0.24	2.15	0.17	0.40	0.01	0.42	5.94	48.39
NL	60.38	11.29	0.43	2.28	0.28	0.16	0.01	0.45	8.59	46.10
PL	61.59	9.29	0.52	2.61	0.16	0.60	0.04	0.08	7.63	51.71
SE	63.29	10.80	0.52	2.10	0.42	0.51	0.02	0.02	16.71	58.11
All	62.03	9.76	0.47	2.17	0.26	0.40	0.03	0.27	9.72	51.29
N	11698	11698	11673	11698	11698	11698	11698	11698	11256	9592

Note: Each variable is averaged over non-missing values. Country codes: AT = Austria, BE = Belgium, CZ = the Czech Republic, CH = Switzerland, DE = Germany, DK = Denmark, ES = Spain, FR = France, GR = Greece, IE = Ireland, IT = Italy, NL = the Netherlands, PL = Poland, SE = Sweden. 'N' denotes the number of non-missing observations.

2.4 Methodology

We exploit mediation analysis to investigate whether the effect of education on labor market attachment can be at least partially attributed to education-induced realized fertility. Figure 2.1 outlines the studied mechanism, where education can directly (path $D \to Y$) and/or indirectly — through fertility (path $D \to M \to Y$) — influence women's participation in the labor market. In this paper, the natural direct effect quantifies the effect of education that is not mediated by fertility decisions. For the natural indirect effect to operate through fertility, significant associations between education and fertility, as well as between fertility and labor market behavior conditional on education are prerequisite (Judd and Kenny, 1981; Baron and Kenny, 1986). However, causal mediation analysis relies on much stronger identification assumptions than simple associations between the variables in the causal chain.

Figure 2.1: Causal diagram with treatment D, mediator M and outcome Y

$$\begin{tikzcd} D \arrow[r] \arrow[rr, bend right] & M \arrow[r] & Y \end{tikzcd}$$

Following the path diagram above, we decompose the total effect of education on labor market attachment into the natural direct and indirect effects (hereinafter *the direct* and *indirect effects*) using a linear structural equation model:

$$M = \alpha_0 + D\alpha_1 + \mathbf{X}'\alpha_2 + U_1, \tag{2.1}$$

$$Y = \beta_0 + D\beta_1 + M\beta_2 + \mathbf{X}'\beta_3 + U_2, \tag{2.2}$$

where Y measures labor market attachment, D is the number of years of education, M denotes the number of biological children, $\mathbf{X}$ represents a set of baseline covariates, including a second-order polynomial in age, an indicator for being foreign born, controls for early-life conditions outlined in Section 2.3, interview-year and country fixed effects, U_1 and U_2 capture mutually independent error terms. Given linearity in parameters and no treatment-mediator interaction, the direct effect corresponds to β_1 and the indirect effect of education is defined as $\alpha_1\beta_2$. The sum of the direct and indirect effects yields the total effect of education on female labor market attachment.

We use OLS to estimate equations (2.1) and (2.2). In order to account for potential dependencies among women born in the same year and living in the same country, we cluster standard errors at the country-cohort level. Inference for the total and direct effects is based on analytical standard errors.[15] A block bootstrap procedure with 999 replications is further used to obtain standard errors from a generated distribution of indirect effects; p-values are based on the corresponding t-statistics.

[15]For the total effect, we regress the outcome of interest Y on the years of schooling D and control variables X.

To causally interpret the estimated direct and indirect effects, the so-called sequential ignorability assumption (Imai et al., 2010) must hold, which involves conditional independence of the treatment and the mediator. In other words, (i) there are no unobservables that causally affect the treatment and the mediator or the outcome conditional on the observed pretreatment covariates, and (ii) there are no unobservables that jointly affect the mediator and the outcome given the observed treatment and pretreatment covariates. In our empirical setting, we bolster the credibility of sequential ignorability with a rich set of controls that capture information on individual early-life conditions related to ability, family background and health — likely to confound the causal mechanism of interest (see Section 2.3).

Sequential ignorability, like many other identifying assumptions in causal inference, is not directly testable with the observed data.[16] As a robustness check, we perform mediation analysis with distinct instruments for the potentially endogenous treatment and mediator. Compulsory schooling reforms are used as a source of exogenous variation in education (Harmon and Walker, 1995) and sibling sex composition as an instrument for the number of children (Angrist and Evans, 1998). Along the lines of Powdthavee et al. (2013), who examined the direct and indirect effects of education on subjective well-being in Australia, we consider a system of linear equations for the treatment, mediator and outcome. This approach identifies the direct and indirect effects among compliers if both instruments are conditionally independent of unobservables across equations.[17] While valid instruments rule out unobserved confounders not causally affected by the treatment, we cannot in general eliminate unmeasured post-treatment mediator-outcome confounders.[18] As a special case, the natural direct and indirect effects are still identified under no treatment-mediator interaction in the outcome equation.[19]

[16]Unlike the ignorability of the treatment, the ignorability of the mediator may not hold even in randomized experiments (Imai et al., 2010). Thus, several sensitivity analysis techniques have been developed to assess the robustness of empirical findings to potential violations of sequential ignorability. While earlier methodological contributions cover unobserved pretreatment confounders (e.g. Hafeman, 2011; Imai et al., 2010; VanderWeele, 2010; Imai et al., 2010), more recent tools of sensitivity analysis deal with post-treatment confounders of the mediator-outcome relationship (e.g. Albert and Nelson, 2011; Tchetgen Tchetgen and Shpitser, 2012; Imai and Yamamoto, 2013; VanderWeele and Chiba, 2014; Hong et al., 2018).

[17]We obtain the direct effect on treatment compliers and the indirect effect on mediator compliers among treatment compliers. The total effect corresponds to the local average treatment effect among treatment compliers. Using distinct instruments for the treatment and the mediator, Frölich and Huber (2017) have suggested sufficiently strong assumptions to identify the direct and indirect effects on all treatment compliers. However, their nonparametric identification is limited to settings with a binary treatment and continuous or discrete instruments for a mediator.

[18]Likewise, the identification assumptions of Frölich and Huber (2017) do not in general allow estimating the natural direct and indirect effects when unobserved mediator-outcome confounders are causally influenced by the treatment. Under their control function approach, only partial (path specific) direct and indirect effects are identified if unobserved post-treatment confounders are controlled for.

[19]Under the assumption of no interaction effects between the treatment and the mediator on the outcome, controlled and natural direct effects coincide. In this case, controlled and natural direct effects can be estimated even when there are mediator-outcome confounders affected by the treatment (e.g. VanderWeele, 2009). See Pearl (2001)

Moreover, the ignorability of the mediator appears less credible if the mediator does not occur shortly after the treatment. Given a possibly large time span between full-time education and realized fertility, it is plausible that education impacts, for instance, women's health behaviors and/or partner choices, which in turn (directly or indirectly) confound the relationship between realized fertility and labor market outcomes. Still, the homogeneous natural direct and indirect effects of schooling can be identified under the assumption of no education-fertility interaction effects on female labor market decisions later in life.

To relax functional form assumptions imposed in equations (2.1) and (2.2), we apply Huber's (2014) semiparametric approach based on inverse probability weighting (IPW) by the treatment propensity score.[20] In doing so, we permit nonlinear model specifications for the mediator and the outcome, and in particular the possibility of interaction effects between the treatment and the mediator on the outcome. The latter captures potential heterogeneity of the natural direct and indirect effects with respect to the treatment state. For the purposes of this robustness check, we consider an alternative measure of educational attainment, i.e. an indicator for having achieved at least upper secondary according to the ISCED 1997 classification.

In the end, a major challenge for our empirical strategy is the potential sample selection on mortality (enhanced by schooling). On the one hand, older cohorts with fewer biological children may live longer (see, for example, Hurt et al., 2006, for a survey) and thus are overrepresented in our sample. On the other hand, the older cohorts might have had more children who did not survive until the interview. Education itself may influence the mortality of females and/or their children. World War II might also have affected cohorts differentially. We address all these issues with a couple of robustness checks in Section 2.5.

2.5 Results

We decompose the total effect of education into the indirect and direct effects, with the former running through realized fertility and the latter capturing all other channels. Table 2.2 presents our main results when controlling for a full set of characteristics described in Section 2.3. For each outcome of interest, we report the estimated total effect (column 1), direct effect β_1 (column 4) and indirect effect $\alpha_1\beta_2$ (column 7) with the corresponding clustered standard errors and p-values (columns 2–3, 5–6, 8–9, respectively). Furthermore, column 10 shows the contribution of the education-induced fertility to the total effect of schooling for each labor market outcome under consideration.

for the conceptual distinction between the controlled direct effect and the natural direct effect.

[20]The propensity score is defined as a probability of receiving the treatment either given pretreatment covariates or given covariates and the mediator. Hence, the estimation procedure rests on two propensity scores. In addition to sequential ignorability, the identification requires that, conditional on pretreatment covariates, the mediator is not a deterministic function of the treatment, i.e. the so-called common support restriction must hold.

Table 2.2: Effects of education on labor market attachment

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.012	0.001	0.000	0.012	0.001	0.000	0.003	0.001	0.000	2.7	11698
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.008	0.001	0.000	9.7	11698
Unemployed	-0.001	0.001	0.011	-0.001	0.001	0.010	0.000	0.000	0.882	-0.4	11698
Homemaker	-0.016	0.001	0.000	-0.015	0.001	0.000	-0.012	0.002	0.000	7.2	11698
Hours worked	0.399	0.052	0.000	0.385	0.052	0.000	0.144	0.034	0.000	3.6	11256
Labor survival age	0.309	0.040	0.000	0.285	0.040	0.000	0.239	0.047	0.000	7.7	9592

Note: 'te, 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

In general, education positively relates to female labor market attachment both at the extensive and intensive margins. More precisely, an additional year of schooling increases the employment probability by 1.2 percentage points on average, and the indirect effect accounts for 2.7% of this effect. The estimated coefficients are highly significant. Furthermore, the total effect of schooling on retirement is significantly positive but rather modest in size.[21] The contribution of education-induced fertility amounts to 9.7%. We also find that the unemployment rate significantly drops by 0.1 percentage points with an extra year of education. However, the corresponding indirect effect appears economically and statistically negligible. Consistent with previous results, the likelihood to stay at home falls by 1.6 percentage points for women with one additional year of schooling. The mediation effect explains 7.2% of this reduction.

At the intensive margin, the number of hours worked per week on average increases by 0.4 (equivalent to 24 minutes), out of which 3.6% are attributed to education-induced fertility (about 0.9 minutes). Regarding labor survival age, women with one more year of schooling remain employed 0.3 years (3.6 months) longer, and the indirect effect contributes 7.7% (about 0.3 months). These effects are highly significant. The above conclusions remain robust once we re-estimate the effects without controlling for early-life conditions (see Appendix Table 2.A2). Note that a large part of the estimated total effects remains unexplained by the education-induced fertility.[22]

In addition, Table 2.3 reports the estimated effects of education on fertility (column 1) and the effects of fertility on the labor market outcomes (column 4) that produce the indirect effects in the main results discussed above. On the one hand, more educated women have significantly fewer children. On the other hand, females with more children engage in paid jobs significantly less. They are also significantly less likely to be retired. Overall, our results suggest that the total effects of education on female labor market attachment are partially driven by reduced fertility.

[21] More educated older women may be more likely to afford retiring, for example, because of higher (household) net worth.

[22] This suggests possible avenues for future research of other potential channels beyond the scope of this paper, which primarily focuses on the indirect effect of education operating though fertility on female labor market behavior at an older age.

Hence, there is some evidence in our data that the substitution effect dominates the income effect if children are normal goods.

Table 2.3: Decomposition of indirect effects from Table 2.2

	Education (1)	st. error (2)	p-value (3)	Fertility (4)	st. error (5)	p-value (6)	N (7)
Employed	-0.034	0.004	0.000	-0.010	0.002	0.000	11698
Retired	-0.034	0.004	0.000	-0.023	0.003	0.000	11698
Unemployed	-0.034	0.004	0.000	-0.000	0.001	0.880	11698
Homemaker	-0.034	0.004	0.000	0.034	0.003	0.000	11698
Hours worked	-0.033	0.004	0.000	-0.432	0.098	0.000	11256
Labor survival age	-0.030	0.004	0.000	-0.809	0.106	0.000	9592

Note: 'Education' denotes the effect of years of schooling on the number of children in equation 2.1. 'Fertility' refers to the effect of the number of children on the labor market outcomes in equation 2.2. Standard errors are clustered at the country-cohort level. Two-sided p-values are based on t-statistics. 'N' denotes the number of observations.

To check sensitivity of our findings to the sample composition, we perform mediation analyses by excluding one country after the other from the evaluation sample. Appendix Tables 2.A3–2.A16 provide the results of these estimations. By and large, the estimated coefficients across the leave-one-country-out specifications are comparable to the results in Table 2.2. This points to the robustness of our main conclusion about positive relation between education and labor market attachment mediated by realized fertility.

2.5.1 Selective mortality

Individuals of advanced age in our sample may be subject to selective mortality. Since mortality is usually higher among people with poor health, survivors must have better health than the average individual of the same age. One potential channel involves positive effect of education on health behaviors, which in turn translates into higher survival probability.[23] Further channel may run through fertility in that less fertile women may live longer.[24] In our main specification, we have accounted for the observed age-linked patterns with a second-order polynomial in age.

In addition, cohorts born before or during World War II are particular in that they might have suffered from the wartime destruction, occupation, displacement, hunger or another distress. To capture cohort-specific effects of the war in our sample, we add cohort fixed effects to the analysis. As displayed in Table 2.4, the estimated results are generally consistent with our main findings.

As a next robustness check, we restrict our sample to cohorts born 1946 and afterwards because

[23]Brunello et al. (2016) find that the mediating effects of health behaviors account in the short run for 17.2% and in the long run for 22.8% of the total effect of education on health for women. At the same time, Gathmann et al. (2015), who examine 18 compulsory schooling reforms in Europe, report little or no effect of education on female mortality.

[24]We do not necessarily mean infertile women since there is no information about partners' reproductive abilities or preferences.

Table 2.4: Mediation analysis with cohort fixed effects

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.012	0.001	0.000	0.012	0.001	0.000	0.003	0.001	0.000	2.6	11698
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.008	0.001	0.000	10.5	11698
Unemployed	-0.001	0.001	0.012	-0.001	0.001	0.012	0.000	0.000	0.922	-0.3	11698
Homemaker	-0.016	0.001	0.000	-0.015	0.001	0.000	-0.012	0.002	0.000	7.3	11698
Hours worked	0.404	0.051	0.000	0.390	0.051	0.000	0.138	0.034	0.000	3.4	11256
Labor survival age	0.311	0.039	0.000	0.288	0.039	0.000	0.235	0.047	0.000	7.6	9592

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

younger cohorts must be less prone to selective mortality.[25] The estimates in Table 2.5 confirm our main results that schooling is positively associated with labor market participation. Compared to the main specification, the estimated effects for the employment rate and hours worked have roughly doubled, whereas the effect on the retirement has dropped. As before, the indirect effects are significant for all outcomes except for the unemployment rate.

Table 2.5: Mediation analysis for cohorts born after 1945

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.025	0.003	0.000	0.025	0.003	0.000	0.008	0.002	0.000	3.2	5161
Retired	0.002	0.001	0.117	0.002	0.001	0.199	0.004	0.001	0.006	17.1	5161
Unemployed	-0.002	0.001	0.022	-0.002	0.001	0.024	-0.000	0.001	0.596	1.7	5161
Homemaker	-0.021	0.002	0.000	-0.020	0.002	0.000	-0.011	0.002	0.000	5.5	5161
Hours worked	0.767	0.092	0.000	0.731	0.092	0.000	0.356	0.092	0.000	4.6	4883
Labor survival age	0.367	0.050	0.000	0.340	0.050	0.000	0.268	0.069	0.000	7.3	4401

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

The detected decline in employment and hours worked with age is unlikely to reflect the mortality patterns because older cohorts in our main sample might be more educated and healthier than the cohort average. One possible explanation of the observed differences by age group is the deterioration of physical health and cognitive abilities with age, which leads to lower employability of older cohorts.

So far, we have solely discussed selective mortality among respondents themselves. Given that the fertility measure in the main specification only captures children still alive at the time of the interview, selective mortality of respondents' children poses another concern. There might have been some children who had passed away before the interview took place. If the number of alive children is not affected by respondents' education, the previous sensitivity analyses are again

[25] Previous studies have also tackled selective mortality with a restricted sample of younger cohorts (e.g. Schneeweis et al., 2014; Brunello et al., 2016; Fort et al., 2016). The results do not substantially alter when we restrict the sample to cohorts born 1940 and later or to females at most 69 years old.

applicable. However, if better educated mothers tend to follow healthier behaviors and to care more about their children's health, mothers' education may reduce child mortality. This type of confounding being a severe problem, we would have instead observed positive relationship between schooling and fertility. Nevertheless, similar to Fort et al. (2016), we conduct a robustness check by replacing the number of alive children with an indicator variable of whether a person ever had a biological child. Table 2.6 presents the estimated effects. In general, the results are in line with the main findings. But the indirect effects of schooling are now much smaller. This might indicate that the indirect effects primarily operate through the intensive fertility margin.

Table 2.6: Mediation analysis using an indicator for ever having biological children

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.012	0.001	0.000	0.012	0.001	0.000	0.001	0.000	0.014	1.0	11686
Retired	0.008	0.001	0.000	0.008	0.001	0.000	0.002	0.001	0.003	2.7	11686
Unemployed	-0.001	0.001	0.012	-0.001	0.001	0.012	0.000	0.000	0.683	-0.5	11686
Homemaker	-0.016	0.001	0.000	-0.016	0.001	0.000	-0.004	0.001	0.001	2.3	11686
Hours worked	0.396	0.052	0.000	0.388	0.052	0.000	0.080	0.030	0.008	2.0	11244
Labor survival age	0.313	0.040	0.000	0.304	0.040	0.000	0.087	0.029	0.002	2.8	9581

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

2.5.2 Heterogeneous effects

We have previously assumed that schooling effects are homogeneous across different age groups. However, the results in Table 2.5 rather point to effect heterogeneity. In order to better understand how effects vary with age, we split our sample into three subsamples of females aged 45–54, 55–64 and 65 plus. The corresponding point estimates are reported in Appendix Tables 2.A17–2.A19.

We find that the estimated total effects of schooling on the employment rate, the probability of staying at home and hours worked weaken with age, whereas the effect on the retirement rate increases. Furthermore, the relationship between education and the unemployment rate exhibits a U-shaped pattern, and a reverse U-shaped association emerges for labor survival age.

The 45–54 and 55–64 year olds have significantly different total and direct effects of schooling on the probability of employment ($p < 0.05$, two-sided tests) and retirement ($p < 0.01$, two-sided tests); the total effects on labor survival age only marginally differ ($p < 0.1$, two-sided tests). Differences in the indirect effects are significant for the retirement rate ($p < 0.05$, two-sided tests), the employment rate and hours worked ($p < 0.1$, two-sided tests).

Moreover, we document highly significant differential total, direct and indirect effects between females aged 45–54 and those aged 65 plus for the employment and retirement rates as well as hours worked ($p < 0.01$, two-sided tests). The estimated total and direct effects also significantly differ between these age groups for the likelihood of staying at home ($p < 0.01$, two-sided tests).

Comparing the two older groups (55–64 vs. 65 plus), we find significant differences in the estimated total and direct effects for the employment and unemployment rates, hours worked ($p < 0.01$, two-sided tests), the probability of staying at home ($p < 0.05$, two-sided tests) and labor survival age ($p < 0.1$, two-sided tests). Despite the documented statistical differences, the estimated effects across three age groups are economically not very distant, in particular for those below 65.

Our linear mediation model also assumes no interaction between education and fertility in their effects on female labor supply, implying homogeneous schooling effects on more and less educated women. To investigate potential heterogeneous effects between more and less educated women, we use a binary measure for educational attainment, i.e. an indicator for having completed at least upper secondary education according to the ISCED 1997 classification. Following Huber's (2014) approach, we relax linear model assumptions for realized fertility and labor market outcomes by allowing for the interaction effects of schooling and fertility on the outcomes. This yields separate direct and indirect effects for more and less educated women.[26]

Table 2.7 displays the results of this robustness check. Here, females with at least upper secondary education belong to the *high* group, and all others to the *low* group. We report the estimated direct and indirect effects for both groups. On the whole, our main findings remain unchanged: more educated women generally exhibit higher participation in the labor market, and education-induced fertility explains a modest part of the total effects. More precisely, the attainment of upper secondary education and above increases employment by 9.3 percentage points and reduces unemployment by 1 percentage point. For more educated women, the likelihood of retirement is higher by 6.5 percentage points, but the probability of staying at home decreases by 11.8 percentage points. Females with at least upper secondary schooling on average work 3 hours more and 2 years longer.

In addition, we observe mainly homogeneous direct and indirect effects under the high and low schooling states. In terms of size, effect heterogeneity is most pronounced for the indirect effects on the retirement rate and hours worked. On the one hand, education-induced fertility leads to 2.5 times higher retirement rate among less educated females. On the other hand, more educated women work two times more hours a week in a paid job. However, the absolute differences between more and less educated females are rather small in economic terms and statistically insignificant. The spotted (negligible) effect heterogeneity may arise from differences in occupational choices and health status enhanced by schooling.

We have previously assumed the absence of post-schooling confounders in our mediation analyses. Though this assumption is not testable, we may assess sensitivity of our results to the inclusion of a potential post-schooling confounder. An indicator variable of whether a person has never

[26] As explained in Section 2.4, we apply inverse probability weighting by the treatment propensity score. The estimation is based on the "medweight" function from the "causalweight" package for the statistical software "R" by Bodory and Huber (2018), which we slightly modified to facilitate the inclusion of interview-year and country fixed effects. Propensity scores are estimated using logit regression.

Table 2.7: Heterogeneous effects of education on labor market attachment

	Total effect	Direct effect		Indirect effect (x10)		Indirect/Total %		N
		high	low	high	low	high	low	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Employed	0.093	0.090	0.089	0.048	0.038	5.1	4.1	11673
st. error	0.012	0.012	0.012	0.019	0.017			(145)
p-value	0.000	0.000	0.000	0.014	0.025			
Retired	0.065	0.060	0.063	0.020	0.050	3.0	7.6	11673
st. error	0.012	0.012	0.013	0.028	0.018			(145)
p-value	0.000	0.000	0.000	0.486	0.007			
Unemployed	-0.010	-0.010	-0.011	0.005	-0.005	-5.1	4.8	11673
st. error	0.004	0.004	0.004	0.003	0.005			(145)
p-value	0.014	0.018	0.009	0.113	0.343			
Homemaker	-0.118	-0.110	-0.110	-0.082	-0.079	7.0	6.7	11673
st. error	0.012	0.012	0.012	0.024	0.017			(145)
p-value	0.000	0.000	0.000	0.001	0.000			
Hours worked	3.346	3.249	3.143	2.031	0.965	6.1	2.9	11233
st. error	0.423	0.433	0.418	0.734	0.498			(140)
p-value	0.000	0.000	0.000	0.006	0.053			
Labor survival age	2.109	1.875	1.905	2.043	2.344	10.0	11.1	9570
st. error	0.370	0.362	0.385	0.715	0.578			(51)
p-value	0.000	0.000	0.000	0.004	0.000			

Note: 'high' refers to females with at least upper secondary education, 'low' to all others. Standard errors are clustered at the country-cohort level using 999 block-bootstrap replications. Two-sided p-values are based on t-statistics. 'N' denotes the number of observations. Trimmed observations with propensity scores below 0.01 or above 0.99 are reported in parentheses.

been married is a plausible candidate because education may affect marital status, which in turn influences fertility and labor supply decisions. For example, additional education may increase the likelihood of marriage (with an educated partner), which can positively affect fertility because of higher household income but then (also) potentially leads to lower female labor market participation. Appendix Table 2.A20 presents the estimated effects. In summary, accounting for marital status does not substantially alter the total effects of schooling on labor market attachment, but not surprisingly the indirect effects drop. We only observe highly significant partial indirect effects on labor survival age and the likelihood of being a homemaker for both more and less educated females. This sensitivity analysis reminds that the omission of important post-treatment confounders may lead to biased estimates.

2.5.3 Endogenous schooling and fertility

We have so far assumed that education and fertility are conditionally exogenous. This untestable assumption relies on the richness of pre-schooling characteristics that are controlled for. We now consider an alternative strategy that allows for endogenous schooling and fertility. That is, two distinct instruments are used to handle potential endogeneity of education and fertility. In doing so, we attempt to rule out not only pre-schooling unobserved confounders but also post-schooling unobserved confounders that are not affected by educational attainment.

This strategy combines variation in compulsory schooling laws across countries and cohorts with parental preferences for a mixed gender composition of children. More precisely, we instrument individual years of schooling with the number of years of compulsory education and the number of children with an indicator variable for having the first two children of the same sex. The two instruments are commonly used in the economic literature on (total) causal effects of education and fertility (see Section 2.2), respectively.

We implement this empirical approach by making some sample adjustments. First, we focus on seven countries where respondents in our sample experienced at least one compulsory school reform: Austria, the Czech Republic, Denmark, France, Italy, the Netherlands and Sweden.[27] Appendix Table 2.A21 presents a brief summary of the reforms under consideration. Second, we restrict our sample to cohorts born up to 10 years before and after each reform (e.g. Gathmann et al., 2015; Brunello et al., 2016).[28] Third, we analyze individuals with at least two children.

The usual IV assumptions must hold at least conditional on pre-schooling covariates (see Imbens and Angrist, 1994). While relevance of the instruments can be empirically checked, monotonicity and exclusion restriction are not testable. Without assuming effect homogeneity in the population, the IV-based approach allows identifying average effects among compliers (Angrist et al., 1996). In our mediation analysis, we can only estimate the total and direct effects on the reform compliers as well as the indirect effects on the sibling sex-mix compliers among the reform compliers. Though the first stage estimates appear statistically significant at the 1% level (see Appendix Table 2.A22), we do not find any statistically significant effects of education on labor market attachment in Appendix Table 2.A23 because the IV estimates are imprecise. In absolute terms, the IV strategy yields larger estimated effects than the main results, but the differences are statistically insignificant ($p > 0.1$, two-sided tests). Because of a relatively small number of observations available per each country-cohort (20 on average), the IV results should be taken with caution.

2.6 Conclusion

In this paper, we study long-term relationship between education and female labor market attachment mediated by realized fertility. Mediation analysis is applied to decompose the total effect of schooling into the indirect effect operating through fertility and the direct effect capturing all other mechanisms (e.g. health behaviors, household total net worth, marital

[27]We exclude Germany and Switzerland because the implementation of educational reforms took place at the regional level there and the SHARE data lack precise information on the region where individuals completed compulsory schooling.

[28]Due to data limitations, the estimation windows are not always symmetric. In addition, the first cohorts potentially affected by the reforms are included into the sample. However, the exclusion of the pivotal cohorts does not alter the conclusions. The results are available on request.

stability).[29] Our identification strategy relies on the sequential ignorability assumption, implying that all confounders of education and fertility decisions are observed, supported with the rich information on individual early-life conditions provided by the SHARE.

We find that labor market attachment of older women significantly increases with higher educational attainment. This positive relationship holds both at the extensive and intensive margins of labor supply. That is, better educated females on average exhibit higher employment rates, are less likely to be homemakers, work more hours a week and exit the labor market at an older age. Empirical evidence also suggests that education-induced fertility explains about 3–8% of the total schooling effects, depending on the outcome. These indirect effects are driven by fertility reduction. Thus, our findings provide some evidence that the substitution effect dominates the income effect if children are normal goods. A number of robustness checks with respect to sample composition, selective mortality and functional form corroborate our main results.

Our findings have two important implications. First, human capital investments yield positive returns in terms of female labor market attachment at an older age. Second, a decrease in realized fertility enhanced by schooling modestly contributes to female labor force participation. Therefore, incentives are necessary to facilitate motherhood for women pursuing additional education or paid employment. This paper leaves scope for further research to investigate other channels that might mediate the effect of schooling on the labor supply of older women.

[29] Note that fertility decisions likely confound other channels. For instance, mothers may tend to follow healthier lifestyles than childless women, and household total net worth obviously depends on the number of dependent children. In addition, couples with no or a large number of children may more often experience marital dissolution than couples with a moderate number of children (e.g. Heaton, 1990).

2.A Additional tables

Table 2.A1: Descriptive statistics of early-life characteristics by country

Country	N rooms	Household size	Mother	Father	Siblings	Grandparents	Hot water	Few books
AT	3.14	5.13	0.92	0.77	0.74	0.20	0.25	0.45
BE	5.20	5.88	0.97	0.94	0.84	0.14	0.30	0.46
CZ	2.40	4.77	0.97	0.90	0.84	0.19	0.23	0.14
CH	4.85	5.77	0.97	0.97	0.90	0.10	0.65	0.30
DE	3.74	4.95	0.96	0.84	0.79	0.17	0.31	0.28
DK	4.44	5.11	0.95	0.90	0.77	0.04	0.49	0.23
ES	3.37	6.28	0.95	0.90	0.91	0.19	0.14	0.64
FR	4.03	5.44	0.94	0.89	0.82	0.15	0.43	0.45
GR	2.74	5.40	0.99	0.97	0.90	0.17	0.17	0.64
IE	4.64	6.94	0.97	0.93	0.94	0.11	0.46	0.32
IT	2.97	6.22	0.98	0.94	0.88	0.19	0.19	0.74
NL	4.59	6.44	0.97	0.93	0.86	0.06	0.50	0.27
PL	2.00	5.73	0.97	0.89	0.88	0.13	0.06	0.59
SE	3.57	4.93	0.95	0.89	0.81	0.09	0.58	0.21
All	3.61	5.62	0.96	0.91	0.85	0.14	0.31	0.45
N	11575	11626	11493	11493	11493	11493	11650	11588

	Skilled breadwinner	Maths skills	Language skills	Vaccinated	Poor health	Hospitalized	Parent smoke	Parent drank/ mental issues
AT	0.11	3.18	3.47	0.95	0.12	0.10	0.52	0.10
BE	0.13	3.32	3.47	0.97	0.09	0.04	0.70	0.11
CZ	0.13	3.33	3.47	0.96	0.06	0.09	0.45	0.06
CH	0.16	3.28	3.61	0.95	0.12	0.08	0.53	0.12
DE	0.12	3.23	3.47	0.99	0.12	0.09	0.57	0.12
DK	0.17	3.41	3.63	0.99	0.07	0.07	0.81	0.18
ES	0.07	3.05	3.18	0.91	0.12	0.02	0.60	0.09
FR	0.16	3.06	3.32	0.97	0.13	0.04	0.50	0.10
GR	0.05	3.02	3.17	0.89	0.01	0.01	0.59	0.06
IE	0.18	3.23	3.50	0.97	0.04	0.09	0.63	0.16
IT	0.06	3.06	3.21	0.96	0.07	0.04	0.59	0.11
NL	0.19	3.29	3.34	0.90	0.11	0.09	0.84	0.08
PL	0.06	3.22	3.36	0.95	0.09	0.07	0.62	0.08
SE	0.17	3.44	3.61	0.97	0.09	0.09	0.51	0.08
All	0.12	3.21	3.38	0.95	0.08	0.06	0.61	0.10
N	11442	11300	11291	11529	11662	11652	11658	11658

Note: Each variable is averaged over non-missing values. Country codes: AT = Austria, BE = Belgium, CZ = the Czech Republic, CH = Switzerland, DE = Germany, DK = Denmark, ES = Spain, FR = France, GR = Greece, IE = Ireland, IT = Italy, NL = the Netherlands, PL = Poland, SE = Sweden. 'N' denotes the number of non-missing observations.

Table 2.A2: Mediation analysis without controlling for early-life conditions

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.014	0.001	0.000	0.014	0.001	0.000	0.004	0.001	0.000	3.1	11698
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.011	0.002	0.000	14.6	11698
Unemployed	-0.002	0.000	0.000	-0.002	0.000	0.000	-0.000	0.001	0.977	1.0	11698
Homemaker	-0.017	0.001	0.000	-0.016	0.001	0.000	-0.016	0.002	0.000	9.3	11698
Hours worked	0.480	0.049	0.000	0.461	0.048	0.000	0.192	0.047	0.000	4.0	11256
Labor survival age	0.314	0.035	0.000	0.281	0.035	0.000	0.328	0.056	0.000	10.5	9592

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A3: Mediation analysis excluding Austria

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.013	0.001	0.000	0.012	0.001	0.000	0.003	0.001	0.000	2.6	11321
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.007	0.001	0.000	9.2	11321
Unemployed	-0.001	0.001	0.014	-0.001	0.001	0.013	0.000	0.000	0.787	-0.8	11321
Homemaker	-0.016	0.001	0.000	-0.015	0.001	0.000	-0.011	0.002	0.000	6.7	11321
Hours worked	0.398	0.053	0.000	0.383	0.053	0.000	0.143	0.035	0.000	3.6	10883
Labor survival age	0.300	0.040	0.000	0.277	0.040	0.000	0.228	0.046	0.000	7.6	9260

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A4: Mediation analysis excluding Belgium

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.013	0.002	0.000	0.013	0.002	0.000	0.004	0.001	0.000	3.0	10437
Retired	0.007	0.001	0.000	0.007	0.001	0.000	0.008	0.001	0.000	10.5	10437
Unemployed	-0.001	0.000	0.185	-0.001	0.000	0.193	-0.000	0.000	0.709	2.4	10437
Homemaker	-0.017	0.001	0.000	-0.015	0.001	0.000	-0.011	0.002	0.000	6.7	10437
Hours worked	0.415	0.057	0.000	0.398	0.056	0.000	0.166	0.040	0.000	4.0	10043
Labor survival age	0.296	0.041	0.000	0.274	0.041	0.000	0.222	0.047	0.000	7.5	8562

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A5: Mediation analysis excluding Switzerland

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.013	0.001	0.000	0.012	0.001	0.000	0.003	0.001	0.000	2.5	11256
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.009	0.001	0.000	10.4	11256
Unemployed	-0.001	0.001	0.009	-0.001	0.001	0.008	0.000	0.000	0.821	-0.6	11256
Homemaker	-0.016	0.001	0.000	-0.015	0.001	0.000	-0.012	0.002	0.000	7.3	11256
Hours worked	0.407	0.054	0.000	0.394	0.053	0.000	0.134	0.035	0.000	3.3	10822
Labor survival age	0.312	0.041	0.000	0.288	0.041	0.000	0.234	0.046	0.000	7.5	9182

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A6: Mediation analysis excluding the Czech Republic

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.013	0.001	0.000	0.012	0.001	0.000	0.003	0.001	0.000	2.6	10861
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.008	0.001	0.000	10.1	10861
Unemployed	-0.001	0.001	0.010	-0.001	0.001	0.010	0.000	0.000	0.775	-0.8	10861
Homemaker	-0.017	0.001	0.000	-0.015	0.001	0.000	-0.012	0.002	0.000	7.3	10861
Hours worked	0.426	0.052	0.000	0.410	0.052	0.000	0.155	0.036	0.000	3.6	10441
Labor survival age	0.323	0.042	0.000	0.298	0.042	0.000	0.248	0.050	0.000	7.7	8771

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A7: Mediation analysis excluding Germany

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.013	0.001	0.000	0.012	0.001	0.000	0.003	0.001	0.000	2.5	11006
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.008	0.001	0.000	10.5	11006
Unemployed	-0.001	0.001	0.015	-0.001	0.001	0.015	-0.000	0.000	0.911	0.3	11006
Homemaker	-0.016	0.001	0.000	-0.015	0.001	0.000	-0.012	0.002	0.000	7.1	11006
Hours worked	0.404	0.053	0.000	0.390	0.053	0.000	0.137	0.036	0.000	3.4	10585
Labor survival age	0.322	0.041	0.000	0.298	0.041	0.000	0.244	0.049	0.000	7.6	8954

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A8: Mediation analysis excluding Denmark

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.011	0.001	0.000	0.011	0.001	0.000	0.004	0.001	0.000	3.2	10850
Retired	0.009	0.001	0.000	0.008	0.001	0.000	0.008	0.002	0.000	9.1	10850
Unemployed	-0.001	0.001	0.017	-0.001	0.001	0.016	0.000	0.000	0.945	-0.2	10850
Homemaker	-0.017	0.002	0.000	-0.016	0.001	0.000	-0.012	0.002	0.000	7.3	10850
Hours worked	0.372	0.054	0.000	0.357	0.054	0.000	0.154	0.039	0.000	4.1	10450
Labor survival age	0.298	0.042	0.000	0.272	0.042	0.000	0.266	0.051	0.000	8.9	8770

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A9: Mediation analysis excluding Spain

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.012	0.002	0.000	0.012	0.002	0.000	0.004	0.001	0.000	2.9	10820
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.007	0.001	0.000	8.6	10820
Unemployed	-0.001	0.001	0.023	-0.001	0.001	0.023	-0.000	0.000	0.905	0.3	10820
Homemaker	-0.016	0.001	0.000	-0.014	0.001	0.000	-0.011	0.002	0.000	7.0	10820
Hours worked	0.403	0.054	0.000	0.387	0.054	0.000	0.154	0.038	0.000	3.8	10422
Labor survival age	0.307	0.041	0.000	0.283	0.040	0.000	0.242	0.046	0.000	7.9	9077

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A10: Mediation analysis excluding France

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.014	0.002	0.000	0.013	0.002	0.000	0.003	0.001	0.000	2.3	10684
Retired	0.009	0.001	0.000	0.008	0.001	0.000	0.008	0.002	0.000	9.7	10684
Unemployed	-0.002	0.001	0.004	-0.002	0.001	0.004	-0.000	0.000	0.840	0.5	10684
Homemaker	-0.018	0.002	0.000	-0.016	0.002	0.000	-0.012	0.002	0.000	6.9	10684
Hours worked	0.438	0.057	0.000	0.425	0.057	0.000	0.134	0.039	0.001	3.0	10269
Labor survival age	0.331	0.043	0.000	0.311	0.043	0.000	0.192	0.042	0.000	5.8	8655

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A11: Mediation analysis excluding Greece

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.012	0.002	0.000	0.011	0.002	0.000	0.003	0.001	0.002	2.4	10155
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.009	0.002	0.000	12.6	10155
Unemployed	-0.001	0.001	0.036	-0.001	0.001	0.034	0.000	0.000	0.823	-0.8	10155
Homemaker	-0.015	0.001	0.000	-0.014	0.001	0.000	-0.012	0.002	0.000	8.3	10155
Hours worked	0.394	0.055	0.000	0.382	0.055	0.000	0.125	0.036	0.000	3.2	9749
Labor survival age	0.357	0.043	0.000	0.331	0.043	0.000	0.264	0.052	0.000	7.4	8697

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A12: Mediation analysis excluding Ireland

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.012	0.001	0.000	0.012	0.001	0.000	0.003	0.001	0.000	2.5	11482
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.008	0.001	0.000	9.4	11482
Unemployed	-0.001	0.001	0.011	-0.001	0.001	0.009	0.000	0.000	0.706	-1.0	11482
Homemaker	-0.016	0.001	0.000	-0.015	0.001	0.000	-0.011	0.002	0.000	6.9	11482
Hours worked	0.398	0.053	0.000	0.385	0.052	0.000	0.129	0.034	0.000	3.2	11052
Labor survival age	0.307	0.040	0.000	0.284	0.040	0.000	0.227	0.044	0.000	7.4	9406

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A13: Mediation analysis excluding Italy

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.011	0.001	0.000	0.011	0.001	0.000	0.003	0.001	0.000	3.2	10559
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.006	0.001	0.000	8.3	10559
Unemployed	-0.001	0.001	0.009	-0.001	0.001	0.009	-0.000	0.000	0.886	0.4	10559
Homemaker	-0.014	0.001	0.000	-0.013	0.001	0.000	-0.010	0.001	0.000	7.2	10559
Hours worked	0.369	0.053	0.000	0.355	0.053	0.000	0.138	0.036	0.000	3.8	10130
Labor survival age	0.292	0.041	0.000	0.269	0.041	0.000	0.237	0.048	0.000	8.1	8830

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A14: Mediation analysis excluding the Netherlands

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.011	0.001	0.000	0.011	0.001	0.000	0.003	0.001	0.001	2.7	10758
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.008	0.002	0.000	10.2	10758
Unemployed	-0.002	0.001	0.006	-0.002	0.001	0.005	0.000	0.000	0.779	-0.8	10758
Homemaker	-0.015	0.001	0.000	-0.014	0.001	0.000	-0.012	0.002	0.000	7.8	10758
Hours worked	0.357	0.054	0.000	0.343	0.053	0.000	0.137	0.037	0.000	3.8	10382
Labor survival age	0.259	0.039	0.000	0.236	0.039	0.000	0.237	0.045	0.000	9.1	8778

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A15: Mediation analysis excluding Poland

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.013	0.001	0.000	0.012	0.001	0.000	0.003	0.001	0.000	2.5	10846
Retired	0.008	0.001	0.000	0.008	0.001	0.000	0.007	0.001	0.000	8.7	10846
Unemployed	-0.001	0.001	0.008	-0.001	0.001	0.007	0.000	0.000	0.454	-1.9	10846
Homemaker	-0.017	0.001	0.000	-0.016	0.001	0.000	-0.011	0.002	0.000	6.7	10846
Hours worked	0.419	0.054	0.000	0.405	0.053	0.000	0.139	0.034	0.000	3.3	10486
Labor survival age	0.331	0.042	0.000	0.308	0.041	0.000	0.228	0.050	0.000	6.9	8806

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A16: Mediation analysis excluding Sweden

	te	se	p-val	de	se	p-val	ie (x10)	se (x10)	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.013	0.001	0.000	0.012	0.001	0.000	0.004	0.001	0.000	3.1	11039
Retired	0.008	0.001	0.000	0.007	0.001	0.000	0.008	0.001	0.000	10.0	11039
Unemployed	-0.001	0.001	0.009	-0.001	0.001	0.008	0.000	0.000	0.736	-1.0	11039
Homemaker	-0.016	0.001	0.000	-0.015	0.001	0.000	-0.013	0.002	0.000	7.6	11039
Hours worked	0.403	0.053	0.000	0.386	0.053	0.000	0.174	0.039	0.000	4.3	10614
Labor survival age	0.310	0.042	0.000	0.282	0.041	0.000	0.276	0.048	0.000	8.9	8948

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A17: Mediation analysis for the age group 45–54

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.026	0.003	0.000	0.025	0.003	0.000	0.009	0.003	0.003	3.5	2872
Retired	0.001	0.001	0.174	0.001	0.001	0.224	0.002	0.001	0.107	10.6	2872
Unemployed	-0.001	0.001	0.484	-0.001	0.001	0.478	0.000	0.001	0.898	-1.2	2872
Homemaker	-0.022	0.003	0.000	-0.021	0.003	0.000	-0.010	0.003	0.001	4.7	2872
Hours worked	0.699	0.103	0.000	0.658	0.106	0.000	0.403	0.148	0.006	5.8	2724
Labor survival age	0.309	0.053	0.000	0.290	0.053	0.000	0.185	0.070	0.008	6.0	2461

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A18: Mediation analysis for the age group 55–64

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.017	0.003	0.000	0.017	0.002	0.000	0.003	0.002	0.031	2.5	4445
Retired	0.008	0.002	0.001	0.007	0.002	0.005	0.010	0.003	0.000	12.8	4445
Unemployed	-0.003	0.001	0.003	-0.003	0.001	0.003	-0.000	0.001	0.669	1.5	4445
Homemaker	-0.018	0.002	0.000	-0.017	0.002	0.000	-0.014	0.003	0.000	7.7	4445
Hours worked	0.618	0.099	0.000	0.606	0.098	0.000	0.116	0.063	0.064	1.9	4225
Labor survival age	0.449	0.065	0.000	0.417	0.064	0.000	0.316	0.082	0.000	7.0	3758

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A19: Mediation analysis for the age group 65 plus

	te (1)	se (2)	p-val (3)	de (4)	se (5)	p-val (6)	ie (x10) (7)	se (x10) (8)	p-val (9)	ie/te % (10)	N (11)
Employed	0.001	0.001	0.272	0.001	0.001	0.273	-0.000	0.000	0.869	-1.1	4381
Retired	0.012	0.002	0.000	0.011	0.002	0.000	0.012	0.003	0.000	9.4	4381
Unemployed	0.000	0.000	0.873	0.000	0.000	0.901	0.000	0.000	0.394	24.8	4381
Homemaker	-0.012	0.002	0.000	-0.011	0.002	0.000	-0.011	0.003	0.000	8.9	4381
Hours worked	0.062	0.032	0.054	0.061	0.032	0.062	0.010	0.024	0.659	1.7	4307
Labor survival age	0.262	0.072	0.000	0.243	0.073	0.001	0.191	0.074	0.010	7.3	3373

Note: 'te', 'de' and 'ie' denote total, direct and indirect effects, respectively. Standard errors ('se') are clustered at the country-cohort level. Standard errors for indirect effects are based on 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

Table 2.A20: Heterogeneous effects with never married as a post-schooling confounder

	Total effect	Direct effect		Indirect effect (x10)		Indirect/Total %		N
		high	low	high	low	high	low	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Employed	0.093	0.089	0.088	0.024	0.029	2.6	3.2	11672
st. error	0.011	0.012	0.011	0.015	0.015			(148)
p-value	0.000	0.000	0.000	0.107	0.052			
Retired	0.065	0.059	0.063	0.016	0.032	2.5	5.0	11672
st. error	0.012	0.012	0.012	0.023	0.016			(148)
p-value	0.000	0.000	0.000	0.484	0.041			
Unemployed	-0.010	-0.010	-0.011	0.004	-0.006	-3.7	5.7	11672
st. error	0.004	0.004	0.004	0.003	0.004			(148)
p-value	0.012	0.013	0.008	0.201	0.190			
Homemaker	-0.118	-0.107	-0.108	-0.052	-0.048	4.4	4.0	11672
st. error	0.013	0.013	0.013	0.020	0.014			(148)
p-value	0.000	0.000	0.000	0.010	0.001			
Hours worked	3.337	3.179	3.073	0.936	0.533	2.8	1.6	11232
st. error	0.423	0.433	0.412	0.523	0.457			(148)
p-value	0.000	0.000	0.000	0.073	0.244			
Labor survival age	2.110	1.852	1.888	1.390	1.809	6.6	8.6	9569
st. error	0.376	0.367	0.393	0.606	0.517			(49)
p-value	0.000	0.000	0.000	0.022	0.000			

Note: 'high' refers to females with at least upper secondary education, 'low' to all others. Standard errors are clustered at the country-cohort level using 999 block-bootstrap replications. Two-sided p-values are based on t-statistics. 'N' denotes the number of observations. Trimmed observations with propensity scores below 0.01 or above 0.99 are reported in parentheses.

Table 2.A21: Compulsory school reforms

Country	Reform year	Pivotal cohort	Changes in years of compulsory education
Austria	1962/66	1951	8–9
Czech Republic	1948	1934	8–9
	1953	1939	9–8
	1960	1947	8–9
Denmark	1958	1947	4–7
France	1959/67	1953	8–10
Italy	1963	1949	5–8
Netherlands	1942	1929	7–8
	1947	1933	8–7
	1950	1936	7–8
Sweden	1962	1950	8–9

Note: See Brunello et al. (2016) and Fort et al. (2016) for further details.

Table 2.A22: First stage effects for endogenous schooling and fertility

	Reform IV	st. error	p-value	Sex-mix IV	st. error	p-value	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Employed	0.267	0.082	0.001	0.147	0.030	0.000	2829
Hours worked	0.250	0.092	0.007	0.128	0.030	0.000	2721
Labor survival age	0.255	0.096	0.008	0.138	0.029	0.000	2502

Note: 'Reform IV' refers to the effect of compulsory schooling on individual education, 'sex-mix IV' refers to the effect of the sibling sex composition IV on the number of children. Standard errors are clustered at the country-cohort level using 999 block-bootstrap replications. Two-sided p-values are based on t-statistics. 'N' denotes the number of observations.

Table 2.A23: Mediation analysis with instruments for endogenous schooling and fertility

	te	se	p-val	de	se	p-val	ie	se	p-val	ie/te %	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Employed	0.119	0.102	0.241	0.117	0.105	0.267	0.002	0.015	0.877	2.0	2829
Retired	-0.011	0.098	0.912	-0.020	0.101	0.845	0.009	0.019	0.637	-84.0	2829
Unemployed	-0.005	0.025	0.831	-0.000	0.031	0.990	-0.005	0.018	0.793	92.4	2829
Homemaker	-0.079	0.159	0.619	-0.072	0.152	0.634	-0.007	0.041	0.867	8.7	2829
Hours worked	1.100	5.737	0.848	1.197	5.721	0.834	-0.097	1.083	0.929	-8.8	2721
Labor survival age	1.351	3.252	0.678	1.221	3.357	0.716	0.130	0.802	0.872	9.6	2502

Note: 'te', 'de' and 'ie' denote partial total, direct and indirect effects among compliers, respectively. Standard errors ('se') are clustered at the country-cohort level using 999 block-bootstrap replications. Two-sided p-values ('p-val') are based on t-statistics. 'N' denotes the number of observations.

3 Linguistics and time preferences: The role of language in shaping intertemporal choices[1]

Overview

Recent evidence suggests that the grammatical association of the present and the future in a language correlates with patience across language groups, but the mechanisms driving this relationship remain unclear. We provide novel evidence from incentivized choice experiments assessing the extent to which potential differences are driven by differences in present bias. To this end, we measure time preferences of French and German speakers — two language groups that differ in their encoding of time — from a bilingual region of Switzerland where institutions are shared and socioeconomic conditions are very similar between the two language groups. We find that French speakers are significantly more impatient than German speakers, and differences are particularly pronounced when payments in the present are involved. Estimates of preference parameters of a quasi-hyperbolic discounting model suggest significant differences in both the long-run discount factor and present bias across language groups.

3.1 Introduction

Some of the most important lifetime decisions, such as the degree of human capital acquisition, decisions about healthy lifestyles, or pension savings, involve intertemporal trade-offs and are shaped by individual time preferences (Chabris et al., 2008; Sutter et al., 2013; Golsteyn et al., 2013; Backes-Gellner et al., 2018). Patience has also been shown to significantly correlate with economic outcomes at the country level, such as GDP per capita, entrepreneurial activities, savings and human capital accumulation (Falk et al., 2018). It is therefore important to gain an in-depth understanding of the nature, determinants and origins of individual time preferences.

One factor that has recently been suggested as a driver of heterogeneity in time preferences is language structure (Chen, 2013). In particular, languages that grammatically associate the future and the present, i.e. languages with a so-called weak future-time reference (w-FTR; Thieroff, 2000), are hypothesized to foster future-oriented behavior. Indeed, Chen (2013) finds initial correlational evidence of a positive relationship between w-FTR languages and future-oriented behavior, such

[1] This study was preregistered in the American Economic Association's registry for randomized controlled trials: `https://www.socialscienceregistry.org/trials/2021/history/15724`. This chapter is co-authored with Holger Herz, Martin Huber and Tjaša Maillard-Bjedov. We thank the Department of Education, Culture and Sport in the canton of Fribourg, which approved the study, as well as the administration of the lower secondary school in Murten for supporting this study. This research was funded by the Department of Economics at the University of Fribourg and the Swiss Distance Learning University.

as higher savings rates, wealth and a healthier lifestyle. Further corroborating the relationship between language structure and patience is correlational cross-country evidence in Falk et al. (2018), who find that speakers of w-FTR languages are on average more patient. Sutter et al. (2018) also find stronger discounting among children speaking a strong future-time reference (s-FTR) language in a bilingual Italian city, using incentivized methods for time preference elicitation.

However, our knowledge about the precise mechanisms that explain this relationship remains limited. First, differences in language correlate with differences in culture — language can in fact be seen as a proxy for or one aspect of culture — making it difficult to differentiate whether the observed effects are a direct consequence of language or of other cultural influences on future-oriented behavior. Indeed, Roberts et al. (2015) find that the effect of language on behavior is weaker, but does not disappear, when controlling for the relatedness of languages and culture. Second, time preferences are complex. In particular, impatience could be the result of general differences in patience or in present bias (Frederick et al., 2002; Strotz, 1956; Laibson, 1997; O'Donoghue and Rabin, 1999). However, the effects of language structure have never been analyzed in a way that allows isolating differences in present bias.

We provide the first evidence on this issue by measuring time preferences of students from a bilingual lower secondary school in Murten, a bilingual Swiss city that is partly German and partly French speaking. While German is considered a w-FTR language, French is a s-FTR language. In Murten, individuals across the two language groups are highly integrated, similar in terms of socioeconomic conditions, and they share social and cultural institutions. Thus, we take advantage of a setting that allows us to keep cultural aspects as constant as possible. By systematically varying the time horizon in our experimental preference elicitation, we are able to measure individual preference parameters, assuming quasi-hyperbolic discounting (Laibson, 1997; O'Donoghue and Rabin, 1999), and to assess potential differences in these parameters across the two language groups.

Why should language structure affect discounting and/or present bias? Chen (2013) hypothesizes that speaking about future events as if they were happening now leads w-FTR speakers to perceive future events as less distant, and hence to discount less.[2] However, a change in tense only occurs for comparisons between the present and the future, but not for comparisons between two

[2]German is a good example of a w-FTR language because German speakers frequently refer to the future using present tense forms of verbs, e.g. "Es regnet morgen" which literally translates to "It rains tomorrow". On the contrary, s-FTR languages require speakers to use future tense forms of verbs when referring to future events. Compared to German, French is classified by linguists to have a stronger future-time reference. For example, "Il pleuvra demain" would literally translate to "It will rain tomorrow". Chen (2013) bases his linguistic-savings hypothesis on the so-called Sapir-Whorf hypothesis (Boas, 1940; Sapir, 1949; Whorf, 1956) that postulates that language fundamentally affects our thinking. The related work of Boroditsky (2001) shows that the native language one speaks influences how one thinks, especially about abstract domains like time. Her results suggest that language does not fully define individuals' thoughts and thinking in the strong sense of the Sapir-Whorf hypothesis, but that language is a powerful tool, and certainly plays a role, in shaping thoughts about time.

future events.[3] Consequently, if the perceived difference in distance — induced by the language structure — is truly the cause of measured differences in discounting across language groups, then such differences should primarily be present for intertemporal trade-offs that involve the present. It should be absent, or less pronounced, for intertemporal trade-offs between two future events. Yet, the evidence in Chen (2013), Falk et al. (2018) or Sutter et al. (2018) cannot discriminate whether the observed differences in patience stem from uniform differences in discount rates, or are primarily a manifestation of differences in present bias.

Our results show that, consistent with Chen's (2013) original hypothesis, French speakers indeed discount more strongly. However, differences are more pronounced when present payments are involved. For example, while 41% of French-speaking students prefer CHF 16 today over CHF 20 in four weeks, only 22% of the German-speaking students do so, a highly significant difference of 19 percentage points. When faced with a choice between CHF 16 in four weeks and CHF 20 in eight weeks, the difference in the fraction of students accepting the earlier payment between the two language groups shrinks to 9 percentage points.

Moreover, we find that students in French-speaking classes on average demand CHF 1.69 less to be willing to switch to the earlier amount when the earlier amount is paid today, which is a highly significant difference. But this difference significantly decreases to CHF 1.06 when the trade-off only involves future payoffs.

Using our individual preference parameter estimates, we consistently find that the difference in the long-run discount factor (δ) between language groups is pronounced and statistically significant across all our main specifications and robustness checks. We also find a pronounced and marginally significant difference in present-biasedness (β). In particular, our data suggests that the fraction of present-biased individuals is around 10 percentage points larger among the French-speaking student population. Our data therefore show significant behavioral differences in intertemporal choice behavior across language groups. Moreover, these differences are particularly evident in situations in which immediate payoffs are involved, precisely where the linguistic difference in the encoding of time is present as well.

We also collected a vast array of socioeconomic and demographic characteristics of students to control for alternative factors that may systematically differ between the two language groups, and may in turn affect revealed time preferences. We find that the differences in the exponential discount factor as well as in present-biasedness across language groups remain robust once these factors are controlled for.

Our results might provide a microfoundation for other observed behavioral differences between language groups. First, Eugster et al. (2017) show that French speakers display significantly

[3] To illustrate this point more clearly, "It will rain tomorrow" and "It will rain the day after tomorrow" translate to "Es regnet morgen" vs. "Es regnet übermorgen" in German and "Il pleuvra demain" vs. "Il pleuvra après-demain" in French. In German, both events are referred to in present tense, whereas in French both events are referred to in future tense. Hence, there is no differential treatment *within* a language.

longer unemployment spells than German speakers along the language border in Switzerland, despite an integrated labor market and cross-language-border labor mobility. DellaVigna and Paserman (2005) theoretically and empirically demonstrate that longer unemployment spells can be related to differences in present bias, and Backes-Gellner et al. (2018) provide empirical evidence that more present-biased apprentices are indeed less likely to obtain job offers a few months before completion of their apprenticeship program. Our finding of differences in present-biasedness across language groups may therefore provide a potential microfoundation for the observed differences in unemployment spells in Eugster et al. (2017).

Second, Guin (2017) finds that residents on the German-speaking side of the Swiss language border are more than 11 percentage points more likely to save than similar households on the French-speaking side, and at least 9 percentage points less likely to ever have smoked. Again, both these behaviors are consistent with higher discount rates and present-biasedness.

Third, Erhardt and Haenni (2018) show significant differences in firm start-up rates between founders in Switzerland with German or French-speaking origin. Individuals with ancestry from the German-speaking side of the language border found 20% more firms than individuals with ancestry from the French-speaking side. Theories of entrepreneurship predict that higher patience is associated with a higher likelihood to become an entrepreneur (Doepke and Zilibotti, 2008, 2014), and Andersen et al. (2014) find that entrepreneurs are indeed more patient than non-entrepreneurs. Our evidence therefore suggests that one reason behind the differences in Erhardt and Haenni (2018) could be the difference in patience demonstrated in our data.

Our results also relate to recent studies that examine differences in other economic outcomes between language groups. Also using the Swiss language border as a study setting, Eugster et al. (2011) show that the demand for social insurance is higher among the Latin Swiss population than among the German Swiss population. Brown et al. (2018) conduct a survey among 15-year-old children in Switzerland and find higher financial literacy among German speakers compared to French speakers. German-speaking students also display more patience, but the difference is statistically insignificant.[4] There is further evidence of a direct effect of language on economic outcomes, such as cooperation (Clist and Verschoor, 2017), cognitive biases (Keysar et al., 2012; Costa et al., 2014), moral judgement (Costa et al., 2014) and identity (Aspachs-Bracons et al., 2008; Clots-Figueras and Masella, 2013).

The remainder of the paper is organized as follows. Section 3.2 describes the institutional background. Section 3.3 describes our hypotheses, preference measures and procedures. Section 3.4 reports the results. Section 3.5 concludes.

[4]Their patience measure is a weighted average of survey questions as well as non-incentivized choice experiments.

3.2 Institutional background

There are four official languages in Switzerland: (Swiss) German is mainly spoken in the northeast, center, and parts of the east, French in the west, Italian in the southeast, and Romansh in some parts of the east. The majority of the population speak either German (63%) or French (23%).[5] Neither geographical barriers nor borders between cantons, i.e. states, strictly define the language border between French and German-speaking regions. Three cantons (Bern, Fribourg, and Valais) are bilingual with the majority speaking either German or French. Historically, the language border has been very stable since the late 18th century (Büchi, 2001).

Policies and institutions are predominately set on the federal or cantonal level in Switzerland. Hence, people living in a bilingual canton experience the same political and institutional environment despite belonging to different language groups. We exploit this setting by studying students from a lower secondary school located in Murten, a bilingual city in the bilingual canton of Fribourg.

In the canton of Fribourg, 69% of the population speak French and 27% speak German.[5] The region has been bilingual for centuries. German was the official language from 1483 to 1798. Between 1798 and 1856, German and French alternated as official languages. Since 1857, both languages are recognized as official languages of the canton.[6] Murten is the main city and administrative center of the Lake district, both bilingual and predominantly German speaking. In the Lake district with 35,377 inhabitants, 64% of its population speak German and 34% speak French.[7] Because the region has been bilingual for centuries, the two language groups are deeply intertwined and share the same social and cultural institutions. Catholicism (25%) and Protestantism (44.5%) are the two main religions.[8] The majority of the French-speaking population is Catholic, whereas the majority of the German-speaking population is Protestant.[9]

In school, German is the first foreign language taught to French speakers, and vice versa, starting from the fifth grade, at the age of 8–9. As a consequence, while most inhabitants clearly identify with either German or French as their native language, the great majority of them speak the other official language very well, implying that language is hardly a barrier for social mobility in the region. Residents celebrate the same festivals such as, for instance, the Murten Lights

[5] See Federal Statistical Office, `https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/sprachen-religionen/sprachen.assetdetail.7466554.html`, retrieved 30 September 2019.

[6] See `http://www.fr.ch/ww/de/pub/andere_links/zweisprachigkeit.cfm#i118897`, retrieved 16 August 2018.

[7] See Federal Statistical Office, `https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/tables.assetdetail.7726978.html`, retrieved 30 September 2019.

[8] See `http://www.murten-morat.ch/de/portrait/zahlenundfakten/zahlenfakten/`, retrieved 8 April 2019.

[9] Unfortunately, we were not allowed to elicit religious denominations at the individual level in school. However, we asked students whether religiosity was encouraged by their parents. Only 7% of the participants agreed. We discuss the implications of differences in religion in more depth in the conclusion.

Festival or the Youth Festival "Solennität"[10] that are equivalently animated in both languages.[11] They also attend the same sport and social clubs.[12] The labor market is integrated across the language border as well. For example, there is substantial commuting over the language border, and there are no differences in labor market indicators such as earnings, the job separation rate, unemployment inflow rate, vacancies per worker or job growth (see Eugster et al., 2017).

Murten and nine neighboring municipalities in the Lake district officially belong to the school district of the lower secondary school in Murten, where 68% of students attend classes in German and 32% in French, which closely represents the language composition of the local population. Children between 4 and 15 years attend compulsory school that comprises two years of kindergarten, six years of primary school and three years of lower secondary school. The lower secondary education introduces four tracks of classes (A, B, C and E) that differ by curriculum complexity. Teaching is either in German or in French. Compulsory school attendance depends on the place of residence. Consequently, students in our sample school are not selected on other criteria than residence.

In summary, the school in Murten provides an ideal setting to assess behavioral differences between language groups in an environment in which social and cultural institutions as well as the labor market are shared across language groups, and in which socioeconomic conditions of language groups are highly comparable.

3.3 Theory, hypotheses and data

3.3.1 Time discounting and language

Since Samuelson (1937), exponential discounting has been commonly used in economics to analyze intertemporal choices. Exponential discounting implies that future events are discounted by a constant factor for every unit of time until the event occurs. The exponential discount function is given by $D(t) = \delta^t$. Consequently, individuals discount future outcomes by a factor that increases exponentially over time. A crucial feature of the exponential discount function is that it implies time consistency. The model allows for individual heterogeneity in patience through differences in the exponential discounting parameter δ.

Recently, Chen (2013) proposed that one cause of individual differences in discounting is language. He postulates that languages that grammatically associate the future and the present, i.e. languages with a so-called weak future-time reference (w-FTR; Thieroff, 2000),

[10]See `https://www.festivaldeslumieres.ch` or `https://www.murtenlichtfestival.ch/`, `https://www.regionmurtensee.ch/en/P7969/youth-festival-solennitaet`, retrieved 12 December 2018.

[11]This is nicely illustrated by a short video of the 2018 Murten Lights Festival `https://www.youtube.com/watch?v=_otvXFXbbpU`, retrieved 12 December 2018.

[12]See, for instance, the bilingual websites of the soccer club, tennis club, and for yoga classes `https://www.fcmurten.ch/fr/`, `http://www.tsc-murten.ch/index.php?fr_platzvermietung`, `https://www.yoga-murten-morat.ch/bienvenue/`, retrieved 12 December 2018.

foster future-oriented behavior in terms of savings and other economic outcomes. Indeed, Chen finds initial correlational evidence of a positive relationship between w-FTR languages and future-oriented behavior, such as higher savings rates, wealth and a healthier lifestyle. Contrariwise, languages that grammatically distinguish the future and the present are referred to as languages with a strong future-time reference (s-FTR).

Chen's hypothesis is based on the so-called Saphir-Whorf hypothesis (Boas, 1940; Sapir, 1949; Whorf, 1956) that postulates that language fundamentally affects our thinking. Speaking about future events as if they were happening now leads w-FTR speakers to perceive future events as less distant, which in turn manifests itself in more future-oriented behavior. Consequently, it is hypothesized that perceptional effects triggered through the use of present and past tense in the language induce lower discount rates among w-FTR speakers.

If exponential discounting is a correct representation of individual time preferences, and if language affects these preferences, any difference in time preferences between w-FTR and s-FTR speakers would be represented in the exponential discounting parameter δ. More precisely, one would expect that w-FTR speakers display larger exponential discount factors. This observation leads to our first hypothesis:

Hypothesis 1 (Exponential discounting) *w-FTR (German) speakers have higher estimated exponential discount factors than s-FTR (French) speakers:* $\delta_{DE} > \delta_{FR}$.

Moreover, because of time consistency, one would expect that discount factors only depend on the time distance between the relevant events, independent of how distant these events are from the moment of decision making. However, it appears that people consistently consume more or exercise less tomorrow than they anticipate today. Such behavior is inconsistent with the exponential discounting model. Still, behaviors revealing such present bias can be explained by the fact that people put additional weight on immediate outcomes relative to outcomes occurring in the future. To capture such behavior theoretically, hyperbolic discounting has been introduced into the literature by Strotz (1956) and Laibson (1997). A functional form satisfying the assumptions of hyperbolic discounting is $D(t) = (1 + \alpha t)^{-1}$. Such a discount function implies a declining rate of time preference (see, for example, Frederick et al., 2002, for a survey). In the following, we will focus on a particularly simple version of hyperbolic discounting, the so-called quasi-hyperbolic discounting model (Phelps and Pollak, 1968; Laibson, 1997; O'Donoghue and Rabin, 1999), given by:

$$D(t) = \begin{cases} 1 & \text{if } t = 0 \\ \beta\delta^t & \text{if } t \geq 1 \end{cases} \tag{3.1}$$

This model differentiates between two time discounting parameters, a "classical" exponential discounting parameter δ and a present-bias parameter β. Individuals with $\beta = 1$ are time-consistent exponential discounters, those with $\beta < 1$ are present biased, and those with $\beta > 1$ are future biased.

The essence of Chen's linguistic-savings hypothesis is that the forced *change in tense* in the s-FTR languages causes an increase in perceived distance, and hence an increase in discounting and a decrease in future-oriented behavior. But since such a change only occurs for comparisons between the present and the future, and not for comparisons between two future events, one can argue that behavioral differences should primarily be present for intertemporal trade-offs that actually involve the present. While an exponential discounting model cannot capture potential differences in discount factors over identical time spans depending on their realization relative to today, the quasi-hyperbolic model can. In fact, the quasi-hyperbolic model has the exact same emphasis on the present as the linguistic-savings hypothesis. Discounting between present and future events is exacerbated through β, but discounting between two future events is not.

We therefore hypothesize that the effects of language on behavior should primarily be visible for trade-offs that actually involve the present. Only in such instances do s-FTR languages actually affect the perception of distance between relevant events. Translated into the preference parameters of a quasi-hyperbolic discounting model, one would therefore expect an effect of language on β.

Hypothesis 2 (Quasi-hyperbolic discounting) *w-FTR (German) speakers are less present-biased than s-FTR (French) speakers:* $\beta_{DE} > \beta_{FR}$.

3.3.2 Preference measures

To assess our hypotheses, we elicit individual time preferences, using incentivized choice experiments. To this end, we employed two multiple price lists in which participants faced multiple choices between a (smaller) sooner and a (larger) later reward. Using time-dated monetary rewards to measure time preferences has advantages and disadvantages (see, e.g., Andreoni et al., 2015; Andersen et al., 2008; Andreoni and Sprenger, 2012a,b; Augenblick et al., 2015). In particular, future payments need to be credible and should not involve non-negligible transaction costs, and utility function curvature should be controlled for. Furthermore, using time-dated monetary rewards assumes that subjects treat money like consumption.

Our decision to use time-dated monetary rewards and multiple price lists was driven by important constraints inherent in the field setting. In particular, our access to the subject pool was restricted to school hours. Consequently, one session had to be completed within 95 minutes, including payments. This severely constrained our ability to collect more sophisticated measures of time preferences that do not rely on timed monetary rewards, for example, using real effort in the spirit of Augenblick et al. (2015), or that rely on many more individual decisions than our elicitation method such as convex time budgets (Andreoni and Sprenger, 2012a). Multiple price lists had the obvious advantage that they were implementable in our study setting.

Moreover, Dohmen et al. (2017) find no evidence that choice patterns can be explained by the potential confounds mentioned above in a representative sample of adults in Germany. Further, Balakrishnan et al. (2017) show that measures elicited using multiple price lists and convex time budgets (Andreoni and Sprenger, 2012a) are strongly and highly correlated. So even if one takes the view that other methods are superior, the evidence strongly suggests that timed monetary rewards do provide meaningful proxies for time preferences and are hence useful in environments where more complicated elicitation procedures are unfeasible.

Our study design addresses the specific concerns of credibility, transaction costs, utility function curvature and arbitrage. In particular, we explicitly guaranteed credibility of future payments by an official statement from the University of Fribourg and we had the official endorsement from the school administration. Future payments were mailed in cash to participants' homes on the specified day, and envelopes were already inscribed on the day of the study. Further, we control for risk aversion using additional behavioral measures, and we include an explicit question on credit constraints in our analyses.

Each price list consists of 12 decisions between sooner and delayed payoffs. The sooner payoffs vary between CHF 9 and 20 in steps of CHF 1, whereas the delayed payoff is fixed at CHF 20. In the first price list, participants choose between an immediate payoff and a payoff in four weeks. In the second price list, students decide between a payoff in four weeks and a payoff in eight weeks. We use all decisions as well as switch points in both price lists for reduced form analyses of the relationship between language and revealed time preferences.

We also use individual switch points to infer preference parameters for a quasi-hyperbolic discounting model. To do so, we need to make assumptions on individual utility functions. In particular, we assume that individual intertemporal preferences are represented by the following utility function:

$$U^t(u_t, u_{t+1}, ...) = \delta^t u_t + \beta \sum_{\tau=t+1}^{\infty} \delta^\tau u_\tau, \tag{3.2}$$

where t measures 4-week intervals. Further, we assume that the instantaneous utility in period t is equal to the monetary amount received in that period. We therefore assume that subjects perceive the reception of some monetary amount as an instantaneous utility flow equivalent to this amount.[13] Moreover, the fact that time preference measurements based on timed monetary rewards have substantial predictive power for real world decisions suggests that, as a descriptive model of behavior, the assumption has some validity (see, for example, Sutter et al., 2013; Backes-Gellner et al., 2018; Meier and Sprenger, 2010, 2012; Golsteyn et al., 2013). Halevy (2014) also provides an extensive discussion and argument in favor of using timed monetary rewards to measure time preferences.

To estimate lower and upper bounds for δ and β in equation (3.2), we exploit the switch point in individual decisions between sooner and later payoffs as an indicator of individual preferences.[14]

[13] Such an assumption could, for example, be justified by narrow bracketing in the experiment (see Rabin and Weizsäcker, 2009).

[14] This method can only be applied when subjects displayed a unique switch point, i.e. when their choices were consistent. More than 95% of participants indeed displayed consistent choices. In case of multiple switch points,

To illustrate this, let us consider a price list with all payments in the future. For each subject, we observe the lowest payoff x_1 in four weeks that is preferred to CHF 20 in eight weeks. This implies, using equation (3.2), that $x_1 \geq \delta 20$ and we obtain $\delta \leq x_1/20$. Similarly, we observe the highest payoff x_2 in four weeks that is not preferred to CHF 20 in eight weeks. This implies that $x_2 \leq \delta 20$ and we obtain $\delta \geq x_2/20$. We focus on the upper bound in our analysis.

Now, consider the first price list with immediate and future payments. For each subject, we observe the lowest immediate payoff x_0 that is preferred to CHF 20 in four weeks. This means that $x_0 \geq \beta\delta 20$. Substituting the expression for δ from above yields $\beta \leq x_0/x_1$, i.e. we derive an upper bound on β.[15] To capture present-biased individuals, we additionally create a dummy variable β^* that takes value one for subjects with $\beta < 1$ and zero otherwise.

Students also completed a lottery task with real payoffs, which allows us to elicit individual risk aversion. Participants made 10 decisions between certain payoffs and a coin toss. The certain payoffs vary between CHF 1 and 10 in steps of CHF 1. The coin toss yields CHF 10 in case of "heads" and CHF 0 in case of "tails". A revealed preference for a lower certain payoff over the coin toss indicates stronger risk aversion. For each subject, we observe the highest certain payoff x at which the coin toss is still preferred and define risk aversion by $\rho = 10 - x$.

Finally, we administered a socio-demographic questionnaire including questions on gender, age, migration background, family structure and material conditions, parental background, schooling and cultural values. We use the collected characteristics as additional controls in our analysis.

3.3.3 Procedures

Preferences were measured in nine sessions, conducted in April 2017 at the lower secondary school in the bilingual city of Murten, Switzerland. 496 students aged between 12 and 17 years participated in our study. This corresponds to 88.4% of all students. 70% of participants followed the curriculum in German, and the remainder in French. The participation rate was slightly higher in the German section (91%) than in the French section of the school (83%).

The study was run with pen and paper. Upon arrival into a study room, students took their places at the desks where questionnaires, pens and envelopes were placed. Participants received instructions in their main schooling language. A trained and bilingual instructor then read instructions aloud.[16] Questionnaires were filled out individually and privately. The order of the tasks and questions was the same for all participants. Participants' understanding of the tasks was checked using control questions.

we determined the switch point that would be "most consistent" with the overall choice pattern, and use this most consistent switch point for our parameter estimation. "Most consistent" is defined as the switch point for which the actual choice pattern displays the fewest errors. In 15 out of 992 cases (1.5%), we could not determine a unique "most consistent" switch point. These observations are dropped from the analysis.

[15] There is a small number of participants who always choose the delayed payoff (around 3% of our observations). Because we do not observe the upper bound for these subjects, we use the observed lower bound instead.

[16] All sessions, German and French, were led by the exact same instructor and research team.

On average, each session lasted about 95 minutes, including payment. Participants got paid anonymously for one randomly chosen decision from the two multiple price lists and one decision from the lottery task.[17] If the delayed payment was drawn, participants received a guarantee letter from the Department of Economics of the University of Fribourg stating the amount to be sent via mail at the specified future date. The purpose of the guarantee letters was to raise credibility of the future payments among participants, which is particularly relevant for the first price list where students decided between immediate and delayed payoffs. All other payments were made immediately. On average, participants received CHF 25.74, that is, CHF 19.39 in the discounting tasks and CHF 6.36 in the lottery task.[18]

Our study and key hypotheses were preregistered at the AEA RCT registry (`https://www.socialscienceregistry.org/trials/2021/history/15724`).

3.4 Results

The analyses in Sections 3.4.1 and 3.4.2 were not explicitly preregistered, but we consider them illustrative for the reader and all tests presented are fully in line with the preregistered general hypotheses. The analyses in Sections 3.4.3 and 3.4.4 follow the preregistration. We report two-sided p-values in all tables, but given the preregistration and the clearly one-sided nature of our hypotheses, we often refer to one-sided p-values in the main text.

3.4.1 Descriptive statistics

We start our analysis by comparing choices in the two price lists across French and German-speaking classes. We focus on class language for our main analyses for two reasons. First, the

[17]At the end of the study, each participant rolled first a 24-sided cube to randomly select one decision out of 24 in the two multiple price lists, and then a 10-sided cube to randomly select one decision out of 10 in the lottery task.

[18]After the study was finished, we learned that, in some German classes, teachers encouraged students to contribute the earnings from our study to the class budget, without knowing the details of our study. Obviously, this unwanted intervention made us worry about the validity of our preference measures. In particular, this intervention could *increase* measured patience and hence distort our preference measurement. However, teachers could not enforce contributions because they observed neither participants' choices nor their earnings directly, i.e. what students earned and at what time. If asked, the dominant strategy of students would have been to simply claim to have gotten the lowest possible amount, and in fact to reveal their true preferences in the experiment. Nevertheless, the teacher intervention could potentially have influenced students' decisions during the study. Fortunately, we can identify three otherwise identical pairs of German classes (parallel classes in the same grade level) in the 10th and 11th grades with the feature that one of them was affected by such an intervention and the other one was not. This allows us to test whether student behavior was more patient in classes with teacher interventions. The average switch point in the first price list was 18.6 without teacher interventions and 18.1 in classes with teacher interventions. The average switch point in the second price list was 18.9 without teacher interventions and 18.2 in classes with teacher interventions. We can therefore reject the hypothesis that the intervention increased patience (p-values of one-sided t-tests are $p = 0.117$ and $p = 0.066$, respectively). Hence, any behavioral consequences of this intervention would rather work *against* our hypothesis, which postulates that German speakers are more patient.

class language was also the language used during the study and in the instructions. Consequently, the class language was the language used at the moment of decision making. Second, students form strong social ties within their classes. Therefore, it is conceivable that class language also plays an important part in students' lives outside class, even if class language is not the native language. Finally, we preregistered that we would analyze differences in preferences and behavior by class language. In Section 3.4.4, we provide robustness checks for our results using alternative specifications, in particular native language.

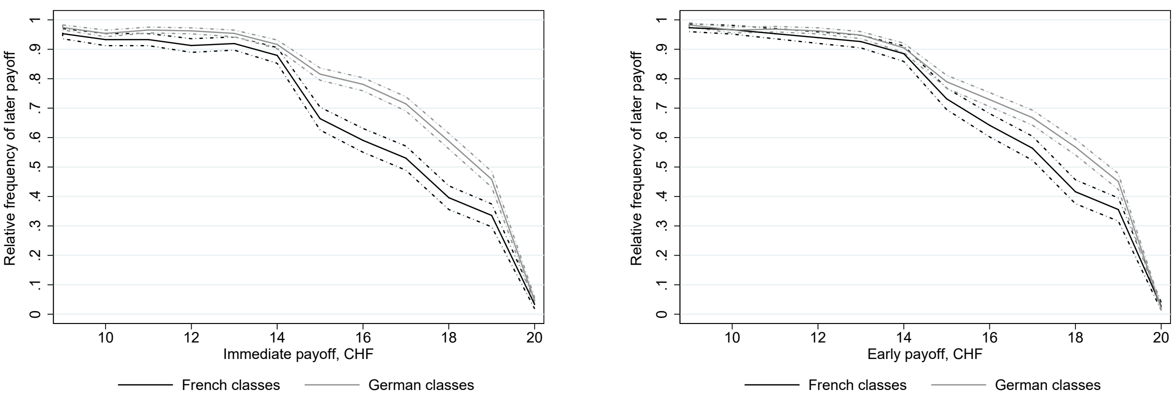
Figure 3.1a illustrates the distribution of individual choices between immediate and delayed payoffs. First, and as one would obviously expect, the immediate payoff is chosen more frequently as the immediate payoff becomes larger. It also appears that students in French-speaking classes are more likely than those in German-speaking classes to switch to the immediate payoffs as the immediate payoff increases. This difference is particularly pronounced for immediate payoffs between CHF 15 and 19.

Figure 3.1b displays the average behavior in each language group for all 12 decisions in the second price list, when students faced a choice between a (weakly) smaller amount in four weeks and an amount of CHF 20 in eight weeks. As before, the majority of students prefer to wait when the immediate amount is small, but they increasingly switch to the earlier payoff as the earlier payoff gets larger. Again, French speakers appear to be less likely than German speakers to wait for a delayed payoff in eight weeks, and the difference becomes most pronounced for sooner payoffs between CHF 15 and 19.

Figure 3.1: Relative frequency of delayed payoff choices by class language

(a) Today vs. 4 weeks

(b) 4 weeks vs. 8 weeks



Note: Each figure includes 496 observations. Dotted lines display one standard deviation of the mean. Appendix Figure 3.A1 presents the relative frequency of delayed payoff choices by native language. Patterns are very similar.

Table 3.1 provides further statistical support for differences in discounting between the two language groups. The reported differences in the relative frequency of the early choice between the two language groups are obtained by regressing the choice of the earlier option on a German class language dummy for each individual choice that students faced. The left panel of Table 3.1 shows the difference by class language for each decision in the first price list, which involved a choice between a varying immediate payoff today and CHF 20 in four weeks. The right panel shows the difference by class language for each decision in the second price list, which involved a choice between a varying payoff in four weeks and CHF 20 in eight weeks.

Table 3.1: Differences in the relative frequency of earlier payoff choices by class language

	Today vs. 4 weeks			4 weeks vs. 8 weeks		
Decision	DE-FR	st. error	p-value	DE-FR	st. error	p-value
9 vs. 20	-0.021	0.033	0.525	-0.010	0.022	0.668
10 vs. 20	-0.021	0.038	0.577	0.001	0.028	0.971
11 vs. 20	-0.033	0.039	0.401	-0.015	0.034	0.651
12 vs. 20	-0.050	0.051	0.327	-0.023	0.034	0.498
13 vs. 20	-0.034	0.051	0.500	-0.022	0.039	0.572
12 vs. 20	-0.050	0.051	0.327	-0.023	0.034	0.498
14 vs. 20	-0.037	0.054	0.489	-0.020	0.044	0.652
15 vs. 20	-0.151	0.076	0.048	-0.058	0.061	0.342
16 vs. 20	-0.190	0.087	0.028	-0.087	0.080	0.276
17 vs. 20	-0.184	0.093	0.049	-0.104	0.087	0.233
18 vs. 20	-0.192	0.086	0.026	-0.152	0.091	0.098
19 vs. 20	-0.123	0.085	0.149	-0.095	0.094	0.314
20 vs. 20	-0.013	0.017	0.470	0.007	0.022	0.765

Note: OLS estimates are based on individual choices from the first price list ('Today vs. 4 weeks') and the second price list ('4 weeks vs. 8 weeks'). Columns 'DE-FR' display estimated coefficients for a German dummy, showing differences in the relative frequency of earlier payoff choices between German and French classes. Standard errors are clustered at the class level. Sample includes 496 observations in 29 classes. Appendix Table 3.A1 reports regression results with heteroscedasticity robust standard errors without clustering.

The results reveal that there is no statistical difference between the two language groups as long as the earlier payoffs are sufficiently small. However, when the earlier payoff ranges from CHF 15 to 18, students in French-speaking classes are significantly more likely to choose the earlier payoff when the earlier payoff is paid out today. For the second price list, in which the earlier payoff is paid out in four weeks, differences between the two language groups are less pronounced, and statistical significance diminishes (for the choice of CHF 18 vs. 20, the difference remains largest and significant in a one-sided t-test, $p = 0.049$).[19]

Result 1 *Students in German-speaking classes display more patient behavior than students in French-speaking classes.*

3.4.2 Analysis of switch points

So far, we have solely focused on the 24 binary decisions between earlier and later amounts, but have ignored other factors that might affect these preferences. While our setting is carefully chosen

[19]Appendix Table 3.A1 reports similar regression results with robust standard errors instead of clustered standard errors. It is not clear whether clustering at the class level is truly warranted in this individual decision-making task. Without clustering, differences in the second price list become significant as well.

such that French and German speakers are as similar as possible, it could be that they nonetheless differ in important dimensions that in turn affect their time preferences. To rule out that our observed differences in discounting are driven by such omitted factors, we now turn to regression analysis. First, we investigate the effects of language on individual switch points in the two price lists, using tobit regressions. For each subject, the most consistent switch point is used as an observation for each price list, implying that the regression includes two observations per subject, the switch point in the first price list and the switch point in the second price list.[20] Moreover, a few students never switch. They either always prefer the earlier or always prefer the later amount.[21] Consequently, their true switch point is either censored above or below, due to our experimental elicitation procedure. To take into account this censoring, we use two-limit tobit regressions and focus on the estimated latent switch point in our regression interpretations.

Table 3.2 provides the results of these estimations. The first regression model (column 1) only includes a dummy for students from German-speaking classes (*German*), a dummy for the delayed price list (4 *weeks*), and an interaction term of these two dummies. The second specification (column 2) additionally controls for risk aversion. Risk preferences are frequently discussed as an important characteristic that can affect time preferences, because future payments may be perceived as inherently risky. Controlling for risk preferences therefore allows us to assess whether differences in discounting are due to differences in time preferences or differences in risk preferences.

In our third specification (column 3), we further add an array of important socio-demographic characteristics that were identified in the literature as potentially related to intertemporal preferences (see, for example, Dohmen et al., 2010; Sutter et al., 2018; Backes-Gellner et al., 2018; Brown et al., 2018). These include age, gender, Swiss citizenship, the number of siblings in the family, and parents' age. We also add dummies for the class grade (9, 10 or 11) as well as track (A, B, C or E). We use several proxies for household income, such as housing conditions (type of housing and availability of an own room) and weeks of holidays spent in the previous year. To control for liquidity constraints that might drive individual intertemporal choices, we asked students to assess the difficulty they have to raise CHF 100 on a 5-point Likert scale (the higher the number indicated the smaller the difficulty) along with their pocket money per week in CHF. Finally, we add a dummy for correctly answering comprehension questions on both price lists during the experiment, and some proxies of personality traits: how trusting participants are (measured using the trust question from the World Values Survey), and whether respondents claimed that their parents encouraged them to be independent, responsible, hardworking, unselfish and religious.[22]

[20] In 15 out of 992 cases (1.5%), we could not determine a most consistent switch point. These observations are dropped.

[21] 2.3% of students switch at the smallest amount in the early price list and 1.6% of students switch at the smallest amount in the delayed price list. 3.5% of students always choose the delayed amount in the early price list, and 2.2% of students always choose the delayed amount in the delayed price list. Using OLS instead of tobit specifications does not substantially alter any of our results, which are available on request.

[22] In Appendix Table 3.A6, we present and compare our socio-demographic controls across language groups. Com-

Table 3.2: Tobit regressions on switch points

	None (1)	Risk (2)	All controls (3)
German	1.688	1.642	1.983
st. error	0.439	0.439	0.491
p-value	0.000	0.000	0.000
4 weeks	0.370	0.390	0.636
st. error	0.298	0.297	0.322
p-value	0.214	0.190	0.048
German x 4 weeks	-0.628	-0.667	-0.804
st. error	0.353	0.353	0.385
p-value	0.076	0.060	0.037
Constant	17.922	17.191	26.966
st. error	0.369	0.680	6.737
p-value	0.000	0.000	0.000
Observations	977	969	824

Note: Two-limit tobit regressions on individual most consistent switch points. Two observations per individual are included (one for each price list), except for 1.5% cases when no most consistent switch point is determined. Standard errors are clustered at the individual level. Results for the full specification of columns 2 and 3 are reported in Appendix Table 3.A2. Regression results with standard errors clustered at the class level are reported in Appendix Table 3.A3.

Column 1 in Table 3.2 provides estimation results for our most basic specification without any further controls. Students in French-speaking classes on average switch to the immediate payment in the early price list when they are offered CHF 17.92 today. Students in German-speaking classes require CHF 1.69 more before they switch, and this difference is highly significant. When all payments are delayed, French-speaking students become slightly more patient, but the effect is not statistically significant. However, the interaction between the German class language dummy and the delayed price list is negative and significant. This implies that students in German-speaking classes only require CHF 1.06 more than students from French-speaking classes before they switch to the earlier payoff in the second price list, which is still a statistically significant effect ($p = 0.014$). The decrease in the difference between the two language groups of CHF 0.63 is also statistically significant.

When additional controls are included in the regression specification (columns 2 and 3), the basic patterns from our simplest specification are confirmed. Controlling for important socioeconomic characteristics, we find that students in French-speaking classes become significantly more patient when all payments are delayed. Moreover, the decrease in the difference in patience between German and French class students is robust to controls and remains significant.[23]

parisons across language groups display interesting patterns. German speakers turn out to be more risk averse and trusting, and they have higher socioeconomic status, proxied by home ownership. Parents of German speaking students are a bit older, and the German-speaking community is composed of fewer foreigners. Hence, these variables constitute important controls to assess in the most accurate way the direct effect of language on time preference parameters.

[23]In Appendix Table 3.A3, we report results for the same tobit specifications with standard errors more conservatively clustered at the class level. Using one-sided tests, the coefficients on German remain significant at the 5%

Taken together, the reduced form analyses in Sections 3.4.1 and 3.4.2 provide strong evidence that students in French-speaking classes display less patience than those in German-speaking classes when immediate payments are involved, but the difference is reduced when all payments are delayed.

Result 2 *The difference in revealed patience between students from French and German-speaking classes is significantly reduced when intertemporal trade-offs only involve future payments.*

3.4.3 Analysis of individual preference parameters

A main feature of our study design is that it enables us to uncover individual preference parameters, assuming time preferences with quasi-hyperbolic discounting, as explained in Section 3.3.2. We can therefore also assess the relationship between language and preference parameters directly.

Table 3.3 presents the mean values of our time preference parameter estimates by class language.[24] The average switch point in French-speaking classes in the second price list (only concerning future payoffs) was 17.4, which translates into a discount factor of $\delta_{FR} = 0.87$, whereas the average switch point in German-speaking classes was 18, which translates into a discount factor of $\delta_{DE} = 0.90$. The difference in discount factors is significant without clustering, but once standard errors are clustered at the class level, significance is weak ($p = 0.121$ in a one-sided test).

Table 3.3: Mean values of time preference parameter estimates by class language

	FR	DE	DE-FR	p-value robust	p-value clustered	Observations
δ	0.868	0.899	0.031	0.019	0.243	492
	(0.138)	(0.126)	(0.027)	0.009	0.121	
β	0.997	1.021	0.024	0.120	0.180	481
	(0.164)	(0.140)	(0.018)	0.060	0.090	
β^*	0.306	0.198	-0.109	0.014	0.042	481
	(0.462)	(0.399)	(0.053)	0.007	0.020	

Note: Outcome variables with most consistent choices. 'DE' and 'FR' stand for German and French, respectively. 'p-value' denotes the significance level of mean difference t-tests of German vs. French ('DE-FR'). The first (second) row of each outcome contains two-sided (one-sided) p-values that are heteroscedasticity robust without clustering and clustered at the class level, respectively. Values in parentheses are standard deviations of outcome values for language groups and standard errors of mean differences. Appendix Table 3.A4 presents mean values of time preference parameter estimates by class language for consistent choices only.

When comparing average estimates of β, we find that French-speaking students have lower β compared to German-speaking students, and the difference is weakly significant ($p = 0.090$, one-sided test).[25] A striking difference across the two language groups appears when comparing the fraction of the respective population that displays some present-biasedness. We define a student as

level in all specifications. The interaction of German with the delayed price list remains marginally significant in specification 2, and significant at the 5% level in specification 3.

[24] As explained in Section 3.3.2, we use the most consistent unique switch point to determine these values. Appendix Table 3.A4 reports the same statistics using only consistent answers. The conclusions remain unchanged.

[25] It also appears that, on average, neither students in German-speaking classes nor students in French-speaking

present biased if the student displays stronger discounting in the first price list (involving immediate payments) compared to the second price list (involving only future payments). Specifically, β^* is defined as a dummy variable that equals one in case a student is present biased in such a manner. The last row in Table 3.3 reveals that 31% of students in French-speaking classes display present-biasedness, whereas only 20% of students in German-speaking classes do. This 11 percentage point difference is statistically significant ($p = 0.020$, one-sided test)

To rule out that our observed differences in individual preference parameters are driven by omitted factors, we now again turn to regression analysis. The average effects of school language on individual time preference parameters are estimated using the following OLS regression:

$$Y = \beta_0 + 1_{\{German\}}\beta_1 + \mathbf{X}'\beta_2 + U, \tag{3.3}$$

where Y is the outcome of interest, i.e. individual time preference parameters δ and β, $1_{\{German\}}$ is an indicator that takes value one for German being a class language and zero otherwise, $\mathbf{X}'$ is a vector of the control variables defined in Section 3.4.2, and U is the error term. For a dummy variable that indicates some degree of present bias β^*, we estimate probit regression to take into account the binary nature of the outcome. In order to allow for potential dependencies among students studying in the same class, we always cluster standard errors at the class level.

In our first regression model, we only control for individual risk preferences. In our second specification, we add the set of control variables specified in Section 3.4.2. Detailed results for all regressors of this specification are presented in Appendix Table 3.A7. While the effects on δ and β in these two specifications are estimated by OLS, the reported effects on β^* correspond to the average marginal probit effects.

In our third specification, we select control variables from our full set of potential control variables (see Table 3.A6) as well as several higher order and interaction terms[26] in a data-driven way (to overcome the curse of dimensionality) by applying the so-called post-regularization approach (see, for example, Belloni et al., 2014; Chernozhukov et al., 2015). The aim is to select from a potentially large pool of variables those controls that importantly predict either the linguistic treatment and/or the outcome. This ensures that regressors that are non-negligibly associated with both the treatment and the dependent variable are taken into account.[27] As before, standard errors are clustered at the class level.

classes are particularly present biased in our study. But there is considerable heterogeneity across students, some displaying considerable present bias, while others have moderate future bias. Overall, 23% of the students display present bias, 26% display future bias, and 51% are time consistent. Meier and Sprenger (2012) find comparable degrees of future bias in their data.

[26]This includes the square and cube of age, the number of siblings squared, as well as gender interacted with age, the number of siblings, pocket money per week, and being born in Switzerland, respectively.

[27]The post-regularization works as follows. First, linear lasso regression is applied separately for predicting the linguistic treatment and the outcome. In contrast to OLS, lasso selects variables by setting the coefficients of less important predictors to zero, based on a penalty term that restricts the sum of absolute values of slope coefficients in the model. Next, the treatment and outcome are predicted by standard OLS using the respective lasso-selected

Table 3.4 shows the estimated coefficients for the German class language dummy in regressions on our parameters of interest.[28] The top panel focuses on the estimated effects on δ. When controlling for socioeconomic variables (column 2), students in German classes reveal a discount factor that is on average 3.8 percentage points larger than that of students in French-speaking classes, which is a sizable and statistically significant difference. In the lasso specification, the coefficient is comparable in size but not statistically significant.[29]

Table 3.4: Effects of class language

	Risk (1)	Controls (2)	Lasso (3)
δ estimate	0.031	0.038	0.034
st. error	0.027	0.018	0.035
2-sided p-value	0.247	0.034	0.321
1-sided p-value	0.123	0.017	0.160
observations	488	415	360
β estimate	0.021	0.020	0.047
st. error	0.017	0.012	0.022
2-sided p-value	0.222	0.099	0.032
1-sided p-value	0.111	0.049	0.016
observations	477	407	354
β^* estimate	-0.101	-0.101	-0.144
st. error	0.050	0.063	0.061
2-sided p-value	0.041	0.110	0.018
1-sided p-value	0.021	0.055	0.009
observations	477	407	354

Note: Sample contains observations with most consistent choices and non-missing values in the respective outcome, language dummy, and control variables. 'Risk': controlling for risk aversion. 'Controls': same controls as in column 3 of Table 3.2. All controls are shown in Appendix Table 3.A7. Columns 1–2 show OLS estimates for β and δ and average marginal probit effects for β^*. 'Lasso': OLS regression with lasso-selected controls. Standard errors are clustered at the class level

The middle and bottom panels display estimated effects of the German class language on β and β^*, respectively. It can be seen that the effects of language on present bias are also pronounced and statistically significant, at least once important socioeconomic controls are included in the regression specification. Column 2 shows that being in a French-speaking class decreases β by 2 percentage points, on average. When looking at the overall incidence of present-biasedness, the bottom panel shows that students in French-speaking classes are 10 percentage points more

regressors, which presents a post-lasso step. The lasso coefficients generally do not correspond to the OLS coefficients obtained in the post-lasso step even when both procedures rely on selected predictors only. The reason is that even among selected predictors, lasso may shrink and thus bias some coefficients towards zero (relative to OLS) to obey the penalization. Finally, δ, β, or β^* are estimated by regressing the residual of the outcome equation on the residual of the treatment equation, thus purging any associations with the control variables. We implement the post-regularization using the "pdslasso" package for the statistical software "STATA" by Ahrens et al. (2018).

[28] Full regression results for specification 2 can be found in Appendix Table 3.A7. The left panel of Appendix Table 3.A8 replicates all three specification using only consistent choices. The conclusions remain similar.

[29] The lasso-selected controls include: whether students attended preschool, the number of years spent in preschool, parents' native languages (French and German).

likely to be present biased than those in German-speaking classes. Both these effects are, at least marginally, significant. In the lasso specification, the point estimates increase further and are highly significant.

Result 3 *Estimated quasi-hyperbolic preference parameters show significant differences in long-run discount factors (δ) across language groups. Moreover, students in French-speaking classes are more likely to be present biased, and their present bias (β) is on average more pronounced.*

3.4.4 Robustness checks

One concern regarding our analysis so far is that we focus on the class language as a predictor of time preferences. However, not all participants in our study are either native French or native German speakers, for example, due to migration background. Table 3.5 only includes students who indicate that German or French are their native languages, which is the case for 89% of our initial sample.[30]

Table 3.5: Effects of class language for native French and German speakers only

	Risk (1)	Controls (2)	Lasso (3)
δ estimate	0.023	0.034	0.026
st. error	0.023	0.018	0.028
2-sided p-value	0.318	0.059	0.348
1-sided p-value	0.159	0.029	0.174
observations	433	372	322
β estimate	0.026	0.028	0.051
st. error	0.019	0.014	0.026
2-sided p-value	0.177	0.049	0.048
1-sided p-value	0.088	0.024	0.024
observations	424	366	318
β^* estimate	-0.114	-0.117	-0.139
st. error	0.055	0.067	0.078
2-sided p-value	0.036	0.079	0.073
1-sided p-value	0.018	0.039	0.037
observations	424	366	318

Note: Sample contains observations with German and/or French native language, most consistent choices, and non-missing values in the respective outcome, language dummy, and control variables. ‘Risk’: controlling for risk aversion. ‘Controls’: same controls as in column 2 of Table 3.4. Columns 1–2 show OLS estimates for β and δ and average marginal probit effects for β^*. ‘Lasso’: OLS regression with lasso-selected controls. Standard errors are clustered at the class level

It can be seen that the point estimates on the German class language dummy remain mostly unchanged. We still find a significant effect of language on the discount factor δ, a moderate and significant positive effect of language on the present-bias parameter β, and a large negative effect on the likelihood of displaying present-biasedness.

[30]Note that it remains possible that a native German speaker attends a French-speaking class, and vice versa. We restrict the sample to native German speakers in German classes and native French speakers in French classes later (see Table 3.8).

A deeper question with respect to Chen's (2013) linguistic-savings hypothesis is whether language inherently changes time preferences through yearlong nurture and exposure, or whether language only triggers a perceptional difference in distance at the moment of decision making. The former hypothesis would suggest that behavioral differences between language groups are inherent and independent of the class language. The latter hypothesis would imply that the class language itself might be the cause of behavioral differences. Using class (and therefore also instruction) language as our primary regressor, we have so far focused on the latter hypothesis. If, however, native language inherently changes time preferences, independently of the currently spoken language, students' native language should be ultimately predictive of differences in time preference parameters.

In our sample, class language is strongly and highly significantly correlated with native language ($\rho = 0.58$ for German-speaking classes and German native language, $\rho = 0.61$ for French-speaking classes and French native language), but still not every student in a German-speaking class is a native German speaker, and vice versa.

Because we are now interested in comparing native German speakers with native French speakers (and not native German speakers with students of any other native language), we again restrict our sample to those students who mention either German or French as one of their native languages. Table 3.6 shows the respective average time preference parameter estimates for French speakers, German speakers and bilinguals (both native French and German speakers).[31] We can see that results are fairly similar compared to our class language distinction. French speakers appear to be less patient (but statistical significance is again weak), and a significantly larger fraction of native French speakers is present-biased in our study (31% vs. 20%). This 11 percentage point difference is again significant ($p = 0.023$, one-sided test). Furthermore, we do not find statistically significant differences in patience or present-biasedness between French-speaking and bilingual students.

The left panel of Table 3.7 shows estimated coefficients for the German native language dummy in regressions on δ, β and β^*, respectively. It can be seen that results are fairly similar to our previous results when using class language as a regressor. Once important socioeconomic characteristics are controlled for, we find a positive and significant effect of German native language on the estimated δ. The effect on β is insignificant in this specification. However, we still find a large negative and significant effect of German native language on the likelihood to be present biased. Compared to our earlier results, only the direct effect of language on the estimated β appears to be weaker in this specification. We also do not find that bilingual students are significantly more patient or less often present biased than native French speakers.

Since the class language does not always coincide with the native language of a student, we finally restrict our sample to those students for whom native language and class language match.

[31] We use the most consistent unique switch point to determine these values. Appendix Table 3.A5 reports the same statistics using only consistent answers. The conclusions remain unchanged.

Table 3.6: Mean values of time preference parameter estimates by native language

	French	German	Bilingual	DE-FR	p-value robust	p-value clustered	bi-FR	p-value robust	p-value clustered	Observations
δ	0.886	0.904	0.900	0.018	0.204	0.301	0.014	0.550	0.532	437
	(0.127)	(0.124)	(0.125)	(0.017)	0.102	0.150	(0.023)	0.275	0.266	
β	1.002	1.016	1.004	0.015	0.424	0.457	0.002	0.940	0.942	428
	(0.166)	(0.145)	(0.103)	(0.020)	0.212	0.228	(0.025)	0.470	0.471	
β^*	0.312	0.198	0.222	-0.114	0.025	0.046	-0.090	0.287	0.255	428
	(0.465)	(0.399)	(0.422)	(0.057)	0.012	0.023	(0.079)	0.143	0.127	

Note: Outcome variables with most consistent choices. 'DE', 'FR' and 'bi' stand for German, French and bilingual, respectively. 'p-value' denotes the significance level of mean difference t-tests of German vs. French ('DE-FR') and bilingual vs. French ('bi-FR'). The first (second) row of each outcome contains two-sided (one-sided) p-values that are heteroscedasticity robust without clustering and clustered at the class level, respectively. Values in parentheses are standard deviations of outcome values for language groups and standard errors of mean differences. Appendix Table 3.A5 presents mean values of time preference parameter estimates by native language for consistent choices only.

Table 3.7: Effects of native language

	German vs. French			Bilingual vs. French		
	Risk	Controls	Lasso	Risk	Controls	Lasso
	(1)	(2)	(3)	(4)	(5)	(6)
δ estimate	0.018	0.027	0.044	0.013	0.006	0.025
st. error	0.018	0.016	0.028	0.024	0.023	0.021
2-sided p-value	0.321	0.085	0.111	0.574	0.797	0.253
1-sided p-value	0.160	0.042	0.056	0.287	0.399	0.126
observations	433	372	322	433	372	322
β estimate	0.008	0.000	-0.004	-0.000	0.004	-0.004
st. error	0.018	0.015	0.036	0.025	0.021	0.035
2-sided p-value	0.658	0.991	0.916	0.986	0.858	0.897
1-sided p-value	0.329	0.495	0.542	0.507	0.429	0.552
observations	424	366	318	424	366	318
β^* estimate	-0.101	-0.105	-0.201	-0.067	-0.065	-0.071
st. error	0.052	0.061	0.097	0.061	0.057	0.092
2-sided p-value	0.054	0.083	0.039	0.275	0.255	0.437
1-sided p-value	0.027	0.042	0.020	0.138	0.128	0.218
observations	424	366	318	424	366	318

Note: Sample contains observations with German and/or French native language, most consistent choices, and non-missing values in the respective outcome, language dummies, and control variables. 'Risk': controlling for risk aversion. 'Controls': same controls as in column 2 of Table 3.4. Columns 1–2, 4–5 show OLS estimates for β and δ and average marginal probit effects for β^*. 'Lasso': OLS regression with lasso-selected controls. Standard errors are clustered at the class level

This allows us to eliminate students who deliberately attend school in the non-native language. Table 3.8 presents results for our three regression specifications when only native German speakers in German-speaking classes and native French speakers in French-speaking classes are compared. Here, the German language dummy therefore represents class language and native language at the same time.

The results again confirm the overall picture from the previous analyses. In column 2, which includes our vector of socioeconomic controls, we find a significant effect of the German language

Table 3.8: Effects of class and native languages

	Risk (1)	Controls (2)	Lasso (3)
δ estimate	0.022	0.037	0.022
st. error	0.022	0.017	0.026
2-sided p-value	0.334	0.028	0.381
1-sided p-value	0.167	0.014	0.191
observations	394	339	294
β estimate	0.020	0.013	0.016
st. error	0.018	0.010	0.024
2-sided p-value	0.255	0.210	0.504
1-sided p-value	0.127	0.105	0.252
observations	386	333	290
β^* estimate	-0.119	-0.106	-0.152
st. error	0.056	0.072	0.097
2-sided p-value	0.033	0.144	0.116
1-sided p-value	0.016	0.072	0.058
observations	386	333	290

Note: Sample contains observations with German native language in German classes and French native language in French classes, most consistent choices, and non-missing values in the respective outcome, language dummy, and control variables. 'Risk': controlling for risk aversion. 'Controls': same controls as in column 2 of Table 3.4. Columns 1–2 show OLS estimates for β and δ and average marginal probit effects for β^*. 'Lasso': OLS regression with lasso-selected controls. Standard errors are clustered at the class level

on δ, a moderate but insignificant effect on β, and a large and marginally significant effect on β^*.[32]

Taken together, when imposing a quasi-hyperbolic discounting model and inferring preference parameters at the individual level, we find systematic differences in long-run discounting between the two language groups. Moreover, consistent with the finding that differences in observed discounting were more pronounced in the short run, we find evidence for the existence of more present-biased individuals in the francophone subsample, as well as support for stronger present bias on average among the French-speaking students.

3.5 Conclusion

In this paper, we study whether the encoding of time in a language is correlated with differences in revealed time preferences across language groups. In particular, we refine the original hypothesis by Chen (2013), who suggested that languages that grammatically associate the future and the present, i.e. languages with a so-called weak future-time reference (w-FTR; Thieroff, 2000), lead w-FTR speakers to perceive future events as less distant, and hence to discount less. Since the grammatical association between the present and the future only makes a grammatical distinction when actually comparing the present and the future, it seems plausible that such linguistic differences also only translate into behavioral differences when trade-offs involve the present and the future. If this is the case, the encoding of time in a language should primarily affect the degree of

[32]The right panel of Appendix Table 3.A8 replicates the analyses presented here, but only including participants with consistent choices in the elicitation of time preferences. Again, results remain qualitatively unchanged.

present bias, rather than patience universally.

Our results show that French speakers in general discount more strongly than German speakers. However, the effect is more pronounced and statistically significant when immediate payments are involved, i.e. for trade-offs between present and future payments. When only future payments are involved, differences become significantly less pronounced. Consistent with these findings, when estimating structural preference parameters for a quasi-hyperbolic discounting model, we find that French speakers have on average significantly smaller δ, are significantly more likely to display present-biasedness and are on average more present biased. Our data therefore suggest that language indeed affects future orientation, and the effects are particularly pronounced in trade-offs that involve immediate rewards, reflecting a stronger present bias.

Our results inform observed behavioral differences between language groups. For example, Eugster et al. (2017) document that, at the Swiss language border, French speakers have significantly longer unemployment spells than German speakers, despite access to the same labor market. DellaVigna and Paserman (2005) show theoretically that stronger present bias leads to extended unemployment spells. Since our data reveal stronger present-biasedness among French speakers at the language border, language might be the microfoundation that can explain the differences observed in Eugster et al. (2017), mediated through present-biasedness. Similarly, Erhardt and Haenni (2018) show differences in the propensity for entrepreneurship based on ancestry from the German or French-speaking side of the language border in Switzerland. Again, our data provide a potential explanation for these observations, based on differences in patience and present bias.

One potential issue in the analysis of the effects of language on behavior is that language can be seen — and is often used — as a proxy for culture. Indeed, Roberts et al. (2015) find that the effect of language on behavior is weaker, but does not disappear, when controlling for the relatedness of languages and culture. In the same vein, one could argue that religion, which might be also seen as a component of culture, could be a potential driver of behavioral differences between the two language groups. At the language border under consideration, the French-speaking population has traditionally been predominantly Catholic, whereas the German-speaking population has predominantly been Protestant.

While we acknowledge these difficulties as inherent to the task of assessing the pure effect of language on time preferences across different language groups, we believe that our setting has unique advantages to test differences in preferences between language groups. We exploit natural language differences among students in a closely confined geographical area, in which both language groups have been sharing the same cultural and institutional environment for centuries. Moreover, given that lower secondary education is compulsory and residence dependent, there is no selection of students in terms of other socioeconomic characteristics.

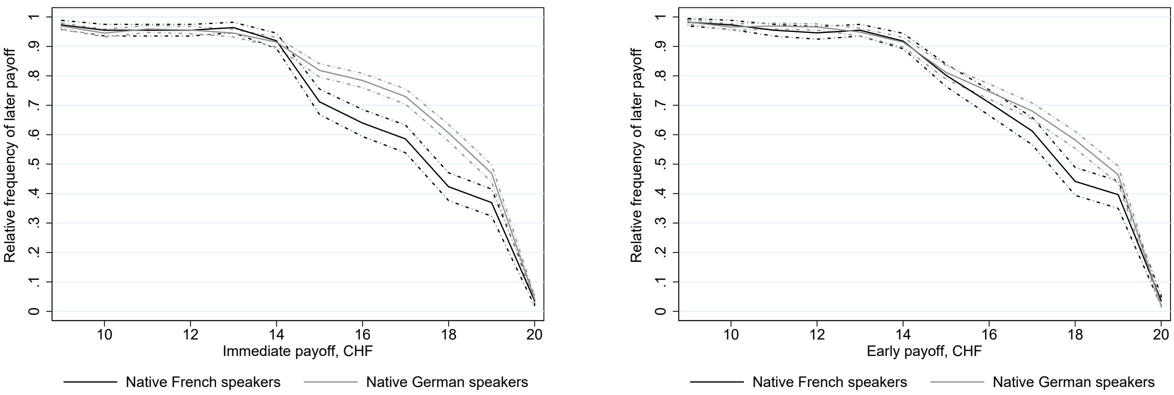
Moreover, religiosity plays little role for our study participants because only 7% of our sample indicate religiosity as something that is emphasized by their parents. Nonetheless, it has been argued since Weber (1930) that Protestantism may foster patience, and cross-country evidence suggests that Protestant countries are more patient (see, for example, Falk et al., 2018). However, recent work suggests that it is economic conditions, rather than religiosity, that foster patience over time (Doepke et al., 2005; Doepke and Zilibotti, 2008). Indeed, Cantoni (2015) finds no evidence for long lasting causal effects of Protestantism on economic development, also arguing that regional differences in economic prospects rather than religiosity are causing the emergence of differences in patience. Since economic prospects were constant for French and German-speaking inhabitants of the region, the latter argument cannot explain observed differences in time preferences.

Ultimately, language and culture remain inseparably intertwined. Even though we cannot ultimately rule out that our findings might be caused by some other difference in culture that we are unable to control for, our study provides the first clean evidence on differences in present-biasedness across w-FTR and s-FTR language groups, which is consistent with the hypothesis that they are caused by differences in language. Further empirical research is needed to continue improving our understanding of the origins of the apparent differences in preferences across language groups.

3.A Additional figures and tables

Figure 3.A1: Relative frequency of delayed payoff choices by native language

(a) Today vs. 4 weeks (b) 4 weeks vs. 8 weeks



Note: Each figure includes 403 observations. Dotted lines display one standard deviation of the mean.

Table 3.A1: Differences in the relative frequency of earlier payoff choices by school language

	Today vs. 4 weeks			4 weeks vs. 8 weeks		
Decision	DE-FR	st. error	p-value	DE-FR	st. error	p-value
9 vs. 20	-0.021	0.019	0.279	-0.010	0.015	0.526
10 vs. 20	-0.021	0.024	0.372	0.001	0.018	0.954
11 vs. 20	-0.033	0.023	0.155	-0.015	0.020	0.442
12 vs. 20	-0.050	0.025	0.051	-0.023	0.022	0.301
13 vs. 20	-0.034	0.025	0.171	-0.022	0.025	0.374
14 vs. 20	-0.037	0.031	0.226	-0.020	0.031	0.521
15 vs. 20	-0.151	0.044	0.001	-0.058	0.043	0.174
16 vs. 20	-0.190	0.046	0.000	-0.087	0.046	0.060
17 vs. 20	-0.184	0.048	0.000	-0.104	0.048	0.031
18 vs. 20	-0.192	0.048	0.000	-0.152	0.049	0.002
19 vs. 20	-0.123	0.047	0.010	-0.095	0.048	0.047
20 vs. 20	-0.013	0.019	0.501	0.007	0.015	0.667

Note: OLS estimates are based on individual choices from the first price list ('Today vs. 4 weeks') and the second price list ('4 weeks vs. 8 weeks'). Columns 'DE-FR' display estimated coefficients for a German dummy, showing differences in the relative frequency of earlier payoff choices between German and French classes. Standard errors are heteroscedasticity robust without clustering. Sample includes 496 observations.

Table 3.A2: Tobit regressions on switch points: full estimation results

	None			Risk			All controls		
	coef	st. error	p-value	coef	st. error	p-value	coef	st. error	p-value
German	1.688	0.439	0.000	1.642	0.439	0.000	1.983	0.491	0.000
4 weeks	0.370	0.298	0.214	0.390	0.297	0.190	0.636	0.322	0.048
German x 4 weeks	-0.628	0.353	0.076	-0.667	0.353	0.060	-0.804	0.385	0.037
Risk aversion				0.138	0.113	0.222	-0.006	0.132	0.963
Age							-0.686	0.444	0.123
Female							0.106	0.418	0.799
Swiss national							1.113	0.563	0.048
Number of siblings							-0.018	0.212	0.931
Mother's age							-0.020	0.052	0.702
Father's age							0.017	0.047	0.717
Lives in owned house							1.002	0.724	0.167
Lives in rented house							1.450	0.943	0.125
Lives in rented flat							0.213	0.787	0.787
Home ownership unknown							-0.469	0.995	0.637
Has own room							-1.047	0.805	0.194
Weeks on holiday previous year							-0.240	0.138	0.082
Difficulty to raise CHF 100							0.297	0.182	0.104
Pocket money per week							-0.005	0.008	0.552
Parents encourage independence							-0.482	0.583	0.408
Parents encourage responsible behavior							0.944	0.442	0.033
Parents encourage hard work							0.126	0.459	0.784
Parents encourage unselfishness							0.334	0.444	0.452
Parents encourage religiosity							1.395	0.768	0.070
Trust							-0.751	0.404	0.063
Quiz correct							0.268	0.595	0.652
Class 10							0.239	0.668	0.720
Class 11							1.474	1.006	0.143
Track B							-1.492	0.467	0.001
Track C							-2.047	0.586	0.001
Track E							-3.987	1.976	0.044
Constant	17.922	0.369	0.000	17.191	0.680	0.000	26.966	6.737	0.000
Observations	977			969			824		

Note: Two-limit tobit regressions on individual most consistent switch points, identical to the specifications reported in Table 3.2. Standard errors are clustered at the individual level.

Table 3.A3: Tobit regressions on switch points

	None	Risk	All controls
	(1)	(2)	(3)
German	1.688	1.642	1.983
st. error	0.986	0.972	0.739
p-value	0.087	0.091	0.007
4 weeks	0.370	0.390	0.636
st. error	0.459	0.454	0.435
p-value	0.420	0.390	0.144
German x 4 weeks	-0.628	-0.667	-0.804
st. error	0.495	0.490	0.484
p-value	0.205	0.174	0.097
Constant	17.922	17.191	26.966
st. error	0.924	1.227	7.585
p-value	0.000	0.000	0.000
Observations	977	969	824

Note: Two-limit tobit regressions on individual most consistent switch points. Standard errors are clustered at the class level.

Table 3.A4: Mean values of time preference parameter estimates by class language: consistent choices only

	French	German	DE-FR	p-value robust	p-value clustered	Observations
δ	0.870	0.901	0.031	0.024	0.246	482
	(0.141)	(0.128)	(0.027)	0.012	0.123	
β	0.996	1.020	0.024	0.125	0.199	467
	(0.163)	(0.131)	(0.019)	0.062	0.099	
β^*	0.310	0.200	-0.110	0.015	0.044	467
	(0.464)	(0.401)	(0.054)	0.007	0.022	

Note: 'DE' and 'FR' stand for German and French, respectively. 'p-value' denotes the significance level of mean difference t-tests of German vs. French ('DE-FR'). The first (second) row of each outcome contains two-sided (one-sided) p-values that are heteroscedasticity robust without clustering and clustered at the class level, respectively. Values in parentheses are standard deviations of outcome values for language groups and standard errors of mean differences.

Table 3.A5: Mean values of time preference parameter estimates by native language: consistent choices only

	French	German	Bilingual	DE-FR	p-value robust	p-value clustered	bi-FR	p-value robust	p-value clustered	Observations
δ	0.886	0.904	0.900	0.018	0.204	0.301	0.014	0.550	0.532	437
	(0.127)	(0.124)	(0.125)	(0.017)	0.102	0.150	(0.023)	0.275	0.266	
β	1.002	1.016	1.004	0.015	0.424	0.457	0.002	0.940	0.942	428
	(0.166)	(0.145)	(0.103)	(0.020)	0.212	0.228	(0.025)	0.470	0.471	
β^*	0.312	0.198	0.222	-0.114	0.025	0.046	-0.090	0.287	0.255	428
	(0.465)	(0.399)	(0.422)	(0.057)	0.012	0.023	(0.079)	0.143	0.127	

Note: 'DE', 'FR' and 'bi' stand for German, French and bilingual, respectively. 'p-value' denotes the significance level of mean difference t-tests of German vs. French ('DE-FR') and bilingual vs. French ('bi-FR'). The first (second) row of each outcome contains two-sided (one-sided) p-values that are heteroscedasticity robust without clustering and clustered at the class level, respectively. Values in parentheses are standard deviations of outcome values for language groups and standard errors of mean differences.

Table 3.A6: Mean values of all covariates by class language

	FR	DE	DE-FR	p-value
Age (years)	14.532	14.495	-0.037	0.919
Female	0.470	0.542	0.072	0.337
Born in Switzerland	0.879	0.928	0.049	0.179
Years in Switzerland	13.593	14.017	0.424	0.171
Swiss national	0.752	0.901	0.150	0.011
Number of siblings	1.497	1.464	-0.033	0.712
Number of elder siblings	0.792	0.776	-0.016	0.807
Attended preschool	0.819	0.988	0.170	0.000
Years in preschool	1.461	1.836	0.375	0.000
Attended Swiss preschool	0.764	0.944	0.181	0.000
Mother's age	44.619	45.386	0.767	0.245
Father's age	47.514	48.168	0.655	0.180
Mother's native language: German	0.235	0.749	0.514	0.000
Mother's native language: French	0.530	0.150	-0.380	0.000
Mother's native language: Italian	0.067	0.032	-0.035	0.134
Father's native language: German	0.329	0.744	0.415	0.000
Father's native language: French	0.456	0.108	-0.349	0.000
Father's native language: Italian	0.067	0.041	-0.026	0.251
Lives with both parents	0.812	0.741	-0.071	0.031
Lives with mother only	0.121	0.199	0.078	0.005
Lives with father only	0.034	0.017	-0.016	0.269
Lives in owned house	0.564	0.697	0.134	0.020
Lives in rented house	0.094	0.078	-0.016	0.570
Lives in rented flat	0.275	0.187	-0.088	0.057
Home ownership unknown	0.027	0.023	-0.004	0.820
Has own room	0.905	0.928	0.023	0.273
Weeks on holiday previous year	3.275	3.482	0.207	0.317
Difficulty to raise CHF 100	3.224	3.111	-0.113	0.412
Pocket money per week	17.495	16.722	-0.773	0.816
Receives pocket money: less than once a month	0.047	0.052	0.005	0.824
Receives pocket money: once a month	0.376	0.529	0.153	0.007
Receives pocket money: once a week	0.302	0.195	-0.107	0.068
Receives pocket money: more than once a week	0.067	0.026	-0.041	0.027
Saves part of pocket money	0.738	0.755	0.017	0.782
Savings per week	8.442	9.286	0.844	0.666
Attended course on money use	0.109	0.078	-0.031	0.483
Spends hours on homework	1.154	1.245	0.091	0.532
Parents encourage independence	0.743	0.833	0.090	0.034
Parents encourage responsible behavior	0.804	0.628	-0.176	0.000
Parents encourage imagination	0.419	0.303	-0.116	0.027
Parents encourage respect to others	0.730	0.597	-0.133	0.013
Parents encourage thrift	0.331	0.458	0.127	0.014
Parents encourage perseverance	0.547	0.340	-0.207	0.000
Parents encourage hard work	0.453	0.216	-0.237	0.001
Parents encourage religiosity	0.101	0.052	-0.049	0.010
Parents encourage unselfishness	0.277	0.432	0.155	0.023
Parents encourage obedience	0.595	0.354	-0.240	0.000
Parents encourage self-expression	0.297	0.337	0.040	0.496
Important quality for oneself: patience	1.859	1.818	-0.041	0.460
Important quality for oneself: risk-taking	2.336	2.334	-0.001	0.985
Important quality for oneself: thrift	1.918	1.725	-0.194	0.005
Important quality for oneself: helpfulness	1.732	1.278	-0.453	0.000
Important quality for oneself: future planning	1.784	1.892	0.108	0.191
Important quality for oneself: fairness/equality	1.723	1.353	-0.370	0.000
Important quality for oneself: openness/tolerance	1.698	1.626	-0.072	0.485
Important quality for society: patience	1.711	1.794	0.082	0.347
Important quality for society: risk-taking	2.432	2.437	0.005	0.957
Important quality for society: thrift	1.748	1.729	-0.019	0.799
Important quality for society: helpfulness	1.541	1.246	-0.295	0.000
Important quality for society: future planning	1.757	1.888	0.131	0.144
Important quality for society: fairness/equality	1.534	1.251	-0.283	0.000
Important quality for society: openness/tolerance	1.497	1.609	0.113	0.098

	FR	DE	DE-FR	p-value
Trust	0.434	0.533	0.099	0.048
Risk aversion	5.041	5.485	0.445	0.052
Quiz correct	0.899	0.893	-0.006	0.847
Class 9	0.356	0.354	-0.001	0.995
Class 10	0.349	0.354	0.005	0.979
Class 11	0.295	0.291	-0.004	0.983
Track A	0.450	0.392	-0.058	0.794
Track B	0.463	0.380	-0.083	0.704
Track C	0.087	0.202	0.114	0.288
Track E	0.000	0.026	0.026	0.200
Session 1	0.369	0.366	-0.003	0.988
Session 2	0.403	0.314	-0.089	0.674
Session 3	0.228	0.320	0.092	0.607
Born in Jan	0.087	0.104	0.017	0.504
Born in Feb	0.074	0.072	-0.002	0.954
Born in Mar	0.054	0.107	0.053	0.068
Born in Apr	0.060	0.104	0.044	0.088
Born in May	0.081	0.078	-0.003	0.886
Born in Jun	0.121	0.078	-0.043	0.172
Born in Jul	0.067	0.101	0.034	0.199
Born in Aug	0.101	0.072	-0.028	0.369
Born in Sep	0.054	0.075	0.021	0.408
Born in Oct	0.121	0.058	-0.063	0.044
Born in Nov	0.114	0.084	-0.030	0.245
Born in Dec	0.067	0.066	-0.001	0.982

Note: 'FR' and 'DE' stand for French and German, respectively. 'p-value' stands for two-sided p-values of mean difference t-tests of German vs. French ('DE-FR') and accounts for clustering at the class level. 'Difficulty to raise CHF 100' is measured on a 5-point scale: 1=very difficult, 5=very easy; variables 'important quality for ...' are measured on a 4-point scale: 1=very important, 4=unimportant.

Table 3.A7: Effects of class language: specification 2 detailed results

	δ estimate			β estimate			β^* estimate		
	OLS	st. error	p-value	OLS	st. error	p-value	Probit	st. error	p-value
German class language	0.038	0.018	0.034	0.020	0.012	0.099	-0.101	0.063	0.110
Risk aversion	-0.004	0.005	0.464	0.010	0.005	0.065	-0.011	0.015	0.471
Age	-0.020	0.015	0.194	-0.002	0.019	0.938	-0.002	0.040	0.958
Female	0.003	0.016	0.861	0.003	0.014	0.804	-0.008	0.053	0.886
Swiss national	0.059	0.023	0.011	-0.043	0.026	0.108	0.082	0.058	0.157
Number of siblings	-0.003	0.006	0.669	0.003	0.008	0.654	0.001	0.025	0.967
Mother's age	-0.000	0.002	0.797	0.000	0.002	0.890	-0.004	0.007	0.577
Father's age	0.001	0.002	0.727	-0.000	0.002	0.822	0.001	0.007	0.936
Lives in owned house	0.022	0.032	0.488	0.014	0.059	0.813	-0.105	0.079	0.180
Lives in rented house	0.037	0.037	0.313	0.011	0.055	0.837	-0.070	0.074	0.342
Lives in rented flat	0.005	0.037	0.893	0.017	0.062	0.788	-0.036	0.071	0.611
Home ownership unknown	0.013	0.052	0.800	0.003	0.079	0.968	-0.061	0.164	0.709
Has own room	-0.030	0.025	0.220	0.008	0.028	0.783	0.104	0.057	0.066
Weeks on holiday previous year	-0.005	0.004	0.244	-0.001	0.007	0.881	0.002	0.014	0.874
Difficulty to raise CHF 100	0.008	0.005	0.074	0.001	0.007	0.927	0.005	0.020	0.817
Pocket money per week	0.000	0.000	0.850	-0.000	0.000	0.233	0.001	0.001	0.095
Parents encourage independence	-0.020	0.018	0.260	0.015	0.022	0.497	-0.052	0.064	0.416
Parents encourage responsible behavior	0.033	0.010	0.001	-0.014	0.015	0.374	0.027	0.046	0.555
Parents encourage hard work	-0.005	0.016	0.774	0.012	0.015	0.399	-0.081	0.045	0.072
Parents encourage unselfishness	-0.002	0.012	0.894	0.010	0.013	0.468	-0.024	0.047	0.613
Parents encourage religiosity	0.034	0.023	0.128	-0.013	0.020	0.524	-0.017	0.064	0.789
Trust	-0.028	0.013	0.030	0.036	0.013	0.005	-0.049	0.047	0.302
Quiz correct	0.016	0.019	0.403	-0.019	0.037	0.598	-0.030	0.053	0.568
Class 10	0.012	0.022	0.593	-0.023	0.022	0.308	0.061	0.048	0.199
Class 11	0.055	0.036	0.125	-0.029	0.043	0.493	0.120	0.114	0.290
Track B	-0.039	0.016	0.018	-0.010	0.011	0.350	0.010	0.058	0.859
Track C	-0.077	0.025	0.002	0.064	0.024	0.008	-0.059	0.063	0.350
Track E	-0.123	0.074	0.095	-0.028	0.066	0.674	0.032	0.125	0.798
Constant	1.115	0.230	0.000	0.994	0.322	0.002			
Observations	415			407			407		

Note: Sample contains observations with most consistent choices and non-missing values in the respective outcome, language dummy, and control variables. 'OLS': OLS coefficients. 'Probit': average marginal probit effects. Standard errors are clustered at the class level, p-values correspond to two-sided hypothesis tests.

Table 3.A8: Effects of class and native languages: consistent choices only

	Class language			Class and native languages		
	Risk	Controls	Lasso	Risk	Controls	Lasso
	(1)	(2)	(3)	(4)	(5)	(6)
δ estimate	0.029	0.036	0.032	0.021	0.037	0.026
st. error	0.025	0.018	0.030	0.022	0.017	0.025
2-sided p-value	0.248	0.042	0.288	0.333	0.036	0.291
1-sided p-value	0.124	0.021	0.144	0.166	0.017	0.145
observations	472	403	351	380	329	287
β estimate	0.019	0.021	0.048	0.014	0.007	0.013
st. error	0.017	0.013	0.025	0.019	0.012	0.023
2-sided p-value	0.254	0.100	0.050	0.445	0.560	0.569
1-sided p-value	0.127	0.050	0.025	0.222	0.280	0.285
observations	459	392	342	372	322	282
β^* estimate	-0.113	-0.110	-0.154	-0.126	-0.106	-0.147
st. error	0.052	0.065	0.065	0.058	0.074	0.097
2-sided p-value	0.031	0.092	0.018	0.029	0.154	0.132
1-sided p-value	0.015	0.046	0.009	0.015	0.077	0.066
observations	459	392	342	372	322	282

Note: Both panels contain observations with most consistent choices and non-missing values in the respective outcome, language dummy, and control variables. The right panel only includes native German speakers in German classes and native French speakers in French classes. 'Risk': controlling for risk aversion. 'Controls': same controls as in column 2 of Table 3.4. Columns 1–2, 4–5 show OLS estimates for β and δ and average marginal probit effects for β^*. 'Lasso': OLS regression with lasso-selected controls. Standard errors are clustered at the class level.

4 Intertemporal choice under social comparison: A real-effort experiment[1]

Overview

Does social comparison affect intertemporal choice behavior? We answer this question by recruiting Amazon Mechanical Turk workers for three weeks to perform a real-effort task. In a randomized experiment, we examine subjects' effort allocations between two work dates in response to varying information about effort reallocations by previous participants. We find that social comparisons affect men and women differentially. Observing that 48% of peers made time-consistent choices causes men to behave time-consistently but has an opposite effect on women. On average, men also exhibit significantly smaller estimates of the present-bias parameter of a quasi-hyperbolic discounting model. Observing peer procrastination induces women to behave time-consistently but does not affect men. Our findings suggest that social comparison based on situational similarity affects intertemporal choice, and that gender-tailored social comparison can be an effective solution to dynamic inconsistency even when individual time preferences are not known beforehand.

4.1 Introduction

Intertemporal choices — trading off earlier and delayed outcomes — are inherent in private and professional contexts. For instance, going to the gym, paying bills, searching a job, preparing a report or providing a feedback often involve decisions between doing it now or later. Theory and evidence on intertemporal choice suggest that people are prone to procrastinate, i.e. postpone effortful activities that they wish they would do sooner (e.g. O'Donoghue and Rabin, 1999, 2001; Augenblick et al., 2015). Procrastination can be very costly in the sense that individuals may procrastinate more heavily on important goals than unimportant ones (O'Donoghue and Rabin, 2001). Such self-control problems may lead procrastinators to experience inferior performance, more stress and illness in the long-run (e.g. Tice and Baumeister, 1997). It is therefore important to understand which incentives can reduce procrastination and induce time-consistent behavior.

Individuals often use social groups to overcome self-control problems in achieving personal goals (e.g. study groups). Social influence is particularly relevant at work, where people spend a large part of their lives interacting with colleagues. Social interactions within groups usually

[1]This paper presents the results of an interim experiment. The final experiment was not conducted due to the lack of financial support. The experiment was designed and run in collaboration with Holger Herz. Maria Senkiv's help with JavaScript in online experiments is kindly acknowledged. This research was funded by the Department of Economics at the University of Fribourg.

involve observing each other's behavior. A few recent studies experimentally examine the impact of observing others and/or being observed by others on individual work performance (e.g. Georganas et al., 2015; Gerhards and Gravert, 2016; Buechel et al., 2018; van Veldhuizen et al., 2018; Beugnot et al., 2019), but none of them explores peer effects on intertemporal choice behavior. Our study addresses this issue by examining whether observing others' behavior affects one's own intertemporal choice, and if so, how.

We find a theoretical foundation in Battaglini et al. (2005), who postulate that observing peers' behavior can be informative about one's own ability to handle temptation if there is enough similarity between one's own and peers' willpower. Their model predicts that individual reaction to peers' successes and failures in managing self-control is non-monotonic. Social influence depends on how people perceive their own vs. peers' self-control problems. The ideal peers are those who have slightly worse self-control problems than oneself because this makes their successes more encouraging and their failures less discouraging. Battaglini et al. (2017) indeed find that students with more social ties have better self-control, and that students' self-control is higher when an average self-control of their friends is slightly lower. These results, however, rely on observational data of social contacts that involve both observing others and being observed by others, without separating one mechanism from another.

Clean evidence whether social comparison affects intertemporal choice is actually limited.[2] The only study we know about is the field experiment of Beshears et al. (2015), who find that information about coworkers' behavior on savings leads to oppositional reactions, i.e. reduced savings. The authors argue that discouragement from upward social comparison among low-income individuals potentially drives this result.[3] Our paper contributes to this literature by providing the first evidence on the effect of social comparison on intertemporal choice over effort, which is closely related to consumption.

We conduct a longitudinal real-effort experiment in the spirit of Augenblick et al. (2015). Participants are recruited for three consecutive weeks on Amazon Mechanical Turk (MTurk), an online marketplace. Subjects perform a task for a fixed monetary reward. The task involves counting zeros in tables of randomly distributed zeros and ones (Abeler et al., 2011). In weeks 1 and 2, we ask subjects to allocate tables over two work dates that occur in weeks 2 and 3, respectively. While an initial allocation involves a trade-off between two future dates, a subsequent allocation involves a trade-off between the present and the future. For each subject, we then randomly choose one of the two allocations to implement. Differences between initial and subsequent allocations allow us to analyze subjects' intertemporal choice behavior.

[2] In a field experiment, Kast et al. (2018) find that feedback text messages with peer information significantly increase deposits among microcredit clients in Chili, but the effect of peer information is inseparable by design. While Gerhards and Gravert (2016) and Buechel et al. (2018) investigate the effect of observing peers on perseverance in laboratory experiments, their experimental designs do not allow measuring procrastination.

[3] Upward social comparison means comparing oneself with others who are superior or more fortunate (see Wood and Taylor, 1991).

Before making subsequent allocations, subjects are randomly assigned to one of three conditions with varying information about reallocation decisions of past participants. In an *implicit shift* group, subjects receive information that 48% of past participants made time-consistent choices. In an *explicit shift* group, subjects are additionally informed about the average number of tables procrastinating peers shifted to the final work date. No peer information is provided to a *control* group. We compare subjects' behavior across the conditions to investigate social comparison effects on intertemporal choice. In particular, if subjects in social comparison groups behave differently than subjects in the control group, social comparison based on situational similarity affects intertemporal choice in effort. We also compare the behavior of subjects in the explicit and implicit groups to evaluate whether knowing about peer procrastination impacts dynamic in/consistency.

Guided by Battaglini et al. (2005), we expect that the implicit shift treatment would encourage and the explicit shift treatment would not discourage time-consistent behavior if subjects are sufficiently confident in their own vs. peers' willpower. Our behavioral measures allow assessing the theoretical predictions both at the extensive and intensive margins, i.e. whether subjects behave time-consistently or procrastinate, and if so, to which extent.

We find that social comparisons affect men and women differentially, leading to mostly imprecise effects in the overall sample. Observing that 48% of past participants made time-consistent choices, men are by 18 percentage points significantly more likely to behave time-consistently, but women are by 32 percentage points significantly less likely to do so. Assuming quasi-hyperbolic discounting (Laibson, 1997; O'Donoghue and Rabin, 1999), we also find that estimates of men's present-bias parameters are on average significantly smaller in the implicit shift vs. the control condition.

If men are sufficiently self-confident, their behavior in the implicit shift condition is consistent with the theoretical predictions of Battaglini et al. (2005). For individuals lacking self-confidence, Battaglini et al. (2005) predict procrastination. While the fraction of women behaving time-inconsistently significantly increases in the implicit shift condition, this behavior is largely driven by future-biased choices, where more tables are subsequently allocated to week 2 than initially planned. In line with Lenney (1977) and Falk and Knell (2004), we expect that women could increase their effort allocations in week 2 if they engaged in upward comparison — for the purpose of self-improvement — when receiving implicit information about peers' dynamic inconsistency.

We also find that observing peer procrastination significantly increases by 40 percentage points the fraction of women behaving time-consistently but has no statistically significant impact on men. These results are consistent with Battaglini et al. (2005) for sufficiently self-confident women and possibly even more self-confident men.[4] Likewise, previous studies document no evidence that observing poorly performing peers significantly reduces individual effort (e.g. Mas and Moretti,

[4]Previous studies show that despite similar abilities men are significantly more (over)confident than women in their relative performance (e.g. Niederle and Vesterlund, 2007; Mobius et al., 2011).

2009; Buechel et al., 2018).[5]

Our results suggest that women are more sensitive to the contents of peer information than men. Buechel et al. (2018) similarly report that women respond more strongly than men to successful peers communicating in an encouraging (discouraging) way.[6] In their study, however, peer effects on perseverance emerge only in the presence of peer communication. It is also worth mentioning that with our experimental design we cannot verify whether the mere reference to peers — a kind of social facilitation (see Zajonc, 1965) — could contribute to men's time-consistent behavior in our social comparison conditions

Our findings have two important implications. First, social comparison based on situational similarity affects intertemporal choice in real effort. Second, gender-tailored social comparison can provide an effective solution to dynamic inconsistency even when individual time preferences are not known beforehand.

In the domain of intertemporal choice our paper is related to empirical research tackling dynamic inconsistency in effort (e.g. Ariely and Wertenbroch, 2002; Kaur et al., 2010; Augenblick et al., 2015). These studies focus on designing self-control mechanisms, such as commitment devices. However, the choice of optimal commitment devices is challenging because people are consistently at least partially unaware of their self-control problems (e.g. Augenblick and Rabin, 2019). In addition, Augenblick et al. (2015) show that even people potentially aware of their dynamic inconsistency are willing to pay little for commitment to constraint their future selves. Our contribution is that we introduce social comparison at price zero instead of costly commitment devices to encourage subjects' time-consistent behavior.

In the domain of social comparison closely relevant for our work is the study by DellaVigna and Pope (2018). They find that observing high performance of many past participants significantly increases effort provision of MTurk workers. However, their social comparison experiment does not involve intertemporal trade-offs. Our paper also contributes to a vast research on social comparison, albeit in different context. For instance, recent evidence shows that social comparison can enhance charity donations (Frey and Meier, 2004; Krupka and Weber, 2009), pro-environmental actions such as towel reuse (Goldstein et al., 2008) or energy consumption (Schultz et al., 2007), payment of overdue taxes (Hallsworth et al., 2017) and take-up of modestly paid teaching jobs (Coffman et al., 2017).

Finally, we add to studies that use MTurk for behavioral research in economics (e.g. Kuziemko et al., 2015; DellaVigna and Pope, 2018; de Quidt et al., 2018). Unlike Augenblick et al. (2015), who document significant present bias among students in a real-effort experiment, we find no

[5]Gerhards and Gravert (2016) find that observing a similar or less able peer switch from hard to easy tasks significantly positively affects the observer's propensity to switch as well, but the two types of peers are pooled into one category in regression analysis.

[6]Unlike us, Buechel et al. (2018) expose subjects to peer information without allowing them to first practice with a task.

evidence that MTurk workers procrastinate or display present bias on average. Our finding is consistent with DellaVigna and Pope (2018), who elicit MTurk workers' intertemporal choices by using delayed payments for effort.

The paper proceeds as follows. Section 4.2 describes the work environment. Section 4.3 outlines the experimental design. Section 4.4 presents our behavioral predictions and measures. Section 4.5 reports the results. Section 4.6 discusses our findings and design limitations. Section 4.7 concludes.

4.2 Work environment

We conducted a study using Amazon Mechanical Turk (MTurk), an online labor marketplace where tasks, the so-called Human Intelligence Tasks (HITs), are posted by requesters and performed by workers in exchange for a monetary reward. Tasks are usually simple and require a few minutes of manual work. Requesters are employers who create HITs and advertise them on the MTurk platform. As a requester,[7] we specified the description of the task containing instructions and an external task link,[8] the reward, the number of assignments to be done by unique workers,[9] the time limit per assignment, the expiration and approval time, as well as the required worker qualifications (i.e. the US residence to ensure language comprehension and no previous participation in a similar task in any of our previous HITs to prevent any spillovers).[10]

Once the HIT is posted, workers with the required qualifications can view and accept it. Many HITs are posted on MTurk simultaneously. Workers can search for tasks they are interested in and then select into tasks voluntarily. For task clarification, workers and requesters can communicate in writing on MTurk. In our study, subjects had to provide their MTurk worker ID and to enter the validation code displayed upon completion of the HIT to receive payment. Requesters may also award bonuses to good workers and punish poor performers by refusing payment. The latter negatively affects workers' track records on MTurk. From their side, workers can filter out unfair requesters and share their experience in public online forums.

MTurk has attracted our attention for several reasons. First, we receive access to a large population on MTurk at low cost per subject.[11] Second, the MTurk infrastructure streamlines the process of subject recruitment, data collection and subject compensation compared to traditional physical laboratories. Third, previous literature has documented that MTurk samples are at least

[7]The Chair of Industrial and Behavioral Economics at the University of Fribourg was a requester in this study.

[8]The task was designed in the online platform Qualtrics. Thus, data collection took place outside MTurk. Kuziemko et al. (2015) also used Qualtrics to survey MTurk workers.

[9]Though registered workers have unique IDs on MTurk, some subjects may have multiple accounts. By tracking down IP addresses, we could detect subjects who possibly performed the task several times or at least from the same location.

[10]Qualifications may also include, for instance, a minimum approval rate or a minimum number of previously approved HITs.

[11]In their MTurk studies, Horton and Chilton (2010) find a median reservation wage of USD 1.38 per hour, and Ipeirotis (2010) reports an effective hourly wage of USD 4.80.

as representative of the US population as traditional subject pools (e.g. Paolacci et al., 2010; Buhrmester et al., 2011; Berinsky et al., 2012), subject behavior on MTurk is consistent with findings from traditional laboratory experiments (e.g. Horton et al., 2011; Goodman et al., 2013), and low compensation rates do not appear to affect data quality (e.g. Mason and Watts, 2009; Buhrmester et al., 2011). Consequently, MTurk is increasingly used for behavioral research in economics (e.g. Kuziemko et al., 2015; DellaVigna and Pope, 2018; de Quidt et al., 2018). Finally, MTurk provides a natural setting for our real-effort experiment, which could classify as a field experiment as long as subjects are not aware of their participation in a scientific study (e.g. Horton et al., 2011).

4.3 Experiment

4.3.1 Intertemporal choice task

We conducted a longitudinal experiment that involved a real-effort task performed online over three consecutive weeks, along the lines of Augenblick et al. (2015). Our experiment consists of two stages: a HIT and a bonus task (see Section 4.2 for details on MTurk). Through the HIT, subjects get informed about the type of work and the bonus task. To be eligible for the bonus task, subjects need to successfully complete the HIT. That is, subjects first self-select into the HIT and then into the bonus task. The HIT involves only one work date. In the bonus task, subjects are required to initially allocate units of effort, subsequently allocate them again a week after, and execute one of the two allocations over two work dates. The comparison of initial and subsequent allocations allows identifying dynamically in/consistent choices.

The HIT is rewarded with USD 1.20 in week 1 and the bonus task with USD 10 in week 3.[12] The payments are performed via MTurk conditional on successful completion of the HIT and the bonus task, respectively. Note that our experimental design deviates from Augenblick et al. (2015) in that it accommodates the MTurk features and is simple to explain to subjects. In the following, we describe the effort task, the experimental timeline and the decision environment.

1. Effort task. In the spirit of Abeler et al. (2011), the effort task involves counting zeros in tables of 56 randomly distributed zeros and ones.[13] Per table, subjects have to enter the correct

[12]The offered payments are supposed to observe the federal minimum wage of USD 7.25 per hour in the US. See `www.dol.gov/whd/minimumwage.htm`, retrieved 15 May 2019.

[13]Bounded at 0.2–0.8, the probability of drawing zero varies across tables. This increases the variance in the number of zeros, which makes the effort task more challenging because guessing is harder. Furthermore, the table content is protected from being copied to encourage manual counting. On the one hand, we cannot rule out that any of our subjects used script-writing to automatize the completion of the bonus task. Since no restriction was imposed on the number of incorrect attempts, one could potentially write a script to sequentially enter all possible numbers in the answer box. On the other hand, we expect that script-writing takes more time than performing the effort task manually and thus appears cost inefficient. At least, passing a task comprehension quiz required a human to perform.

number of zeros to proceed with the task. A wrong answer produces an error message asking to count the same table again, with no penalization imposed. On average, it takes 23 seconds to count zeros in one table.[14] Though the effort task is meaningless, it resembles clerical jobs in being simple, repetitive and tiresome.

2. Experimental timeline. Throughout the three-week experiment, subjects perform the effort task on the same day of the week. That is, work dates are always seven days apart. In week 1, subjects complete the HIT and decide whether to participate in the bonus task. The latter requires to allocate tables over two work dates that occur in weeks 2 and 3, respectively. For weeks 2 and 3, subjects who opted for the bonus task receive an email invitation the day before with instructions and a link to the bonus task. On each work date, subjects have 24 hours to log in with their worker ID and complete the required work until midnight.[15] Upon logging in, subjects are first provided instructions about the work to be performed and decisions to be made that day, reminded about the timeline of the bonus task, and then asked to complete the necessary work. Table 4.1 presents the main elements of our experiment that occur in weeks 1–3, as described in detail below.

Table 4.1: Summary of longitudinal experiment

	HIT		Bonus task					
	Minimum work (fixed)	Receive reward	Minimum work (fixed)	Treatment	Allocate effort to weeks 2 & 3	One of two allocations chosen	Implement chosen allocation	Receive bonus
Week 1	✓	✓			✓			
Week 2			✓	✓	✓	✓	✓	
Week 3			✓				✓	✓

Note: 'HIT' stands for Human Intelligence Task. Bonus task comes after HIT completion.

3. Effort allocations. As part of the bonus task, subjects have to allocate 60 tables between weeks 2 and 3. Using a slider, subjects choose how many tables they want to count in week 2, and the remaining tables are consequently allocated to week 3. That is, every table allocated to week 2 reduces the number of tables allocated to week 3 by one. Each subject makes two allocation decisions: first in week 1 and then in week 2. Note that the allocation decision in week 1 involves two future work dates, whereas the decision in week 2 involves a present and a future work date. Prior to deciding in week 1, subjects are informed that they will decide again in week 2 and that only one of their allocation decisions from week 1 or week 2 will be implemented. In week 2, we do not remind subjects of their initial week 1 allocations.

4. Allocation-to-implement. As explained above, each subject makes two allocation decisions: first in week 1 and then in week 2. Just after the week 2 allocation, it is randomly chosen whether the allocation from week 1 or from week 2 will be actually implemented. In weeks 2 and 3, subjects have to complete the bonus task according to the implemented allocation in order to receive the

[14] The average is based on the one-shot trial that involved counting 20 tables. The median completion time of both the HIT and the bonus task was about 59 minutes (see Appendix Figure 4.A1).

[15] Each Qualtrics session is limited to two hours.

payment of USD 10 in week 3. From the outset, subjects are informed that the implemented allocation will come from week 1 with a 10% chance and from week 2 with a 90% chance.[16] Randomization is important to induce incentive compatibility for both allocation decisions.

5. Minimum work. Every week, subjects are required to count a fixed number of 20 tables before making allocation decisions or counting allocated tables. We inform subjects that the same minimum work has to be completed on each work date, independent of 60 tables that subjects freely allocate as explained above.[17] This minimum work serves several objectives. First, subjects get familiar with the task and thus gain a better understanding of how demanding it is before making allocation decisions. Second, subjects experience the same transaction costs of logging in and completing minimum work at all dates. Third, we expect subjects to take minimum work into account when making allocation decisions.

6. Quiz and demographics. To ensure that subjects understand the allocation decisions and the timeline of the study, they first practice with table allocations and then have to correctly complete a quiz in order to proceed with the bonus task in week 1. At the end of the first session of the bonus task, we collect subjects' demographic characteristics, such as gender, age and the highest level of education.[18] In addition, we can detect the location and IP address of each subject observed on the work dates.[19]

4.3.2 Treatments

To examine the impact of social comparison on intertemporal choice, we randomly assign subjects to either a control condition or one of two social comparison conditions that are equally possible. Randomization is done in week 1, but subjects experience their assigned condition after they complete the minimum work of 20 tables in week 2. In the *control* condition in week 2, subjects then

[16] The later decision receives greater probability weight in order to make our study and thus findings as comparable as possible to Augenblick et al. (2015), who used asymmetric probabilities to increase the chance that subjects experienced their own procrastination and to elicit demand for commitment.

[17] The week 1 minimum work of 20 tables is part of the HIT, whereas the bonus task covers another 100 tables. Note that earnings per table are lower in the HIT than in the bonus task in order to encourage subjects to accept the HIT with the goal of completing the bonus task.

[18] Exploring consistency in self-reported demographics on MTurk, Rand (2012) finds that the majority of subjects reported the same gender (96%), age (93%) and education level (81%) across different studies. Mason and Suri (2012) arrive at a similar conclusion. We could have also verified the consistency of the collected demographic data by repeating the demographic survey at the end of week 3. If anything, (unmeasured) false response rates in week 1 are independent of treatment in week 2.

[19] Among subjects, who completed the bonus task, there are four cases with duplicated IP addresses. Though these IP addresses appear public, we cannot rule out the possibility that one or two subjects have multiple MTurk accounts or subjects worked in pairs. Given that subjects with the same IP addresses received different treatments and completed work in week 2 just one after another, any communication between subjects likely violates the stable unit treatment value assumption (SUTVA; Rubin, 1974). Therefore, we exclude two subjects with duplicated IP addresses and later treatment exposure. Hence, our evaluation sample includes only the first observation from a given IP address as recommended by Berinsky et al. (2012).

simply allocate 60 tables between weeks 2 and 3. That is, without receiving any peer information, subjects in the control group decide how many tables they want to complete today and how many to complete in a week. In the *social comparison* conditions, subjects first receive varying information about the allocation decisions of previous participants and then make allocations themselves. Hereinafter we distinguish between the implicit and explicit shift treatments, both containing truthful information about past participants.

In the *implicit shift* condition, subjects get informed that other MTurk workers have already completed a similar bonus task and 48% of them decided to do as many tables "today" as originally planned. The *explicit shift* condition contains the same information and additionally discloses that past participants who shifted more tables to the final week on average intended to solve 23 tables less "today" than originally planned. Hence, the implicit shift treatment explicitly reveals information about the proportion of participants who behaved time-consistently, whereas the explicit shift treatment also mentions an average degree of intended procrastination. Both social comparison treatments contain partial information on the allocation decisions of past participants.[20] Appendix Figure 4.A4 displays the decision-making environments in the three conditions.

4.3.3 Procedures and data

On 5 July 2017, we posted the HIT with 200 assignments on MTurk that were completed by 200 workers within 4 hours 50 minutes (between 6:07 a.m. PDT and 10:57 a.m. PDT). Out of these 200 subjects, 161 agreed to participate in the bonus task, passed the quiz, made initial effort allocations and answered demographic questions.[21] From the outset, we informed subjects of the bonus task worth USD 10 that involved participation on 5, 12 and 19 July 2017 and explicitly asked subjects to opt for the bonus task only if they were able to do the bonus task on the required days. During the experiment, 17 subjects selected out in week 2 and two subjects in week 3. Consequently, 142 subjects completed the bonus task and received the bonus of USD 10 in the final week. Appendix Table 4.A1 displays no significant differences in terms of the initial allocations, discount factor δ

[20] We pooled information on 29 subjects who completed the pilot study between 19 June and 3 July 2017. Then, the effort task was exactly the same, but two social comparison treatments were based on the preceding pilot as follows: "Other MTURK Workers have already completed a similar Bonus Task. 45% of them decided to do as many tables today as originally planned" and "Other MTURK Workers have already completed a similar Bonus Task. 32% of them decided to shift more tables to the final week, intending to do on average 13 tables less today than originally planned". To refine the instructions, we conducted two longitudinal pilots in November–December 2016, without and with social comparison, respectively. Using a qualification flag, we excluded past participants from being recruited for further HITs.

[21] Initially, 181 subjects opted for the bonus task, but 17 did not accomplish the quiz. Furthermore, one subject did not provide an email address and selected out of the bonus task. Whenever an email address was incorrectly specified, we sent notifications for weeks 2 and 3 using the MTurk platform. In addition, we exclude two subjects with duplicated IP addresses and later treatment exposure from the evaluation sample (see footnote 19).

(explained in Section 4.4), gender, age or educational attainment by the dropout status. Note that the completion rate in this study is comparable to Augenblick et al. (2015). Our sample consists of 54 subjects in the control group, 48 subjects in the implicit shift group and 40 subjects in the explicit shift group.

In order to mitigate potential attrition in weeks 2 and 3, we sent out email reminders to subjects who had not performed their work until then. More precisely, we sent one reminder to 69 subjects in week 2, as well as one reminder to 22 subjects and two reminders to 22 subjects in week 3. Hereinafter, subjects are defined as delayers if they started working on the task after being reminded. Appendix Table 4.A2 shows no significant difference between delayers and non-delayers in week 2 in terms of pretreatment characteristics.[22] However, to take into account potential influence of the week 2 reminder on subsequent allocation decisions, we control for individual reminder exposure in regression analyses.[23]

Note that attrition in week 2 is independent of treatment condition because subjects dropped out without prior treatment exposure. On the contrary, attrition in week 3 might have been influenced by treatment but appears rather modest to pose serious concerns for our analyses.[24] In order to assess the quality of treatment randomization, we compare the control and social comparison groups by pretreatment characteristics conditional on finishing the bonus task. Appendix Table 4.A3 shows that subjects in the explicit shift group are on average 4–5 years older than subjects in the control or in the implicit shift group, respectively. Apart from that, we do not find any significant difference either in the initial allocations or in other demographic characteristics. Nevertheless, we take into account the observed heterogeneity of subjects by controlling for demographics in regression analyses.

4.4 Conceptual framework

4.4.1 Behavioral predictions

Classical economic theory assumes that individuals discount future events exponentially (Samuelson, 1937). The exponential discount function given by $D(t) = \delta^t$ implies time-consistent behavior because the same discount factor δ is used for every unit of time. Consequently, individuals make the same choices between sooner and later events no matter when they decide. But exponential discounting cannot explain time-inconsistent choices, such as procrastination (precrasti-

[22]With respect to the week 3 reminders, differences in pretreatment characteristics are mostly insignificant. Subjects who received the first reminder in week 3 are on average 4 years older than non-delayers. The share of females is about 15 percentage points larger among delayers than non-delayers. However, the observed differences are marginally significant at the 10% level. These results are available on request.

[23]The reminder in week 2 seems effective in the short-run (see Appendix Figure 4.A2).

[24]Only two subjects dropped out in week 3: a 35-year-old woman with a 4-year college degree, who completed 35 tables out of 100 in week 2, from the control group and a 28-year-old man with a high school degree, who completed 50 tables, from the explicit shift group.

nation), that give more (less) relative weight to immediate over future events.[25] In other words, individual discount rates are not necessarily constant over time (see Frederick et al., 2002).

To allow for different behavioral patterns, we use a quasi-hyperbolic discounting model (Phelps and Pollak, 1968; Laibson, 1997; O'Donoghue and Rabin, 1999) that describes intertemporal choice with a discount factor δ and a present-bias parameter β as follows:[26]

$$D(t) = \begin{cases} 1 & \text{if } t = 0 \\ \beta\delta^t & \text{if } t \geq 1 \end{cases} \tag{4.1}$$

where individuals with $\beta = 1$ behave time-consistently, those with $\beta < 1$ show present bias or procrastinate, and those with $\beta > 1$ show future bias or precrastinate.

While we expect subjects to display heterogeneous intertemporal choices, we cannot know the exact distribution of individual time preferences due to potential preference shocks (e.g. fatigue, other priorities) that subjects may face on the work dates. Furthermore, evidence suggests that people themselves are consistently at least partially unaware of their self-control problems (e.g. Augenblick and Rabin, 2019). Consequently, it appears difficult for people to choose optimal self-control mechanisms to avoid procrastinating (e.g. Ariely and Wertenbroch, 2002; DellaVigna and Malmendier, 2006).

The question is whether observing others' intertemporal choice behavior can improve willpower even without perfect knowledge of individual time preferences. Battaglini et al. (2005) proposed a theoretical model where people can learn more about their ability to resist immediate temptations from observing how similar others behave in situations that require self-control, such as effort tasks. Similarity between one's own and peers' willpower is important, so that peers' behavior is informative for oneself. Intuitively, peers who are too weak are most likely to procrastinate, while those who are too strong are most likely to exercise self-control. Consequently, failures of the former or successes of the latter tell little (if anything) about one's own willpower.

Assuming quasi-hyperbolic discounting, Battaglini et al. (2005) predict a non-monotonic relationship between peer information and individual behavior. More precisely, observing how peers handle similar temptations can be encouraging or discouraging: the effect depends on how people perceive their own vs. peers' willpower. Accordingly, the ideal peers are those who have slightly worse self-control problems than oneself because this makes their successes more encouraging and their failures less discouraging. Observing such peers can improve self-control in the sense that: "If they can do it, then so can I" (or "Even if they cannot do it, I can still do it").

The idea that social comparison serves the purpose of self-evaluation and thus affects individual behavior dates back at least to Festinger (1954). More recently, Falk and Knell (2004) have shown

[25]Throughout this paper, "procrastination" means shifting effort from a sooner to a later date (postponing), and "precrastination" means shifting effort from a later to a sooner date (doing beforehand).

[26]Quasi-hyperbolic discounting is a simple version of hyperbolic discounting (Strotz, 1956; Laibson, 1997) that implies declining discount rates.

that balancing downward and upward comparisons leads people to compare themselves with similar others in terms of environment and personal characteristics.[27] In a field study, Goldstein et al. (2008) find that people are more likely to follow the behavior of others with whom they share situational rather than (more important) personal similarities. In line with the mentioned theory and evidence, we expect subjects in our study to engage in social comparison when observing effort reallocations of other MTurk workers in a similar bonus task.

Exploiting situational similarity, we implement two social comparison treatments with varying peer information (see Section 4.3.2). In the implicit shift group, subjects receive information that 48% of past participants made dynamically consistent choices. In the explicit shift group, subjects are additionally informed about the average degree of intended procrastination by past participants (hereinafter *peer procrastination*). Guided by the theoretical predictions of Battaglini et al. (2005), we expect that the former peer information would encourage and the latter would not discourage time-consistent behavior if subjects are sufficiently confident in their own vs. peers' willpower. Note that subjects must already have some self-confidence in order to sign up for the bonus task after completing the HIT in week 1 and then to return in week 2. Additionally, Fedyk (2019) documents that individuals hold more optimistic beliefs about their own vs. peers' self-control problems. Taken together, this leads us to the following key hypotheses about the effects of social comparisons on intertemporal choice behavior if subjects' self-confidence is sufficiently high.

Hypothesis 1 *Subjects in the implicit shift condition behave more time-consistently than subjects in the control condition.*

Hypothesis 2 *Subjects in the explicit shift condition do not behave less time-consistently than subjects in the implicit shift condition.*

We explore these hypotheses using the present-bias parameter β and model-free measures of intertemporal choice behavior described in Section 4.4.2. Since subjects are randomly assigned into three parallel groups, treatment effects are independent of potential preference shocks that could affect individual choices. Hence, we can test the first hypothesis by comparing subjects' intertemporal choice behavior in the implicit shift and control groups. This gives us the effect of observing that 48% of peers made time-consistent effort allocations. The second hypothesis can be tested by comparing subjects' behavior in the explicit and implicit shift groups. This allows us to assess the impact of observing peer procrastination, where the average degree of peer procrastination is inseparable from the event of peer procrastination due to our experimental design. In this regard our study mirrors a realistic setting where people can observe whether, and if so, how much others procrastinate (e.g. study groups or jogging groups).

[27] Downward comparison means comparing oneself with others who are inferior or less fortunate, whereas upward comparison means comparing oneself with others who are superior or more fortunate (see Wills, 1981; Wood and Taylor, 1991).

4.4.2 Behavioral measures

To assess our predictions, we introduce behavioral measures using the following notations: e_t is the number of tables allocated to week 2 at time t, t measures 1-week intervals, $m = 60$ is the number of tables that subjects allocate between weeks 2 and 3 (effort budget), $w = 20$ is the minimum number of tables to complete each week (the minimum work).

For each allocation, we calculate effort budget shares for week 2 (hereinafter *budget shares*) as the proportion of tables allocated to week 2, namely $\frac{e_t}{m}$. By definition, budget shares are valued between $[0, 1]$. We expect social comparisons to affect budget shares in week 2, whereas budget shares in week 1 are independent of treatment.[28] We also consider difference measures for allocations to capture time-consistent behavior of subjects who subsequently allocate as many tables as initially planned and time-inconsistent behavior of subjects who subsequently deviate from their initial allocations. An absolute difference in budget shares between subsequent and initial allocations (*budget share difference*) corresponds to $\frac{e_2 - e_1}{m}$, which can vary between $[-1, 1]$. To explicitly account for the initial work plan that includes both the minimum work and the initial allocation, we calculate a relative change in work plans for week 2 between subsequent and initial decisions (*relative change in work plan*) as $\frac{e_2 - e_1}{e_1 + w}$ for each subject. These are our model-free measures of intertemporal choice behavior that allow examining the effects of social comparison at the intensive margin.

We also use effort allocations to infer parameters for a quasi-hyperbolic discounting model by assuming that individual discounted costs of effort are described as follows:

$$C_t(c_t, c_{t+1}, \ldots) = \delta^t c_t + \beta \sum_{\tau=t+1}^{\infty} \delta^\tau c_\tau, \tag{4.2}$$

where c_t denotes effort made in period t. We further assume that, for each subject, instantaneous costs of effort in period t are equal to effort made in that period.[29]

Using equation (4.2), we first infer the discount factor δ from initial allocations. In week 1, a subject is indifferent between counting $e_1 + w$ tables in week 2 and $m - e_1 + w$ tables in week 3, which implies $\delta = \frac{e_1 + w}{m - e_1 + w}$. Next, we use subsequent allocations to infer the present-bias parameter β by assuming that δ is time invariant, at least between two allocation decisions. This allows us to use the above expression for δ to obtain $\beta = \frac{(e_2 + w)(m - e_1 + w)}{(m - e_2 + w)(e_1 + w)}$, which is our model-based measure of dynamic in/consistency.

[28] As discussed in Section 4.3.3, initial allocations do not significantly differ across treatment groups (see Appendix Table 4.A3).

[29] Besides quasi-hyperbolic discounting, Augenblick et al. (2015) assume a convex instantaneous cost of effort function with the power parameter $\gamma > 1$. Note that we observe only two allocations per subject in our study, whereas the estimation of discount factor δ, present bias β and power parameter γ in Augenblick et al. (2015) requires at least three allocations per subject. In the MTurk setting, however, we strongly preferred keeping our experimental design as simple as possible.

There is obviously a link between different behavioral measures. For instance, negative (positive) budget share differences indicate present-biased (future-biased) behavior and values of zero indicate dynamic consistency. In other words, procrastination (precrastination) is associated with $\beta < 1$ ($\beta > 1$) and time consistency with $\beta = 0$. To examine behavioral patterns at the extensive margin, we create indicator variables β^* and $\tilde{\beta}$ that take value one for subjects displaying procrastination and time-consistent behavior, respectively, and zero otherwise.

4.5 Results

4.5.1 Within-group analysis

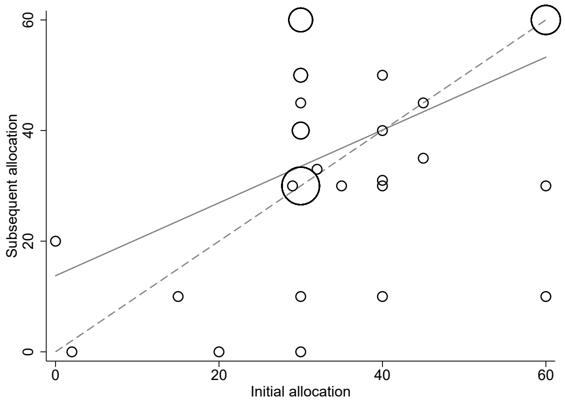
In the first place, we compare each worker's initial and subsequent allocation decisions in weeks 1 and 2, respectively. The decisions might be fairly unrelated given that it is randomly chosen which of them is implemented. Still, both decisions are made by the same person though seven days apart and in different conditions, except for the control group. Accordingly, Figure 4.1 displays significant positive unconditional correlation between initial and subsequent work allocations across the groups. Observations below the 45° line refer to procrastinators, on the line to subjects with time-consistent choices and above the line to precrastinators. Apparently, subsequent allocations often deviate from initial allocations, but consistently splitting effort equally between weeks 2 and 3 appears the most frequent behavior. The best-fit line with a slope of less than 45° confirms imperfect correlation between allocations in all groups. Correlation is the smallest under implicit shift treatment ($\rho = 0.419$, $p < 0.01$) and the largest under control ($\rho = 0.527$, $p < 0.01$). We further investigate statistical significance of the observed behavioral patterns.

Figure 4.2 presents distributions of initial and subsequent budget shares allocated to week 2 for each group. While a modal behavior is to equally split work between weeks 2 and 3 for each allocation, the distribution of subsequent allocations is somewhat shifted to the right vs. the distribution of initial allocations across all groups. This implies that subjects more often allocate more work to week 2 when it is in the future than when it occurs in the present. Hence, we observe a few procrastinators in each group. However, nonparametric tests, such as Kolmogorov-Smirnov test and Wilcoxon signed-rank test, do not detect significant distributional differences between initial and subsequent allocations to week 2 (see Appendix Table 4.A4). In the following, we analyze mean values of behavioral measures introduced in Section 4.4.2.

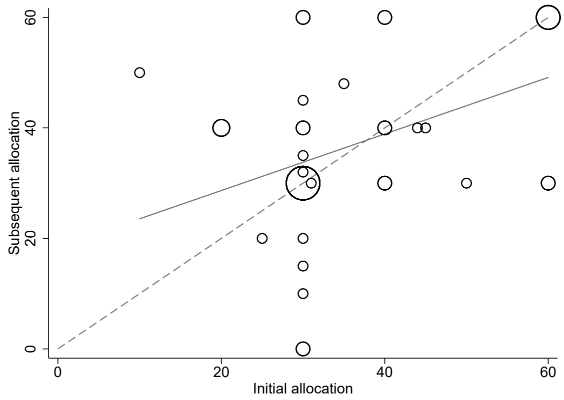
Figure 4.3 displays mean values and standard errors of each behavioral measure in different conditions. An average budget share subsequently allocated to week 2 exceeds 60% across all groups, varying from 61.6% under implicit shift treatment to 66.7% under explicit shift treatment. We find that subsequent budget shares for week 2 are on average 1.2–2.3 percentage points larger but not significantly different from initial budget shares in either group, given large standard errors (see Appendix Table 4.A4). Similarly, average relative changes in work plan for week 2 are

Figure 4.1: Initial vs. subsequent effort allocations to week 2

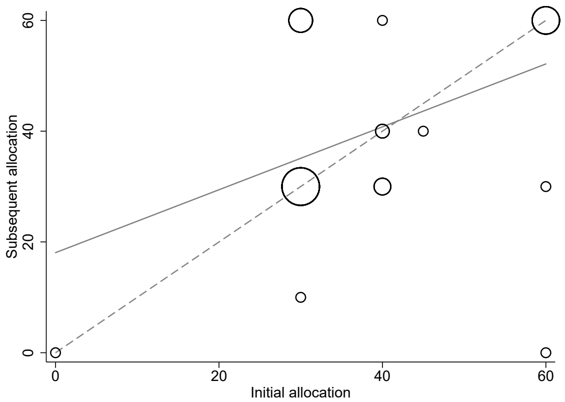
(a) Control



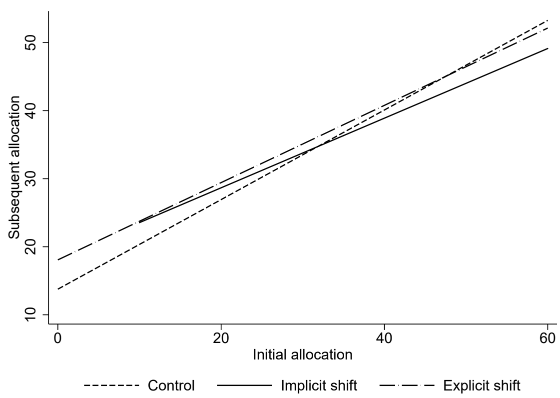
(b) Implicit shift



(c) Explicit shift



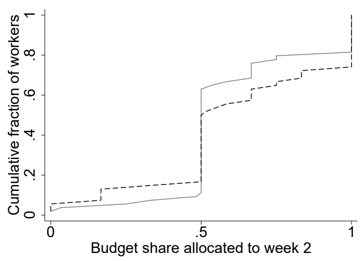
(d) Comparison of the best-fit lines



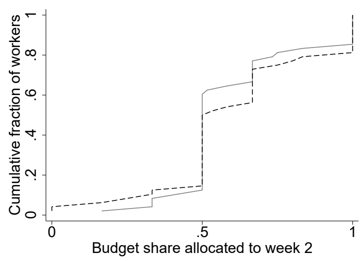
Note: Scatterplot markers are weighted by the number of overlapping data points; the best-fit line is solid and the 45° line is dashed. Linear correlations are: (a) $\rho = 0.527$, (b) $\rho = 0.419$, and (c) $\rho = 0.456$, with $p < 0.01$.

Figure 4.2: Distribution of initial and subsequent budget share allocations to week 2

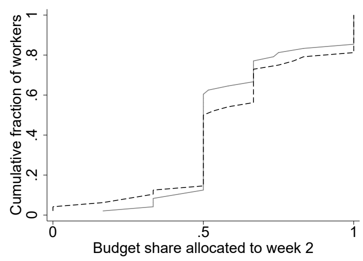
(a) Control



(b) Implicit shift



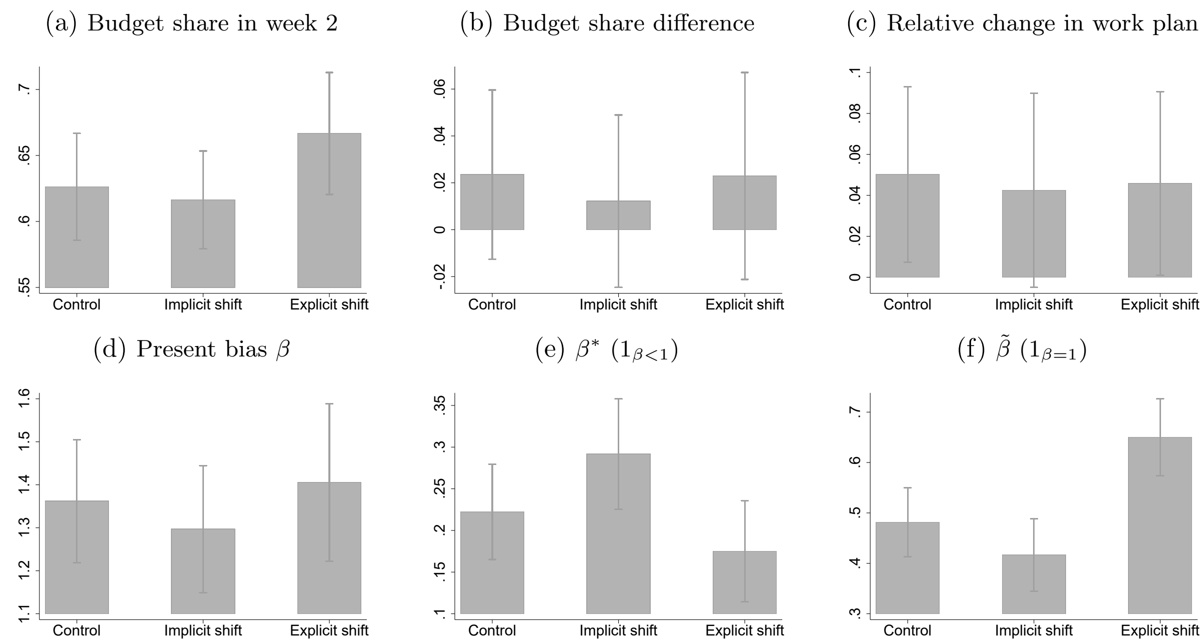
(c) Explicit shift



Note: The solid line displays week 1 allocations and the dash line shows week 2 allocations.

statistically insignificant ($p > 0.10$, two-sided t-tests). With respect to the present-bias parameter β, its mean values significantly exceed one in one-sided t-tests ($p < 0.01$ for control, $p < 0.05$ for social comparisons), which points to future-biased behavior on average. The proportion of procrastinators varies from 17.5% in the explicit shift group to 29.2% in the implicit shift group. We also observe a median pattern of dynamic consistency in each group (see Appendix Figure 4.A3).

Figure 4.3: Mean values of outcome variables by treatment status



Note: Standard error bars are shown around each mean.

We now turn to regression analyses to check whether the observed intertemporal choice behavior is not driven by omitted factors, such as the reminder in week 2 or demographic characteristics. The latter include gender, age and education, which are documented in the literature to correlate with intertemporal choices (e.g. Falk et al., 2018). Table 4.2 presents estimated constant and reminder coefficients for three dynamic measures in the control group. The first regression model (columns 'no x') controls for a reminder dummy only; the second specification (columns 'x') additionally controls for age (years), gender and educational attainment dummies. In either model specification, the estimated conditional mean values for budget share differences and relative changes in work plan are not statistically significant. With respect to the present-bias parameter β, the first specification gives a conditional mean significantly different from one at the 5% level, whereas the second specification yields a larger but statistically insignificant point estimate. Taken together, we conclude that subjects in the control group do not procrastinate and do not display present bias on average.[30]

Table 4.2: Conditional mean estimates

	Budget share difference		Relative change in work plan		Present bias β	
	no x	x	no x	x	no x	x
Constant	0.034	0.041	0.067	0.077	1.341	1.484
st. error	0.043	0.233	0.053	0.277	0.171	1.077
p-value	0.434	0.861	0.213	0.783	0.047	0.653
Reminder in week 2	-0.027	-0.012	-0.044	-0.025	0.055	0.130
st. error	0.080	0.087	0.093	0.104	0.317	0.351
p-value	0.741	0.892	0.640	0.810	0.864	0.713

Note: OLS results for the control group (54 observations). Estimated coefficients for β are tested against one. 'x' denotes demographic control variables. Standard errors are heteroscedasticity robust.

Result 1 *Subjects display heterogeneous intertemporal choices, i.e. not all subjects procrastinate. Subjects do not procrastinate and do not display present bias on average.*

4.5.2 Analysis of treatment effects

We have previously seen that subjects' initial and subsequent effort allocations positively but imperfectly correlate. To analyze treatment effects, we now compare this relationship across different conditions. Figure 4.1d displays upward sloping best-fit lines for the control and two treatment groups at once. Though we observe flatter slopes under social comparisons, especially in the implicit shift group, differences in slope coefficients are not statistically significant across the groups ($p > 0.10$, two-sided tests).

Next, we compare the mean values of our behavioral measures presented in Figure 4.3. Subjects in the implicit (explicit) shift group allocate on average the least (most) effort to week 2 when deciding in the present. The mean values of absolute and relative differences in effort allocations are the smallest in the implicit shift group. The same is also true for the present-bias parameter β. Standard errors of mean outcome values are large across the groups, implying wide confidence intervals that usually overlap. Not surprisingly, we find statistically insignificant differences in either mean values or distributions of these behavioral measures across the groups (see Appendix Table 4.A5).

At the extensive margin, subjects under implicit shift treatment appear the most likely to procrastinate and the least likely to behave time-consistently. The proportions of time-consistent and present-biased choices look the opposite under explicit shift treatment. Finally, we find that

[30] For the two social comparison groups, we reach qualitatively similar conclusions. Furthermore, when we include 17 subjects who dropped out in week 2 (implying zero budget shares, respectively) into the control group, the analysis of budget share differences still provides no support for present-biased behavior on average. Results are available on request.

subjects are significantly more likely to behave time-consistently in the latter than in the former group (see Appendix Table 4.A5).[31]

Though randomization worked pretty well, as discussed in Section 4.3.3, we also use regression analyses to take into account that intertemporal choices may systematically vary with reminder exposure in week 2 and/or demographic characteristics. Regression models for non-binary outcomes are estimated using OLS, but for the likelihood of procrastination β^* and the likelihood of time-consistent behavior $\tilde{\beta}$ we use probit regression.[32] Table 4.3 reports the estimated coefficients on a constant, treatment and reminder dummies for our behavioral measures. The control group is the reference group in the top panel, and the implicit shift group in the bottom panel. As before, we consider two model specifications. In columns 'no x', results are conditional on the reminder dummy; in columns 'x', demographic controls are added. Overall, the estimated coefficients seem quite robust to the list of controls. In the following, we refer to the point estimates from the second specification that is more conservative in terms of included controls.

Table 4.3: Estimated conditional treatment effects

	Budget share in week 2		Budget share difference		Relative change in work plan		Present bias β		β^* $(1_{\beta<1})$		$\tilde{\beta}$ $(1_{\beta=1})$	
	no x	x	no x	x	no x	x	no x	x	no x	x	no x	x
Implicit shift	-0.015	-0.019	-0.017	-0.023	-0.014	-0.023	-0.074	-0.098	0.079	0.080	-0.072	-0.059
st. error	0.056	0.058	0.052	0.054	0.064	0.066	0.208	0.217	0.085	0.085	0.097	0.096
p-value	0.790	0.750	0.747	0.667	0.822	0.732	0.724	0.650	0.352	0.348	0.463	0.536
Explicit shift	0.040	0.054	-0.002	-0.001	-0.006	-0.006	0.042	0.049	-0.046	-0.036	0.167	0.157
st. error	0.062	0.065	0.058	0.060	0.063	0.065	0.237	0.251	0.086	0.086	0.102	0.105
p-value	0.527	0.404	0.979	0.983	0.929	0.923	0.860	0.846	0.595	0.678	0.101	0.136
Reminder	-0.065	-0.069	-0.073	-0.079	-0.088	-0.095	-0.113	-0.130	0.133	0.133	-0.095	-0.081
st. error	0.053	0.057	0.051	0.052	0.056	0.059	0.194	0.202	0.076	0.077	0.085	0.085
p-value	0.220	0.225	0.153	0.135	0.120	0.108	0.562	0.521	0.081	0.086	0.262	0.342
Constant	0.652	0.819	0.052	0.106	0.084	0.121	1.406	1.594				
st. error	0.044	0.117	0.039	0.111	0.047	0.125	0.155	0.487				
p-value	0.000	0.000	0.192	0.340	0.076	0.334	0.009	0.222				
Reference: implicit shift												
Explicit shift	0.054	0.073	0.015	0.022	0.009	0.016	0.116	0.147	-0.115	-0.107	0.236	0.214
st. error	0.061	0.065	0.058	0.059	0.066	0.068	0.239	0.260	0.078	0.081	0.100	0.104
p-value	0.372	0.268	0.791	0.711	0.893	0.808	0.630	0.572	0.141	0.185	0.018	0.039

Note: Sample includes 142 observations with non-duplicated IP addresses. Effects are estimated using OLS, but for β^* and $\tilde{\beta}$ estimates are average marginal probit effects. Estimated coefficients for β under control are tested against one. 'x' denotes demographic control variables. Standard errors are heteroscedasticity robust.

Consistent with our previous analyses, subjects in the control group do not procrastinate on average. We again find that subjects in the implicit shift condition on average allocate less effort to week 2 when deciding in the present and behave more time-consistently than in the other conditions, but the estimated results are very imprecise and statistically insignificant. In addition, subjects in the implicit shift group are about 8 and 11 percentage points more likely to procrastinate than subjects in the control and explicit shift groups, respectively, but these estimates are also

[31] While the number of dynamically inconsistent choices is equally split between procrastinators and precrastinators under social comparisons, we observe more precrastinators than procrastinators under control.

[32] For binary outcomes β^* and $\tilde{\beta}$, multinomial probit regressions yield similar results, which are available on request.

statistically insignificant. In line with our descriptive analyses, we further find that subjects in the explicit shift group are significantly more likely to make dynamically consistent choices. The difference of 21.4 percentage points vs. the implicit shift group — the effect of observing peer procrastination — is significant at the 5% level, and the marginal effect of 15.7 percentage points vs. the control group is significantly positive at the 10% level in a one-sided test.[33]

As to reminder exposure in week 2, the respective estimates indicate that subjects who log in later on the work date subsequently allocate less effort to week 2 when deciding in the present. The associations with differences in budget shares and relative changes in work plan are significantly negative at the 10% level in one-sided tests. Besides, subjects who started working on the effort task after the reminder are 13.3 percentage points more likely to procrastinate than earlier performers. This is a sizable and statistically significant difference. The estimated coefficients of demographic variables are statistically insignificant.[34]

Result 2 *Observing that 48% of previous participants behaved time-consistently does not significantly affect intertemporal choice behavior of subjects in the overall sample.*

Result 3 *Observing that (some) previous participants on average shifted 23 tables to the later date makes subjects significantly more likely to behave time-consistently in the overall sample.*

In order to understand whether the imprecisely estimated treatment effects stem from effect heterogeneity, we now investigate treatment effects across different subgroups defined by gender, age and education. We first examine whether social comparisons influence men and women differentially. Table 4.4 presents the results of our regression models that allow for interactions between female and treatment dummies. Again, we control for reminder exposure in the first specification (columns 'no x') and further add demographic controls in the second specification (columns 'x').

In the control group, we find that men do not procrastinate on average, and women behave significantly more time-consistently in terms of the present-bias parameter β. Besides, men in the implicit shift group are by 18 percentage points more likely to display time consistency than in the control group. This effect, driven by the reduced share of precrastinators, is significant at the 10% level. Contrary, women in the implicit shift group are by 32 percentage points significantly less likely to behave time-consistently than in the control group ($p < 0.01$, two-sided test). A highly significant differential effect of -47.3 percentage points implies that the implicit shift treatment has more negative effect on time consistency of women than men. On the female side, this effect is largely driven by the increased fraction of precrastinators.

Though we also observe directionally opposite effects on intertemporal choice at the intensive margin, gender differences in implicit shift effects appear statistically significant for present bias only. We find that the estimated present-bias parameter β on average significantly decreases by

[33]Under the assumption of symmetric distribution, one-sided p-values correspond to the half of two-sided p-values.
[34]Results are available on request.

Table 4.4: Estimates with female interaction terms

	Budget share in week 2		Budget share difference		Relative change in work plan		Present bias β		β^* ($1_{\beta<1}$)		$\tilde{\beta}$ ($1_{\beta=1}$)	
	no x	x	no x	x	no x	x	no x	x	no x	x	no x	x
Implicit shift (I)	-0.069	-0.065	-0.062	-0.062	-0.101	-0.101	-0.501	-0.496	0.030	0.035	0.184	0.180
st. error	0.085	0.085	0.077	0.077	0.092	0.092	0.297	0.304	0.119	0.116	0.105	0.104
p-value	0.415	0.445	0.422	0.420	0.274	0.275	0.094	0.105	0.801	0.762	0.079	0.082
Explicit shift (E)	0.011	0.025	-0.028	-0.029	-0.059	-0.062	-0.263	-0.262	-0.005	-0.004	0.217	0.219
st. error	0.092	0.094	0.084	0.087	0.096	0.099	0.357	0.373	0.122	0.120	0.128	0.128
p-value	0.906	0.788	0.739	0.736	0.540	0.535	0.462	0.485	0.966	0.975	0.090	0.088
Female x I	0.109	0.096	0.090	0.080	0.175	0.160	0.858	0.809	0.089	0.086	-0.487	-0.473
st. error	0.113	0.118	0.106	0.109	0.130	0.132	0.419	0.434	0.181	0.179	0.102	0.110
p-value	0.335	0.421	0.396	0.463	0.182	0.227	0.043	0.065	0.625	0.633	0.000	0.000
Female x E	0.054	0.057	0.050	0.056	0.104	0.111	0.610	0.627	-0.074	-0.071	-0.112	-0.114
st. error	0.128	0.129	0.123	0.123	0.132	0.133	0.489	0.498	0.152	0.152	0.183	0.184
p-value	0.673	0.658	0.687	0.647	0.433	0.405	0.214	0.210	0.628	0.639	0.542	0.535
Female	-0.086	-0.080	-0.081	-0.077	-0.127	-0.121	-0.608	-0.585	0.063	0.063	0.143	0.129
st. error	0.083	0.087	0.076	0.078	0.089	0.090	0.292	0.303	0.113	0.112	0.112	0.114
p-value	0.301	0.356	0.284	0.326	0.153	0.183	0.039	0.056	0.578	0.576	0.201	0.258
Constant	0.693	0.838	0.091	0.123	0.146	0.155	1.700	1.781				
st. error	0.066	0.114	0.065	0.112	0.077	0.129	0.249	0.493				
p-value	0.000	0.000	0.167	0.271	0.062	0.230	0.005	0.113				
Reference: men, implicit shift												
Explicit shift	0.080	0.091	0.034	0.033	0.043	0.039	0.238	0.234	-0.034	-0.037	0.017	0.024
st. error	0.084	0.084	0.064	0.067	0.073	0.076	0.293	0.311	0.119	0.116	0.141	0.143
p-value	0.345	0.285	0.599	0.621	0.562	0.607	0.419	0.453	0.775	0.747	0.904	0.865
Female x E	-0.055	-0.039	-0.041	-0.024	-0.071	-0.049	-0.247	-0.182	-0.137	-0.134	0.383	0.365
st. error	0.124	0.121	0.122	0.118	0.137	0.136	0.494	0.503	0.125	0.127	0.124	0.136
p-value	0.659	0.751	0.739	0.841	0.607	0.719	0.617	0.718	0.270	0.293	0.002	0.007
Female	0.023	0.015	0.009	0.003	0.047	0.039	0.250	0.224	0.146	0.143	-0.364	-0.357
st. error	0.077	0.079	0.075	0.075	0.096	0.097	0.303	0.307	0.114	0.115	0.093	0.095
p-value	0.768	0.846	0.903	0.967	0.624	0.687	0.411	0.468	0.199	0.211	0.000	0.000

Note: Sample includes 142 observations with non-duplicated IP addresses. Effects are estimated using OLS, but for β^* and $\tilde{\beta}$ estimates are average marginal probit effects. Estimated coefficients for β under control are tested against one. 'x' denotes demographic control variables: age in years and education dummies; all regressions control for the week 2 reminder. Standard errors are heteroscedasticity robust.

about 50 percentage points for men ($p < 0.1$, one-sided tests) but does not significantly change for women ($p > 0.1$, one-sided tests).

In addition, men in the explicit shift group are by 21.9 percentage points significantly more likely to behave time-consistently compared to the control group, which is not a significantly larger difference than the implicit shift effect discussed above. Hence, peer procrastination does not seem to significantly influence intertemporal choices of men. At the same time, it has by 36.5 percentage points significantly more positive effect on the likelihood of time-consistent behavior among women. That is, women under explicit shift treatment are by 40 percentage points more likely to behave time-consistently than under implicit shift treatment, which is a highly significant difference ($p < 0.01$, two-sided test). Negative effects of peer procrastination on the likelihood of present-biased choices are not found to significantly differ by gender. There is also no evidence of heterogeneous explicit shift effects at the intensive margin.

Result 4 *Observing that 48% of previous participants behaved time-consistently affects men and women differentially: men are significantly more likely to behave time-consistently, whereas women* *are significantly less likely to do so; estimates of the quasi-hyperbolic present-bias parameter* β *significantly decrease among men on average.*

Result 5 *Observing that (some) previous participants on average shifted 23 tables to the later date does not significantly affect men but makes women significantly more likely to behave time-consistently.*

Next, we examine treatment effect heterogeneity for younger and older subjects, distinguished by a median age of 36 years. Specifically, we create a dummy variable that takes value one for subjects at least 36 years old and zero otherwise, which is interacted with treatment dummies in our regression models. Appendix Table 4.A6 shows the results. Subjects in neither age group procrastinate on average. Under implicit shift treatment, the proportion of time-consistent choices decreases among the younger, and the proportion of procrastinators increases among the older, implying an opposite impact on precrastination across the age subgroups. However, the point estimates are not statistically significant either at the extensive or at the intensive margin. We further find that the explicit shift effect on time-consistent behavior is by 40.1 percentage points significantly larger among the older than the younger. But there is no strong evidence that peer procrastination produces significantly different effects on the age subgroups.

To explore treatment effect heterogeneity by schooling, we create a dummy variable that takes value one for subjects with at least a 2-year college degree and zero otherwise, and further interact it with treatment dummies. Appendix Table 4.A7 reports regression results. We find that less educated subjects in the control group display significant future-based choices, with $\beta = 1.5$ on average. In the implicit shift group, subjects appear more likely to procrastinate and less likely to behave time-consistently, in particular if they are less educated. However, neither effect is statistically significant. While observing peer procrastination significantly reduces the proportion of procrastinators among less educated subjects by 19.6 percentage points, we do not find significant effect heterogeneity between the two subgroups.

Taken together, our analyses show that treatment effects of our social comparisons are not homogeneous but notably vary by gender. That is, women appear more sensitive to the contents of peer information than men. Differential treatment effects are most pronounced in the likelihood of time-consistent behavior and in the present-bias parameter β.

4.5.3 Robustness checks

We now check whether our main results are robust to estimation approaches. Since our behavioral measures at the intensive margin are by definition bounded, we re-estimate treatment effects on the corresponding outcomes in the overall sample using two-limit tobit regression. We also run logistic regression for our probability measures at the extensive margin. Appendix Table 4.A8 presents the results. The estimated average marginal tobit effects appear robust to the list of included controls. While directionally consistent with the OLS results reported in Table 4.3, our tobit estimation yields somewhat larger effects on the budget share in week 2 but smaller effects on two difference measures and the present-bias parameter β. As before, none of the treatment effects is statistically significant at the intensive margin.

For the likelihood of present-biased and time-consistent choices, the average marginal logistic estimates are comparable to our probit results. That is, observing peer procrastination significantly increases the likelihood of time-consistent behavior in the overall sample. The reminder estimates remain negative and mostly insignificant at the intensive margin, but significantly positive for the likelihood of present-biased behavior.

Similarly, we use tobit and logistic regressions to re-assess treatment effect heterogeneity by gender. The corresponding average marginal tobit effects are reported in Table 4.5. Directionally, our tobit results replicate the OLS estimates presented in Table 4.4. With respect to magnitude, we again observe fairly larger effects on the budget share but smaller effects on the difference measures and on present bias compared to OLS. Nevertheless, we confirm our previous findings that implicit shift effects on the present-bias parameter β are significantly negative for men, and that there is no evidence of significant gender differences in explicit shift effects at the intensive margin.

Table 4.5: Estimates with female interaction terms using tobit/logistic regression

	Budget share in week 2		Budget shares difference		Relative change in work plan		Present bias β		β^* $(1_{\beta<1})$		$\tilde{\beta}$ $(1_{\beta=1})$	
	no x	x	no x	x	no x	x	no x	x	no x	x	no x	x
Implicit shift (I)	-0.086	-0.081	-0.033	-0.033	-0.057	-0.056	-0.135	-0.133	0.031	0.037	0.178	0.173
st. error	0.086	0.085	0.039	0.039	0.050	0.049	0.074	0.074	0.122	0.118	0.102	0.101
p-value	0.318	0.344	0.408	0.398	0.261	0.254	0.068	0.072	0.797	0.755	0.080	0.087
Explicit shift (E)	0.016	0.039	-0.015	-0.015	-0.033	-0.034	-0.071	-0.070	-0.008	-0.005	0.213	0.217
st. error	0.100	0.101	0.043	0.043	0.052	0.053	0.091	0.093	0.125	0.124	0.126	0.127
p-value	0.871	0.704	0.729	0.722	0.525	0.514	0.440	0.454	0.950	0.970	0.091	0.087
Female x I	0.123	0.106	0.048	0.042	0.097	0.089	0.230	0.217	0.083	0.077	-0.477	-0.463
st. error	0.114	0.118	0.055	0.055	0.072	0.071	0.100	0.102	0.184	0.180	0.098	0.106
p-value	0.281	0.369	0.386	0.446	0.176	0.213	0.022	0.034	0.650	0.668	0.000	0.000
Female x E	0.062	0.058	0.024	0.028	0.055	0.059	0.156	0.161	-0.075	-0.076	-0.110	-0.112
st. error	0.137	0.136	0.064	0.062	0.073	0.071	0.124	0.123	0.151	0.150	0.181	0.182
p-value	0.650	0.666	0.701	0.653	0.451	0.411	0.209	0.191	0.620	0.614	0.545	0.539
Female	-0.105	-0.097	-0.043	-0.041	-0.071	-0.067	-0.163	-0.157	0.061	0.061	0.139	0.126
st. error	0.085	0.088	0.039	0.040	0.049	0.049	0.072	0.073	0.115	0.113	0.109	0.112
p-value	0.219	0.273	0.276	0.309	0.147	0.169	0.023	0.032	0.593	0.591	0.204	0.260
Reference: men, implicit shift												
Explicit shift	0.102	0.119	0.018	0.018	0.024	0.022	0.064	0.063	-0.038	-0.040	0.018	0.029
st. error	0.091	0.090	0.032	0.033	0.039	0.040	0.074	0.077	0.121	0.118	0.140	0.143
p-value	0.263	0.183	0.581	0.594	0.541	0.577	0.387	0.409	0.754	0.735	0.897	0.840
Female x E	-0.061	-0.048	-0.023	-0.014	-0.043	-0.030	-0.074	-0.056	-0.132	-0.129	0.374	0.356
st. error	0.131	0.126	0.062	0.059	0.074	0.072	0.127	0.126	0.122	0.125	0.120	0.132
p-value	0.644	0.704	0.706	0.808	0.565	0.674	0.560	0.657	0.280	0.302	0.002	0.007
Female	0.018	0.010	0.005	0.002	0.026	0.022	0.067	0.060	0.140	0.134	-0.359	-0.350
st. error	0.077	0.077	0.038	0.037	0.052	0.051	0.076	0.076	0.114	0.115	0.092	0.094
p-value	0.812	0.899	0.898	0.966	0.609	0.670	0.377	0.432	0.220	0.240	0.000	0.000

Note: Sample includes 142 observations with non-duplicated IP addresses. Average marginal tobit effects are on actual variables. Estimates on β^* and $\tilde{\beta}$ are average marginal logistic effects. 'x' denotes demographic control variables: age in years and education dummies; all regressions control for the week 2 reminder. Standard errors are heteroscedasticity robust.

At the extensive margin, our logistic results are very similar to those obtained using probit estimation. In line with our previous findings, the implicit shift treatment has significantly more negative effect on the likelihood of time-consistent behavior of women compared to men. In addition, peer procrastination has significantly more positive impact on the likelihood of women's time consistency.

Table 4.6: Estimates with female interaction terms using absolute outcome values

	OLS				Tobit			
	Budget shares difference		Relative change in work plan		Budget shares difference		Relative change in work plan	
	no x	x	no x	x	no x	x	no x	x
Implicit shift (I)	-0.126	-0.121	-0.152	-0.146	-0.113	-0.108	-0.131	-0.126
st. error	0.057	0.057	0.072	0.072	0.056	0.055	0.066	0.065
p-value	0.030	0.037	0.036	0.045	0.044	0.049	0.046	0.052
Explicit shift (E)	-0.108	-0.106	-0.151	-0.149	-0.102	-0.101	-0.133	-0.131
st. error	0.065	0.067	0.075	0.078	0.061	0.062	0.069	0.070
p-value	0.100	0.118	0.047	0.059	0.094	0.102	0.054	0.061
Female x I	0.271	0.256	0.327	0.304	0.253	0.239	0.294	0.273
st. error	0.077	0.079	0.100	0.102	0.071	0.074	0.085	0.087
p-value	0.001	0.002	0.001	0.004	0.000	0.001	0.001	0.002
Female x E	0.188	0.187	0.227	0.228	0.136	0.136	0.155	0.156
st. error	0.100	0.101	0.106	0.109	0.097	0.097	0.105	0.105
p-value	0.061	0.066	0.034	0.038	0.162	0.161	0.140	0.139
Female	-0.157	-0.148	-0.197	-0.185	-0.120	-0.111	-0.145	-0.133
st. error	0.054	0.056	0.065	0.066	0.051	0.053	0.059	0.060
p-value	0.005	0.009	0.003	0.006	0.019	0.034	0.014	0.026
Constant	0.212	0.176	0.269	0.244				
st. error	0.048	0.092	0.060	0.108				
p-value	0.000	0.059	0.000	0.025				
Reference: men, implicit shift								
Explicit shift	0.018	0.014	0.000	-0.003	0.011	0.008	-0.002	-0.005
st. error	0.059	0.061	0.067	0.071	0.066	0.066	0.075	0.076
p-value	0.763	0.815	0.996	0.965	0.873	0.909	0.979	0.945
Female x E	-0.083	-0.069	-0.099	-0.076	-0.117	-0.103	-0.139	-0.117
st. error	0.099	0.099	0.113	0.116	0.095	0.094	0.106	0.107
p-value	0.402	0.487	0.381	0.517	0.216	0.274	0.191	0.273
Female	0.115	0.108	0.129	0.118	0.134	0.128	0.149	0.140
st. error	0.054	0.054	0.077	0.077	0.050	0.050	0.062	0.062
p-value	0.036	0.048	0.094	0.128	0.007	0.010	0.016	0.025

Note: Sample includes 142 observations with non-duplicated IP addresses. Estimated OLS coefficients for β under control are tested against one. Average marginal tobit effects are on actual variables. 'x' denotes demographic control variables: age in years and education dummies; all regressions control for the week 2 reminder. Standard errors are heteroscedasticity robust.

We have so far detected significant effect heterogeneity between men and women at the extensive margin and mostly insignificant effects at the intensive margin, except for the present-bias parameter in the implicit shift vs. the control condition. To verify the presence of effect heterogeneity within gender groups at the intensive margin, we run OLS and tobit regressions for absolute values of our difference measures. Significant estimates would indicate heterogeneous reaction (in the opposite direction) to social comparison within gender groups. Table 4.6 presents the results. The estimated average marginal tobit effects are mostly comparable to the OLS estimates in terms of direction, magnitude and significance. The results suggest that the implicit shift treatment on average makes men deviate from their initial choices significantly less. We also find significantly larger positive implicit shift effects on absolute outcome values for women, implying that women deviate from their initial choices significantly more than men in that condition. Besides, there is no statistically significant difference between explicit and implicit shift effects for men. Hence, men do not seem to care about peer procrastination. We also do not find that peer procrastination produces significant effect heterogeneity between men and women. Taken together, our analysis provides some evidence on effect heterogeneity within gender groups at the intensive margin, in particular for the implicit shift treatment.

4.6 Discussion

4.6.1 Treatment effects

Our main results concern social comparison effects on intertemporal choice. For the overall sample, we find that peer procrastination makes subjects significantly more likely to behave time-consistently due to discouraged procrastination and precrastination. We also find that neither observing that 48% of past participants made time-consistent choices nor observing peer procrastination significantly affects individual behavior at the intensive margin. Largely imprecise point estimates have led us to examine treatment effect heterogeneity along demographic dimensions, including gender.

Our heterogeneity analyses reveal that men and women react differentially to social comparisons. Observing that 48% of peers made time-consistent choices, men are more likely to behave time-consistently, while women are less likely to do so. On average, men also exhibit significantly smaller estimates of the present-bias parameter of a quasi-hyperbolic discounting model. Observing peer procrastination makes women more likely to behave time-consistently but does not seem to affect men. Consequently, women appear more sensitive to the contents of peer information than men.[35] While Buechel et al. (2018) document directionally homogeneous peer effects in a non-competitive setting, they also find that information shared by a successful peer has a more pronounced effect on the effort perseverance of women than men.

Effect heterogeneity may stem from gender differences in self-confidence. Psychology literature suggests that women's self-confidence can be more dependent than men's on situational factors, such as the nature of a task, the availability of clear performance feedback and the nature of social comparison (see Lenney, 1977, for a review). In our study, the task is simple and gender-neutral,

[35] One can argue that the found gender differences can be an artifact of sample size. Note that Falk and Ichino (2006) drew conclusions about the presence of peer effects based on a sample of 24 participants, eight in the single treatment and 16 (eight pairs) in the pair treatment, whereas we have 18–27 subjects per gender-condition.

performance feedback is clear, but peer information is incomplete. The latter could potentially weaken women's vs. men's self-confidence.

If men are sufficiently self-confident, their behavior is consistent with Battaglini et al. (2005) because the fraction of men who behaved time-consistently increased in both social comparison conditions, and the explicit information about peer procrastination did not have a discouraging effect. Due to our experimental design, we cannot additionally check whether the mere reference to peers — a kind of social facilitation (see Zajonc, 1965) — could contribute to men's dynamic consistency.

If women lack self-confidence, Battaglini et al. (2005) predict procrastination. However, the fraction of women behaving time-inconsistently significantly increased in the implicit shift condition, and this increase was largely driven by future-biased choices. We suppose that women could engage in upward comparison with past participants who potentially exercised more effort in week 2 because of larger initial and/or subsequent effort allocations to that week. The motive of self-improvement (see Falk and Knell, 2004) could then encourage women to subsequently allocate to week 2 more effort than initially planned.[36] Additional information about peer procrastination in the explicit shift condition could help women gain confidence in their initial allocations, leading to more time-consistent choices.

While we find significant heterogeneous treatment effects on the likelihood of time-consistent choices between men and women, the estimated effects are mostly insignificant at the intensive margin. Examining absolute values of our difference measures for effort allocations, we provide some evidence on heterogeneous responses to social comparisons within gender groups, at least for observing that 48% of peers behaved time-consistently. Response heterogeneity may result from differences in individual beliefs about the whole distribution of peers' choices. Future research can shed light on that. For instance, separate treatments may incentivize subjects to guess the average effort allocations of past participants in weeks 1 and 2, respectively, and the percentage of those who made present-biased (future-biased) choices.

In addition, social comparison with 48% (or implied 52%) of past participants may be too weak to induce directionally homogeneous effects.[37] Unless the observed frequency rates of behavioral types are sufficiently far from 50%, future research may examine whether truthful peer information in the spirit of DellaVigna and Pope (2018) can induce systematic dynamically consistent behavior. For instance: "Previously, [*many*] MTurk workers who completed a similar Bonus Task [*most often*] decided to do as many tables today as originally planned".

Finally, insufficient sample size may at least partly explain imprecise results in our study. To assess this issue, we conduct post-hoc power analysis. The power analysis reveals that a group size

[36]For a more difficult task, upward comparison could have had a discouraging effect on women, resulting in procrastination.

[37]In the field experiment of Frey and Meier (2004), the information that 46% of students contributed to charitable funds in the past insignificantly affected the likelihood of donation.

of 50 subjects is sufficient to detect *larger* effect sizes of 0.5 standard deviations in one-sided t-tests with a power of 0.8 and the 5% significance level. We further duplicate our sample ten times and re-run regressions with female interaction terms to illustrate that our point estimates gain precision in a larger sample. Social comparison effects indeed become statistically significant across different measures at the intensive margin, in particular for observing the fraction of time-consistent choices (see Appendix Table 4.A9).

4.6.2 Time in/consistency

Our analyses also suggest that subjects do not display present bias on average. This finding differs from the result of Augenblick et al. (2015), who document significant present bias in real-effort tasks. Despite some differences in the experimental design, our control condition closely compares to Augenblick et al. (2015) at the task rate of one.[38] In their study, this task rate yields 65.6% time-consistent, 23.7% present-biased and 10.6% future-biased choices. Present-biased choices obviously outweigh future-biased choices leading to significant present bias on average. In our study, the distribution of behavioral types under control looks as following: 48.2% subjects behave time-consistently, 22.2% show present bias and 29.6% show future bias. Compared to Augenblick et al. (2015), we observe almost three times as large proportion of precrastinators, while the proportion of procrastinators is fairly the same.

Stronger self-selection in our study may drive this difference. While our subjects get the task description beforehand, Augenblick et al.'s (2015) subjects find out the task details once in the lab.[39] The physical presence in the lab imposes more restrictions than online participation on MTurk. Another feature of our effort task is that subjects receive immediate feedback on correct or wrong counting, potentially also leading to self-selection. Though subjects who particularly enjoy (are good at) counting may self-select into the bonus task,[40] they still have an incentive to procrastinate because the bonus is paid in the final week only. To gauge self-selection bias in future research, subjects may be asked in an exit survey to indicate how much they enjoyed the task. If self-selection indeed leads to less present-biased behavior on average, employers may mitigate procrastination at work by providing as detailed as possible job descriptions at an early stage of a recruitment process.

With respect to the type and length of the task, our study is comparable to Augenblick et al. (2015), but MTurk workers and UC Berkeley students may perceive routine work differently. While the former can spend a few hours a day doing simple repetitive tasks on MTurk, the latter probably

[38] Augenblick et al. (2015) introduce five task rates ranging from 0.5 to 1.5, so that subjects can minimize the total number of tasks by allocating to a sooner date the cap of 50 tasks at the task rate of 0.5 and zero tasks at the task rate of 1.5. We implement only the task rate of one because it is easy for subjects to understand that every table allocated to a sooner date reduces the number of tables allocated to a later date by one.

[39] In Augenblick et al. (2015), subjects make initial and subsequent allocations under different work environments: in and outside the lab, respectively.

[40] In one of our previous trials, we received a message explicitly stating that a given participant enjoyed counting.

do not perform such tasks on a daily basis and thus procrastinate more. Besides, MTurk workers may procrastinate less if MTurk is an important source of their income. Platform-specific shocks on the effort allocation days (e.g. too few exciting HITs offered in week 2) could also induce future-biased behavior in our study. However, similar to Augenblick et al. (2015), we may underestimate procrastination when subjects choose to substitute other activities with our bonus task.

Finally, the longitudinal design implies that subjects make initial and subsequent allocations under different informational environments. Uncertain about their future schedule, subjects may choose to complete more tasks in the present to resolve uncertainty, which likely leads to future-biased choices (Augenblick et al., 2015; Augenblick and Rabin, 2019). MTurk workers may just carry stronger preference for the resolution of uncertainty than UC Berkeley students.

All in all, different subject pools may exhibit different behavioral patterns. DellaVigna and Pope (2018) also report the lack of support for present bias among MTurk workers in a real-effort task. However, they vary the time distance between the effort and the payment — not between the effort decision and the effort itself — from nearly immediate (within 24 hours) to 2 or 4 weeks later. Hence, their result may also reflect the 24-hour delay in pay (Balakrishnan et al., 2017).

4.6.3 Further insights

We further notice that procrastinators somewhat delay effort on the work date, which relates to the occasionally significant association between the week 2 reminder and the procrastination probability. Note that in week 2 we observe seven subjects who logged in less than two hours before the midnight deadline and only four of them procrastinated. The most severe delayer had at least 35 minutes to work on the task, while a median subject competed work about 12.4 hours before midnight. Hence, there is no evidence that a physical time constraint systematically drives present-biased behavior. Another question is whether subjects who took longer to complete the minimum work in week 2 are more likely to procrastinate. For each subject, we infer an average completion time per table in week 2.[41] A median subject completed one table approximately in 25 seconds, and we find no evidence that procrastinators on average spent more time per table ($p > 0.1$, two-sided t-test). Augenblick et al. (2015) also observe that subjects with present-biased choices may log in somewhat later but not particularly close to the deadline.

We now turn to experimental design challenges. Unlike in laboratory experiments, the experimenter cannot control communication among subjects in online experiments, which jeopardizes the stable unit treatment value assumption (SUTVA; Rubin, 1974). In our sample, subjects are geographically spread across 35 US states and thus are less likely to know each other.[42] Recruiting geographically dispersed MTurk workers creates less threats to design validity than recruiting students at the same university because geographical proximity increases chances of student interaction between effort allocation dates. However, MTurk workers may discuss task contents online.

[41] Qualtrics tracks only the total session time.

[42] In week 2, about 98% of subjects were located in the US, and the remainder logged in from elsewhere.

Browsing online forums, we find only few references to accomplished bonus payments, i.e. nothing to concern about.

Our final point relates to experimenter demand effect. Though subjects are not informed that they participate in a scientific study, they may wish to please the requester by deviating little from initial decisions, mainly in the direction of precrastination. Motives may include, for instance, an expectation of future collaboration or reciprocity from the requester in the form of an extra bonus. Exposing MTurk workers to different demand treatments, de Quidt et al. (2018) estimate bounds on demand-free behavior over canonical tasks, including intertemporal choices. Their findings point to typically modest experimenter demand effects under weak treatment if anything. Hence, experimenter demand must be of minor concern in our quasi-field study. A future experiment may include an exit survey question on what subjects think the requester's behavioral expectations are.

4.7 Conclusion

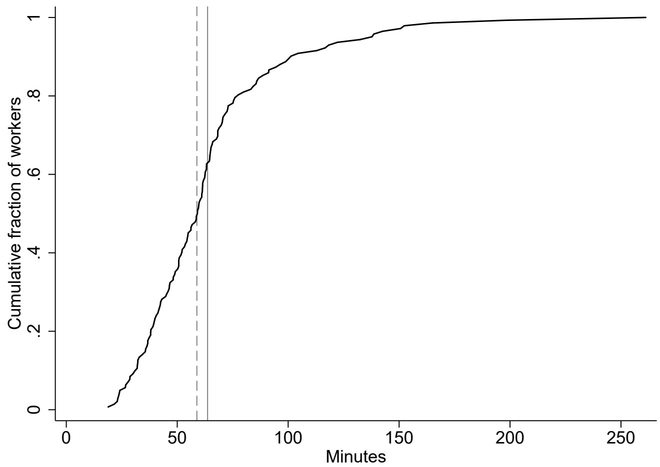
In a real-effort experiment, we recruit MTurk workers for three weeks to examine whether social comparison affects intertemporal choice behavior. Subjects allocate and subsequently allocate again units of effort between two work dates. Prior to subsequent allocations, subjects are randomly assigned to one of three conditions, with or without social comparison. In two social comparison conditions, subjects receive varying information about reallocation decisions of past participants. Behavioral measures based on differences between initial and subsequent allocations enable us to examine social comparison effects on intertemporal choice.

Our findings indicate that social comparisons affect men and women differentially, leading to mostly insignificant effects in the overall sample. More precisely, observing that 48% of past participants displayed time consistency makes men significantly more likely to behave time-consistently and women significantly less likely to do so. On average, men also exhibit significantly smaller estimates of the present-bias parameter of a quasi-hyperbolic discounting model. We additionally find that observing peer procrastination significantly increases the likelihood of time-consistent behavior among women but has no impact on men. These findings suggest that social comparison based on situational similarity affects intertemporal choice in effort. More important, gender-specific social comparison can provide an effective solution to dynamic inconsistency even when individual time preferences are not known beforehand, which is often the case in practice.

Finally, unlike Augenblick et al. (2015), who document significant present bias in effort, we find that subjects in our control condition do not procrastinate on average. Differences in subject pools and self-selection appear plausible explanations thereof. If self-selection indeed mitigates procrastination, it might be desirable at work.

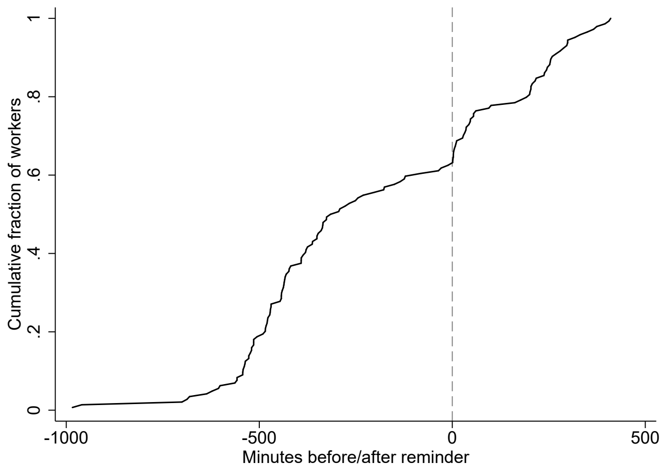
4.A Additional figures and tables

Figure 4.A1: Distribution of task completion time



Note: The vertical dashed line refers to the median and the vertical solid line to the sample mean.

Figure 4.A2: Distribution of time passed between a personal login and the reminder in week 2



Note: The dashed line refers to the reminder time point. Negative distance = started working before the reminder, positive distance = started working after the reminder.

Table 4.A1: Descriptive statistics by dropout status

	Completed	Dropouts	Difference	p-value
Initial allocation (tables)	36.887	36.579	-0.308	0.928
	(13.644)	(16.334)	(3.414)	
Discount factor δ	1.682	1.770	0.089	0.769
	(1.213)	(1.388)	(0.301)	
Female	0.486	0.579	0.093	0.449
	(0.502)	(0.507)	(0.123)	
Age (years)	36.894	35.421	-1.473	0.532
	(9.574)	(9.974)	(2.350)	
At most high school	0.183	0.211	0.027	0.775
	(0.388)	(0.419)	(0.096)	
Some college (at most 2 years)	0.387	0.421	0.034	0.779
	(0.489)	(0.507)	(0.120)	
At least 4-year college	0.430	0.368	-0.061	0.615
	(0.497)	(0.496)	(0.121)	
Observations	142	19	161	

Note: Standard deviations are reported in parentheses.

Table 4.A2: Descriptive statistics by reminder exposure in week 2

	No reminder	Reminder	Difference	p-value
Initial allocation (tables)	36.641	37.130	0.489	0.826
	(14.072)	(13.840)	(2.225)	
Discount factor δ	1.679	1.710	0.031	0.876
	(1.226)	(1.244)	(0.197)	
Female	0.467	0.536	0.069	0.390
	(0.502)	(0.502)	(0.080)	
Age (years)	37.609	35.536	-2.072	0.176
	(10.060)	(8.889)	(1.525)	
At most high school	0.185	0.188	0.004	0.954
	(0.390)	(0.394)	(0.062)	
Some college (at most 2 years)	0.359	0.435	0.076	0.331
	(0.482)	(0.499)	(0.078)	
At least 4-year college	0.457	0.377	-0.080	0.314
	(0.501)	(0.488)	(0.079)	
Observations	92	69	161	

Note: Standard deviations are reported in parentheses.

Table 4.A3: Descriptive statistics by treatment status

	Control (C)	Implicit shift (I)	Explicit shift (E)	I-C p-value	E-C p-value	E-I p-value
Initial allocation (tables)	36.167	36.250	38.625	0.976	0.391	0.407
	(14.298)	(12.630)	(14.096)			
Discount factor δ	1.652	1.593	1.828	0.808	0.491	0.368
	(1.230)	(1.134)	(1.295)			
Female	0.500	0.500	0.450	1.000	0.636	0.645
	(0.505)	(0.505)	(0.504)			
Age (years)	36.278	34.896	40.125	0.460	0.052	0.018
	(8.187)	(8.382)	(11.811)			
At most high school	0.204	0.167	0.175	0.633	0.725	0.919
	(0.407)	(0.377)	(0.385)			
Some college (at most 2 years)	0.352	0.438	0.375	0.381	0.822	0.558
	(0.482)	(0.501)	(0.490)			
At least 4-year college	0.444	0.396	0.450	0.625	0.958	0.613
	(0.502)	(0.494)	(0.504)			
Reminder in week 2	0.389	0.312	0.375	0.427	0.891	0.543
	(0.492)	(0.468)	(0.490)			
Observations	54	48	40			

Note: Standard deviations are reported in parentheses.

Table 4.A4: Initial vs. subsequent budget shares allocated to week 2

	KS test p-value	Wilcoxon test p-value	Paired two-sided t-test Mean difference	Paired two-sided t-test p-value
Control	0.755	0.479	0.023	0.519
Implicit shift	0.847	0.698	0.012	0.743
Explicit shift	0.913	0.444	0.023	0.607

Figure 4.A3: Distribution of present-bias parameter β by treatment status

(a) Control (b) Implicit shift (c) Explicit shift

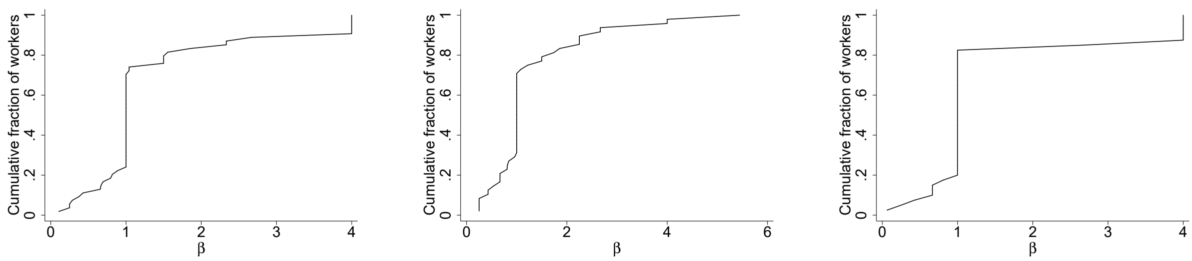


Table 4.A5: Comparison of behavioral measures across treatment groups

	Budget share in week 2			Budget share difference			Relative change in work plan			Present bias β		
	I-C	E-C	E-I	I-C	E-C	E-I	I-C	E-C	E-I	I-C	E-C	E-I
KS test	0.946	0.982	0.579	1.000	0.888	0.928	0.998	0.888	0.928	1.000	0.888	0.928
MW test	0.827	0.666	0.563	0.681	0.750	0.839	0.665	0.718	0.817	0.635	0.762	0.810
t-test	0.857	0.512	0.398	0.827	0.992	0.852	0.904	0.943	0.959	0.753	0.852	0.646
Mean difference	-0.010	0.040	0.050	-0.011	-0.001	0.011	-0.008	-0.004	0.003	-0.065	0.043	0.108
	β^* $(1_{\beta<1})$			$\tilde{\beta}$ $(1_{\beta=1})$								
Proportion test	0.565	0.761	0.304	0.647	0.157	0.049						

Note: Two-sided test p-values unless anything else is specified. 'C', 'I', 'E' denote the control, implicit shift and explicit shift condition, respectively.

Table 4.A6: Estimates with median age interaction terms

	Budget share in week 2		Budget share difference		Relative change in work plan		Present bias β		β^* $(1_{\beta<1})$		$\tilde{\beta}$ $(1_{\beta=1})$	
	no x	x	no x	x	no x	x	no x	x	no x	x	no x	x
Implicit shift (I)	-0.009	-0.000	0.002	-0.021	0.036	0.011	0.104	0.054	0.014	0.038	-0.167	-0.166
st. error	0.087	0.090	0.071	0.077	0.089	0.097	0.287	0.314	0.112	0.114	0.127	0.128
p-value	0.917	0.999	0.983	0.782	0.688	0.914	0.718	0.863	0.900	0.742	0.188	0.197
Explicit shift (E)	0.127	0.128	0.061	0.029	0.087	0.053	0.384	0.302	-0.031	0.018	-0.061	-0.097
st. error	0.108	0.107	0.086	0.096	0.094	0.106	0.400	0.440	0.136	0.142	0.156	0.156
p-value	0.244	0.236	0.475	0.761	0.359	0.616	0.338	0.494	0.820	0.901	0.693	0.536
At least 36 x I	-0.039	-0.050	-0.034	-0.005	-0.099	-0.070	-0.370	-0.319	0.172	0.127	0.156	0.166
st. error	0.108	0.112	0.101	0.111	0.123	0.134	0.412	0.439	0.198	0.196	0.176	0.172
p-value	0.722	0.655	0.734	0.963	0.423	0.603	0.371	0.468	0.387	0.516	0.374	0.333
At least 36 x E	-0.135	-0.139	-0.104	-0.052	-0.156	-0.102	-0.573	-0.447	-0.019	-0.087	0.364	0.401
st. error	0.131	0.127	0.116	0.125	0.127	0.138	0.501	0.553	0.179	0.161	0.154	0.146
p-value	0.305	0.274	0.373	0.681	0.223	0.459	0.255	0.420	0.917	0.588	0.018	0.006
At least 36	-0.038	-0.028	0.030	-0.005	0.073	0.036	0.227	0.150	-0.038	0.005	-0.184	-0.194
st. error	0.085	0.085	0.074	0.083	0.087	0.093	0.296	0.314	0.114	0.116	0.116	0.121
p-value	0.659	0.746	0.689	0.949	0.402	0.699	0.445	0.633	0.739	0.968	0.113	0.108
Constant	0.674	0.679	0.034	0.083	0.042	0.072	1.277	1.313				
st. error	0.070	0.087	0.052	0.087	0.062	0.100	0.218	0.380				
p-value	0.000	0.000	0.512	0.340	0.495	0.471	0.204	0.409				
Reference: <36 years old, implicit shift												
Explicit shift	0.136	0.128	0.060	0.051	0.051	0.043	0.281	0.247	-0.044	-0.019	0.105	0.066
st. error	0.100	0.103	0.088	0.088	0.101	0.102	0.405	0.416	0.126	0.125	0.157	0.155
p-value	0.178	0.217	0.497	0.567	0.615	0.675	0.490	0.553	0.724	0.877	0.503	0.671
At least 36 x E	-0.097	-0.089	-0.069	-0.046	-0.057	-0.033	-0.203	-0.128	-0.147	-0.173	0.227	0.267
st. error	0.119	0.119	0.112	0.109	0.127	0.124	0.495	0.499	0.133	0.124	0.207	0.200
p-value	0.419	0.457	0.536	0.671	0.656	0.792	0.683	0.798	0.269	0.162	0.274	0.182
At least 36	-0.076	-0.078	-0.005	-0.010	-0.026	-0.034	-0.143	-0.169	0.113	0.119	-0.029	-0.029
st. error	0.066	0.068	0.070	0.070	0.088	0.088	0.294	0.293	0.114	0.114	0.148	0.143
p-value	0.249	0.255	0.946	0.882	0.771	0.704	0.626	0.565	0.324	0.299	0.847	0.840

Note: Sample includes 142 observations with non-duplicated IP addresses. Effects are estimated using OLS, but for β^* and $\tilde{\beta}$ estimates are average marginal probit effects. Estimated coefficients for β under control are tested against one. 'x' denotes demographic control variables: female and education dummies; all regressions control for the week 2 reminder. Standard errors are heteroscedasticity robust.

Table 4.A7: Estimates with 2-year college interaction terms

	Budget share in week 2		Budget share difference		Relative change in work plan		Present bias β		β^* $(1_{\beta<1})$		$\tilde{\beta}$ $(1_{\beta=1})$	
	no x	x	no x	x	no x	x	no x	x	no x	x	no x	x
Implicit shift (I)	0.022	0.021	-0.032	-0.033	-0.045	-0.046	-0.281	-0.285	0.150	0.150	-0.113	-0.116
st. error	0.073	0.072	0.068	0.067	0.088	0.087	0.270	0.266	0.132	0.128	0.147	0.145
p-value	0.762	0.770	0.636	0.621	0.612	0.602	0.299	0.285	0.254	0.242	0.444	0.423
Explicit shift (E)	0.109	0.110	0.059	0.060	0.037	0.038	0.260	0.265	-0.078	-0.080	0.133	0.136
st. error	0.105	0.105	0.086	0.087	0.108	0.108	0.425	0.427	0.136	0.135	0.157	0.156
p-value	0.302	0.298	0.496	0.488	0.731	0.722	0.542	0.536	0.569	0.552	0.398	0.385
2-year college x I	-0.064	-0.062	0.028	0.030	0.053	0.055	0.379	0.389	-0.105	-0.107	0.071	0.076
st. error	0.111	0.112	0.106	0.106	0.131	0.132	0.425	0.425	0.129	0.127	0.187	0.185
p-value	0.567	0.583	0.793	0.777	0.687	0.674	0.375	0.362	0.412	0.399	0.704	0.682
2-year college x E	-0.121	-0.126	-0.105	-0.111	-0.075	-0.080	-0.379	-0.402	0.056	0.070	0.060	0.047
st. error	0.132	0.132	0.119	0.120	0.134	0.133	0.511	0.504	0.202	0.204	0.209	0.209
p-value	0.364	0.341	0.379	0.356	0.579	0.547	0.459	0.426	0.782	0.731	0.775	0.822
2-year college	0.082	0.086	-0.013	-0.009	-0.039	-0.035	-0.059	-0.040	0.075	0.065	-0.061	-0.054
st. error	0.082	0.081	0.077	0.077	0.093	0.092	0.315	0.308	0.111	0.112	0.131	0.134
p-value	0.320	0.287	0.865	0.908	0.675	0.706	0.852	0.896	0.500	0.560	0.639	0.688
Constant	0.606	0.622	0.060	0.076	0.107	0.124	1.442	1.511				
st. error	0.058	0.067	0.051	0.059	0.072	0.081	0.225	0.258				
p-value	0.000	0.000	0.244	0.199	0.140	0.131	0.050	0.048				
Reference: no 2-year college, implicit shift												
Explicit shift	0.087	0.089	0.091	0.094	0.082	0.084	0.541	0.550	-0.194	-0.196	0.242	0.248
st. error	0.098	0.098	0.077	0.077	0.091	0.091	0.361	0.367	0.103	0.100	0.150	0.145
p-value	0.378	0.367	0.236	0.229	0.369	0.360	0.136	0.136	0.059	0.050	0.107	0.087
2-year college x E	-0.057	-0.065	-0.133	-0.141	-0.128	-0.136	-0.758	-0.791	0.188	0.206	-0.013	-0.030
st. error	0.127	0.128	0.115	0.115	0.133	0.132	0.490	0.487	0.215	0.214	0.213	0.207
p-value	0.655	0.614	0.249	0.224	0.337	0.305	0.124	0.106	0.381	0.336	0.952	0.886
2-year college	0.018	0.025	0.015	0.021	0.014	0.021	0.320	0.348	-0.041	-0.053	0.010	0.023
st. error	0.074	0.075	0.072	0.073	0.092	0.092	0.284	0.286	0.114	0.113	0.141	0.137
p-value	0.807	0.740	0.839	0.772	0.880	0.823	0.262	0.226	0.717	0.637	0.941	0.865

Note: Sample includes 142 observations with non-duplicated IP addresses. Effects are estimated using OLS, but for β^* and $\tilde{\beta}$ estimates are average marginal probit effects. Estimated coefficients for β under control are tested against one. 'x' denotes demographic control variables: female dummy and age in years; all regressions control for the week 2 reminder. Standard errors are heteroscedasticity robust.

Table 4.A8: Estimates using tobit/logistic regression

	Budget share in week 2		Budget shares difference		Relative change in work plan		Present bias β		β^* $(1_{\beta<1})$		$\tilde{\beta}$ $(1_{\beta=1})$	
	no x	x	no x	x	no x	x	no x	x	no x	x	no x	x
Implicit shift	-0.024	-0.029	-0.009	-0.012	-0.008	-0.013	-0.020	-0.027	0.078	0.078	-0.071	-0.057
st. error	0.059	0.060	0.027	0.027	0.035	0.035	0.055	0.057	0.085	0.085	0.097	0.096
p-value	0.683	0.634	0.739	0.647	0.815	0.713	0.715	0.628	0.360	0.360	0.464	0.550
Explicit shift	0.050	0.068	-0.002	-0.002	-0.005	-0.005	0.008	0.009	-0.049	-0.040	0.168	0.161
st. error	0.067	0.067	0.031	0.031	0.035	0.035	0.061	0.062	0.086	0.087	0.102	0.105
p-value	0.456	0.311	0.956	0.961	0.895	0.888	0.902	0.880	0.570	0.647	0.100	0.126
Reminder	-0.066	-0.071	-0.039	-0.042	-0.050	-0.054	-0.033	-0.038	0.134	0.133	-0.096	-0.082
st. error	0.056	0.058	0.026	0.026	0.031	0.031	0.051	0.052	0.076	0.078	0.085	0.085
p-value	0.243	0.221	0.129	0.102	0.100	0.080	0.520	0.462	0.080	0.086	0.260	0.334
Reference: implicit shift												
Explicit shift	0.074	0.096	0.007	0.011	0.004	0.008	0.028	0.037	-0.115	-0.108	0.236	0.216
st. error	0.065	0.068	0.030	0.029	0.036	0.036	0.063	0.066	0.077	0.080	0.099	0.103
p-value	0.257	0.156	0.809	0.715	0.921	0.824	0.659	0.576	0.133	0.180	0.017	0.036

Note: Sample includes 142 observations with non-duplicated IP addresses. Average marginal tobit effects are on actual variables. Estimates on β^* and $\tilde{\beta}$ are average marginal logistic effects. 'x' denotes demographic control variables. Standard errors are heteroscedasticity robust.

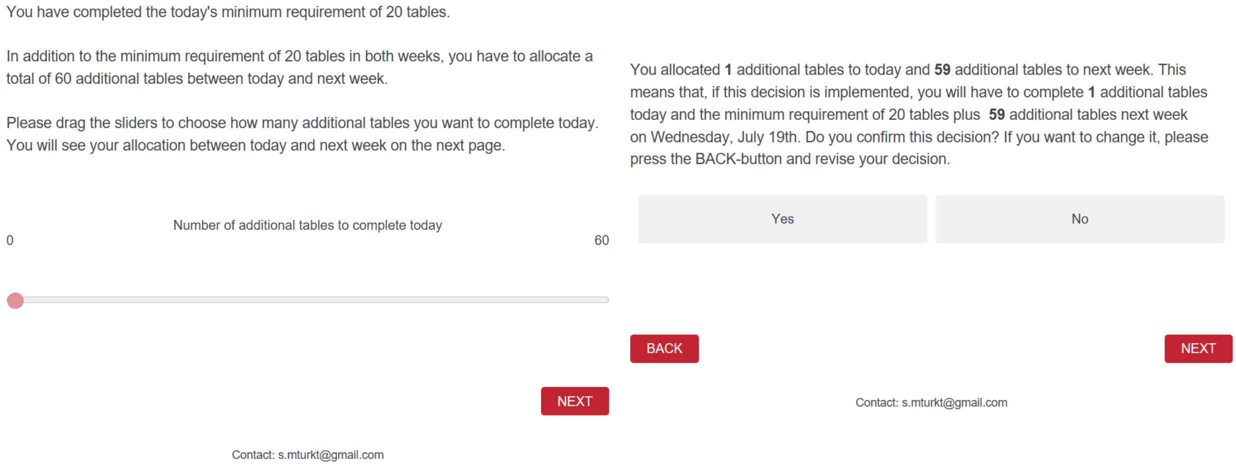
Table 4.A9: Estimates with female interaction terms using sample duplications

	Budget share in week 2		Budget share difference		Relative change in work plan		Present bias β		β^* $(1_{\beta<1})$		$\tilde{\beta}$ $(1_{\beta=1})$	
	no x	x	no x	x	no x	x	no x	x	no x	x	no x	x
Implicit shift (I)	-0.069	-0.065	-0.062	-0.062	-0.101	-0.101	-0.501	-0.496	0.030	0.035	0.184	0.180
st. error	0.026	0.025	0.023	0.023	0.028	0.028	0.090	0.090	0.037	0.037	0.033	0.033
p-value	0.007	0.010	0.008	0.007	0.000	0.000	0.000	0.000	0.424	0.339	0.000	0.000
Explicit shift (E)	0.011	0.025	-0.028	-0.029	-0.059	-0.062	-0.263	-0.262	-0.005	-0.004	0.217	0.219
st. error	0.028	0.028	0.025	0.026	0.029	0.030	0.108	0.111	0.038	0.038	0.040	0.041
p-value	0.697	0.364	0.271	0.257	0.042	0.037	0.015	0.018	0.891	0.921	0.000	0.000
Female x I	0.109	0.096	0.090	0.080	0.175	0.160	0.858	0.809	0.089	0.086	-0.487	-0.473
st. error	0.034	0.035	0.032	0.032	0.039	0.039	0.127	0.129	0.057	0.057	0.032	0.035
p-value	0.001	0.007	0.005	0.014	0.000	0.000	0.000	0.000	0.122	0.131	0.000	0.000
Female x E	0.054	0.057	0.050	0.056	0.104	0.111	0.610	0.627	-0.074	-0.071	-0.112	-0.114
st. error	0.039	0.038	0.037	0.036	0.040	0.039	0.147	0.146	0.048	0.048	0.058	0.058
p-value	0.160	0.133	0.179	0.120	0.009	0.005	0.000	0.000	0.126	0.138	0.054	0.050
Female	-0.086	-0.080	-0.081	-0.077	-0.127	-0.121	-0.608	-0.585	0.063	0.063	0.143	0.129
st. error	0.025	0.026	0.023	0.023	0.027	0.027	0.088	0.090	0.036	0.035	0.035	0.036
p-value	0.001	0.002	0.000	0.001	0.000	0.000	0.000	0.000	0.079	0.077	0.000	0.000
Constant	0.693	0.838	0.091	0.123	0.146	0.155	1.700	1.781				
st. error	0.020	0.033	0.020	0.033	0.024	0.038	0.076	0.144				
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				
Reference: men, implicit shift												
Explicit shift	0.080	0.091	0.034	0.033	0.043	0.039	0.238	0.234	-0.034	-0.037	0.017	0.024
st. error	0.025	0.025	0.019	0.020	0.022	0.023	0.089	0.092	0.038	0.037	0.045	0.045
p-value	0.002	0.000	0.081	0.094	0.054	0.081	0.007	0.011	0.365	0.309	0.704	0.591
Female x E	-0.055	-0.039	-0.041	-0.024	-0.071	-0.049	-0.247	-0.182	-0.137	-0.134	0.383	0.365
st. error	0.037	0.036	0.037	0.035	0.041	0.040	0.148	0.148	0.039	0.040	0.039	0.043
p-value	0.141	0.279	0.266	0.494	0.086	0.223	0.096	0.219	0.000	0.001	0.000	0.000
Female	0.023	0.015	0.009	0.003	0.047	0.039	0.250	0.224	0.146	0.143	-0.364	-0.357
st. error	0.023	0.023	0.023	0.022	0.029	0.029	0.092	0.092	0.036	0.036	0.030	0.030
p-value	0.327	0.512	0.686	0.891	0.104	0.175	0.007	0.015	0.000	0.000	0.000	0.000

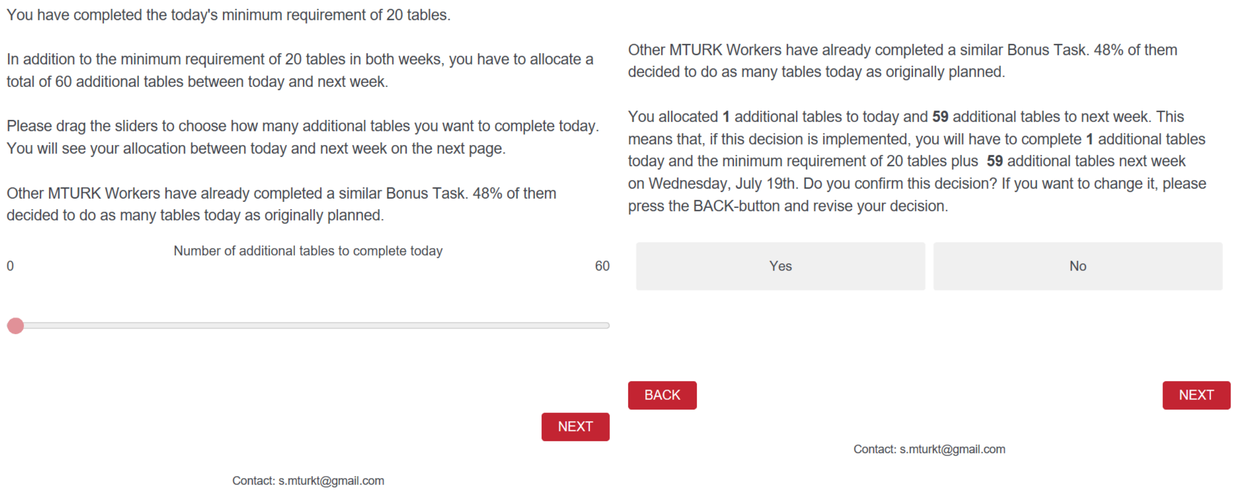
Note: Sample of 142 observations with non-duplicated IP addresses is duplicated ten times. Effects are estimated using OLS, but for β^* and $\tilde{\beta}$ estimates are average marginal probit effects. Estimated coefficients for β under control are tested against one. 'x' denotes demographic control variables: age in years and education dummies; all regressions control for the week 2 reminder. Standard errors are heteroscedasticity robust.

Figure 4.A4: Control and social comparison conditions

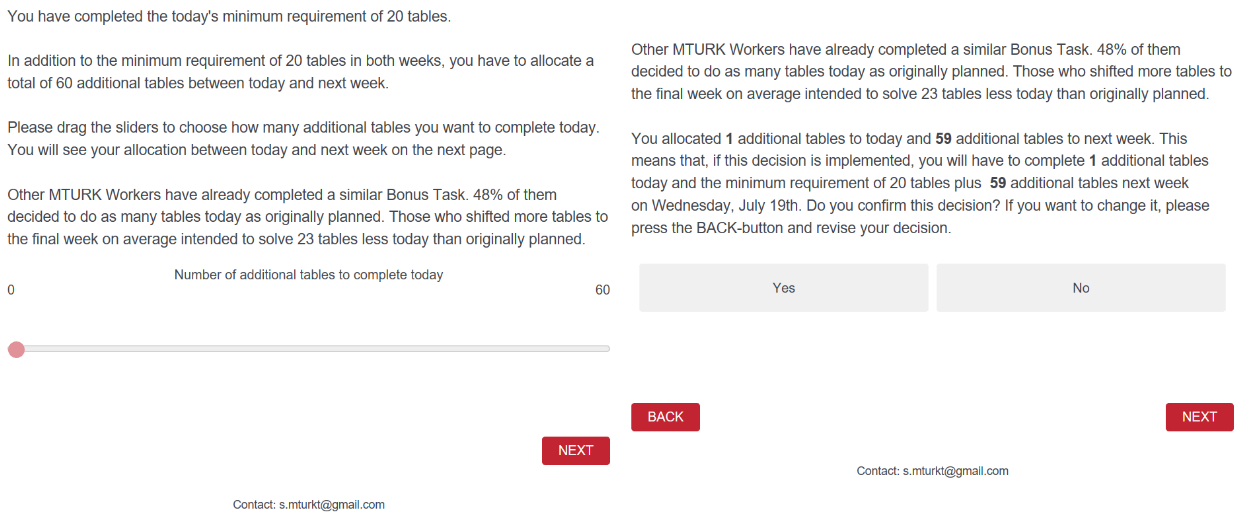
(a) Control



(b) Implicit shift



(c) Explicit shift



References

Aakvik, A., K. G. Salvanes, and K. Vaage (2010). Measuring heterogeneity in the returns to education using an education reform. *European Economic Review 54*(4), 483–500.

Aaronson, D., R. Dehejia, A. Jordan, C. Pop-Eleches, C. Samii, and K. Schulze (2017). The effect of fertility on mothers' labor supply over the last two centuries. *IZA Discussion Paper 10559.*

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*(1), 1–19.

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011). Reference points and effort provision. *American Economic Review 101*(2), 470–492.

Agüero, J. M. and M. S. Marks (2008). Motherhood and female labor force participation: Evidence from infertility shocks. *American Economic Review 98*(2), 500–504.

Agüero, J. M. and M. S. Marks (2011). Motherhood and female labor supply in the developing world: Evidence from infertility shocks. *Journal of Human Resources 46*(4), 800–826.

Ahrens, A., C. B. Hansen, and M. E. Schaffer (2018). PDSLASSO: Stata module for post-selection and post-regularization OLS or IV estimation and inference.

Albert, J. M. and S. Nelson (2011). Generalized causal mediation analysis. *Biometrics 67*(3), 1028–1038.

Amin, V. and J. Behrman (2014). Do more-schooled women have fewer children and delay childbearing? Evidence from a sample of US twins. *Journal of Population Economics 27*(1), 1–31.

Andersen, S., A. Di Girolamo, G. W. Harrison, and M. I. Lau (2014). Risk and time preferences of entrepreneurs: Evidence from a Danish field experiment. *Theory and Decision 77*(3), 341–357.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2008). Eliciting risk and time preferences. *Econometrica 76*(3), 583–618.

Andreoni, J., M. A. Kuhn, and C. D. Sprenger (2015). Measuring time preferences: A comparison of experimental methods. *Journal of Economic Behavior and Organization 116*, 451–464.

Andreoni, J. and C. D. Sprenger (2012a). Estimating time preferences from convex budgets. *American Economic Review 102*(7), 3333–3356.

Andreoni, J. and C. D. Sprenger (2012b). Risk preferences are not time preferences. *American Economic Review 102*(7), 3357–3376.

Angrist, J. D. and W. Evans (1998). Children and their parents' labor supply: Evidence from exogeneous variation in family size. *American Economic Review 88*(3), 450–477.

Angrist, J. D. and W. N. Evans (1996). Schooling and labor market consequences of the 1970 state abortion reforms. Technical report, National Bureau of Economic Research.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 445–455.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics 106*(4), 979–1014.

Angrist, J. D., V. Lavy, and A. Schlosser (2010). Multiple experiments for the causal link between the quantity and quality of children. *Journal of Labor Economics 28*(4), 773–824.

Arel, D. (2006). La face cachée de la Révolution orange: l'Ukraine en négation face à son problème régional. *Revue d'études comparatives Est-Ouest 37*(4), 11–48.

Ariely, D. and K. Wertenbroch (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science 13*(3), 219–224.

Aspachs-Bracons, O., I. Clots-Figueras, J. Costa-Font, and P. Masella (2008). Compulsory language educational policies and identity formation. *Journal of the European Economic Association 6*(2-3), 434–444.

Augenblick, N., M. Niederle, and C. D. Sprenger (2015). Working over time: Dynamic inconsistency in real effort tasks. *Quarterly Journal of Economics 130*(3), 1067–1115.

Augenblick, N. and M. Rabin (2019). An experiment on time preference and misprediction in unpleasant tasks. *Review of Economic Studies 86*(3), 941–975.

Backes-Gellner, U., H. Herz, M. Kosfeld, and Y. Oswald (2018). Do preferences and biases predict life outcomes? Evidence from education and labor market entry decisions. Technical report, Centre for Economic Policy Research.

Balakrishnan, U., J. Haushofer, and P. Jakiela (2017). How soon is now? Evidence of present bias from convex time budget experiments. Technical report, National Bureau of Economic Research.

Balcells, L. and G. Torrats-Espinosa (2018). Using a natural experiment to estimate the electoral consequences of terrorist attacks. *PNAS 115*(42), 10624–10629.

Banks, J. and F. Mazzonna (2012). The effect of education on old age cognitive abilities: Evidence from a regression discontinuity design. *Economic Journal 122*, 418–448.

Baron, R. and D. Kenny (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology 51*(6), 1173–1182.

Barrington, L. and E. Herron (2004). One Ukraine or many? Regionalism in Ukraine and its political consequences. *Nationalities Papers 32*(1), 53–86.

Bateson, R. (2012). Crime victimization and political participation. *American Political Science Review 106*(3), 570–587.

Battaglini, M., R. Bénabou, and J. Tirole (2005). Self-control in peer groups. *Journal of Economic Theory 123*(2), 105–134.

Battaglini, M., C. Díaz, and E. Patacchini (2017). Self-control and peer groups: An empirical analysis. *Journal of Economic Behavior and Organization 134*, 240–254.

Becker, G. S. (1965). A theory of the allocation of time. *Economic Journal 75*(299), 493–517.

Becker, G. S. and H. G. Lewis (1973). On the interaction between the quantity and quality of children. *Journal of Political Economy 81*(2), S279–S288.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Bellows, J. and E. Miguel (2009). War and local collective action in Sierra Leone. *Journal of*

Public Economics 93, 1144–1157.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis 20*(3), 351–368.

Beshears, J., J. J. Choi, D. Laibson, B. C. Madrian, and K. L. Milkman (2015). The effect of providing peer information on retirement savings decisions. *Journal of Finance 70*(3), 1161–1201.

Beugnot, J., B. Fortin, G. Lacroix, and M. C. Villeval (2019). Gender and peer effects on performance in social networks. *European Economic Review 113*, 207–224.

Bhalotra, S. and D. Clarke (2016). The twin instrument. *IZA Discussion Paper 10405*.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2008). Staying in the classroom and out of the maternity ward? The effect of compulsory schooling laws on teenage births. *Economic Journal 118*(530), 1025–1054.

Blattman, C. (2009). From violence to voting: War and political participation in Uganda. *American Political Science Review 103*(2), 231–247.

Bloom, D., D. Canning, G. Fink, and J. Finlay (2009). Fertility, female labor force participation, and the demographic dividend. *Journal of Economic Growth 14*, 79–101.

Blundell, R., A. Bozio, and G. Laroque (2013). Extensive and intensive margins of labour supply: Work and working hours in the US, the UK and France. *Fiscal Studies 34*(1), 1–29.

Blundell, R., L. Dearden, and B. Sianesi (2005). Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*(3), 473–512.

Boas, F. (1940). *Race, Language, and Culture*. University of Chicago Press.

Bodory, H. and M. Huber (2018). causalweight: Estimation methods for causal inference based on inverse probability weighting.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology 43*(1), 1–22.

Braakmann, N. (2011). Female education and fertility — Evidence from changes in British compulsory schooling laws. Technical report.

Bronars, S. and J. Grogger (1994). The economic consequences of unwed motherhood: Using twin births as a natural experiment. *American Economic Review 84*(5), 1141–1156.

Brown, M., C. Henchoz, and T. Spycher (2018). Culture and financial literacy: Evidence from a within-country language border. *Journal of Economic Behavior and Organization 150*, 62–85.

Brunello, G., D. Fabbri, and M. Fort (2013). The causal effect of education on body mass: Evidence from Europe. *Journal of Labor Economics 31*(1), 195–223.

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

Brunello, G., M. Fort, and G. Weber (2009). Changes in compulsory schooling, education and the distribution of wages in Europe. *Economic Journal 119*(536), 516–539.

Brunello, G., M. Fort, G. Weber, and C. T. Weiss (2013). Testing the internal validity of compulsory school reforms as instrument for years of schooling. *IZA Discussion Paper No. 7533*.

Brunello, G., G. Weber, and C. T. Weiss (2017). Books are forever: Early life conditions, education and lifetime earnings in Europe. *Economic Journal 127*, 271–296.

Büchi, C. (2001). *"Röstigraben": das Verhältnis zwischen deutscher und französischer Schweiz: Geschichte und Perspektiven.* Verlag Neue Zürcher Zeitung.

Buechel, B., L. Mechtenberg, and J. Petersen (2018). If I can do it, so can you! Peer effects on perseverance. *Journal of Economic Behavior and Organization 155*, 301–314.

Buhrmester, M., T. Kwang, and S. D. Goslin (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science 6*(1), 3–5.

Butcher, K. F. and A. Case (1994). The effect of sibling sex composition on women's education and earnings. *Quarterly Journal of Economics 109*(3), 531–563.

Bütikofer, A. (2011). Sibling sex composition and cost of children. Technical report.

Cantoni, D. (2015). The economic effects of the Protestant Reformation: Testing the Weber hypothesis in the German lands. *Journal of the European Economic Association 13*(4), 561–598.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160.

Case, A. and C. Paxson (2009). Early life health and cognitive function in old age. *American Economic Review 99*(2), 104–109.

Chabris, C. F., D. Laibson, C. L. Morris, J. P. Schuldt, and D. Taubinsky (2008). Individual laboratory-measured discount rates predict field behavior. *Journal of Risk and Uncertainty 37*, 237–269.

Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review 103*(2), 690–731.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics 7*, 649–688.

Christen, E., O. Fritz, and G. Streicher (2015). Effects of the EU–Russia economic sanctions on value added and employment in the European Union and Switzerland. Technical report, Austrian Institute of Economic Research (WIFO).

Chun, H. and J. Oh (2002). An instrumental variable estimate of the effect of fertility on the labour force participation of married women. *Applied Economics Letters 9*(10), 631–634.

Clist, P. and A. Verschoor (2017). Multilingualism and public goods provision: An experiment in two languages in Uganda. *Journal of Development Economics 129*, 47–57.

Clots-Figueras, I. and P. Masella (2013). Education, language and identity. *Economic Journal 123*(570), F332–F357.

Coffman, L. C., C. R. Featherstone, and J. B. Kessler (2017). Can social information affect what job you choose and keep? *American Economic Journal: Applied Economics 9*(1), 96–117.

Costa, A., A. Foucart, I. Arnon, M. Aparici, and J. Apesteguia (2014). "Piensa" twice: On the foreign language effect in decision making. *Cognition 130*(2), 236–254.

Costa, A., A. Foucart, S. Hayakawa, M. Aparici, J. Apesteguia, J. Heafner, and B. Keysar (2014). Your morals depend on language. *PloS ONE 9*(4), e94842.

Coupé, T. and M. Obrizan (2016a). The impact of war on happiness: The case of Ukraine. *Journal of Economic Behavior and Organization 132*, 228–242.

Coupé, T. and M. Obrizan (2016b). Violence and political outcomes in Ukraine — Evidence from Sloviansk and Kramatorsk. *Journal of Comparative Economics 44*(1), 201–212.

Cristia, J. P. (2008). The effect of a first child on female labor supply: Evidence from women seeking fertility services. *Journal of Human Resources 43*(3), 487–510.

Cruces, G. and S. Galiani (2007). Fertility and female labor supply in Latin America: New causal evidence. *Labour Economics 14*, 565–573.

Cunha, F. and J. Heckman (2007). The technology of skill formation. *American Economic Review 97*(2), 31–47.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature 47*(1), 87–122.

Cygan-Rehm, K. and M. Maeder (2013). The effect of education on fertility: Evidence from a compulsory schooling reform. *Labour Economics 25*, 35–48.

de Quidt, J., J. Haushofer, and C. Roth (2018). Measuring and bounding experimenter demand. *American Economic Review 108*(11), 3266–3302.

DeCicca, P. and H. Krashinsky (2016). The effect of education on overall fertility. Technical report, National Bureau of Economic Research.

DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014). Cross-border media and nationalism: Evidence from Serbian Radio in Croatia. *American Economic Journal: Applied Economics 6*(3), 103–132.

DellaVigna, S. and E. Kaplan (2007). The Fox News effect: Media bias and voting. *Quarterly Journal of Economics 122*(3), 1187–1234.

DellaVigna, S. and U. Malmendier (2006). Paying not to go to the gym. *American Economic Review 96*(3), 941–975.

DellaVigna, S. and M. D. Paserman (2005). Job search and impatience. *Journal of Labor Economics 23*(3), 527–588.

DellaVigna, S. and D. Pope (2018). What motivates effort? Evidence and expert forecasts. *Review of Economic Studies 85*(2), 1029–1069.

Devereux, P. J. and R. A. Hart (2010). Forced to be rich? Returns to compulsory schooling in Britain. *Economic Journal 120*(549), 1345–1364.

Doepke, M. and F. Zilibotti (2008). Occupational choice and the spirit of capitalism. *Quarterly Journal of Economics 123*(2), 747–793.

Doepke, M. and F. Zilibotti (2014). Culture, entrepreneurship, and growth. In *Handbook of Economic Growth*, Volume 2, pp. 1–48. Elsevier.

Doepke, M., F. Zilibotti, et al. (2005). Social class and the spirit of capitalism. *Journal of the European Economic Association 3*(2-3), 516–524.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review 100*(3), 1238–1260.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2017). The robustness and pervasiveness of sub-additivity in intertemporal choice. Technical report.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and political persuasion: Evidence from Russia. *American Economic Review 101*(7), 3253–3285.

Erhardt, K. and S. Haenni (2018). Born to be an entrepreneur? How cultural origin affects entrepreneurship. Technical Report 309, University of Zurich, Department of Economics.

Erikson, R. and L. Stoker (2011). Caught in the draft: The effects of Vietnam draft lottery status on political attitudes. *American Political Science Review 105*(2), 221–237.

Eugster, B., R. Lalive, A. Steinhauer, and J. Zweimüller (2011). The demand for social insurance: Does culture matter? *Economic Journal 121*(556), F413–F448.

Eugster, B., R. Lalive, A. Steinhauer, and J. Zweimüller (2017). Culture, work attitudes, and job search: Evidence from the Swiss language border. *Journal of the European Economic Association 15*(5), 1056–1100.

European Commission (2018). The 2018 ageing report: Economic & budgetary projections for the 28 EU member states (2016-2070). *European Economy, Institutional Paper No 79*.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global evidence on economic preferences. *Quarterly Journal of Economics 133*(4), 1645–1692.

Falk, A. and A. Ichino (2006). Clean evidence on peer effects. *Journal of Labor Economics 24*(1), 39–57.

Falk, A. and M. Knell (2004). Choosing the joneses: Endogenous goals and reference standards. *Scandinavian Journal of Economics 106*(3), 417–435.

Fedyk, A. (2019). Asymmetric naïveté: Beliefs about self-control. Technical report.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations 7*(2), 117–140.

Fort, M., N. Schneeweis, and R. Winter-Ebmer (2016). Is education always reducing fertility? Evidence from compulsory schooling reforms. *Economic Journal 126*, 1823–1855.

Frederick, S., G. Loewenstein, and T. O'Donoghue (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature 40*(2), 351–401.

Frey, B. S. and S. Meier (2004). Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review 94*(5), 1717–1722.

Fricke, H. (2017). Identification based on difference-in-differences approaches with multiple treatments. *Oxford Bulletin of Economics and Statistics 79*(3), 426–433.

Frölich, M. and M. Huber (2017). Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(5), 1645–1666.

Garcia-Ponce, O. and B. Pasquale (2015). How political repression shapes attitudes toward the state: Evidence from Zimbabwe. Technical report, available at: `http://omargarciaponce.com/wp-content/uploads/2013/07/GarciaPonce_and_Pasquale_2014.pdf`, retrieved 6 March 2017.

Gathmann, C., H. Jürges, and S. Reinhold (2015). Compulsory schooling reforms, education and mortality in twentieth century Europe. *Social Science & Medicine 127*, 74–82.

Georganas, S., M. Tonin, and M. Vlassopoulos (2015). Peer pressure and productivity: The role of observing and being observed. *Journal of Economic Behavior and Organization 117*, 223–232.

Gerber, A. S., D. Karlan, and D. Bergan (2009). Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics 1*(2), 35–52.

Gerhards, L. and C. Gravert (2016). Because of you I did not give up — How peers affect perseverance. Technical report.

Geruso, M. and H. Royer (2014). The impact of education on family formation: Quasi-experimental evidence from the UK. Technical report.

Güneş, P. M. (2015). The impact of female education on teenage fertility: Evidence from Turkey. *B.E. Journal of Economic Analysis & Policy 16*(1), 259–288.

Goldstein, N. J., R. B. Cialdini, and V. Griskevicius (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research 35*(3), 472–482.

Golsteyn, B. H., H. Grönqvist, and L. Lindahl (2013). Adolescent time preferences predict lifetime outcomes. *Economic Journal 124*, F739–F761.

Goodman, J. K., C. E. Cryder, and A. Cheema (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making 26*(3), 213–224.

Grenet, J. (2013). Is extending compulsory schooling alone enough to raise earnings? Evidence from French and British compulsory schooling laws. *Scandinavian Journal of Economics 115*(1), 176–210.

Grönqvist, H. and C. Hall (2013). Education policy and early fertility: Lessons from an expansion of upper secondary schooling. *Economics of Education Review 37*, 13–33.

Guin, B. (2017). Culture and household saving. Technical report, European Central Bank.

Hafeman, D. M. (2011). Confounding of indirect effects: A sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. *American Journal of Epidemiology 174*(6), 710–717.

Halevy, Y. (2014). Some comments on the use of monetary and primary rewards in the measurement of time preferences. Technical report.

Hallsworth, M., J. A. List, R. D. Metcalfe, and I. Vlaev (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics 148*, 14–31.

Harmon, C. and I. Walker (1995). Estimates of the economic returns to schooling for the United Kingdom. *American Economic Review 85*(5), 1278–1286.

Heaton, T. B. (1990). Marital stability throughout the child-rearing years. *Demography 27*(1), 55–63.

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hoekstra, C., Z. Z. Zhao, C. Lambalk, G. Willemsen, N. Martin, D. Boomsma, and G. Montgomery (2008). Dizygotic twinning. *Human Reproducation Update 14*(1), 37–47.

Hong, G., X. Qin, and F. Yang (2018). Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics 43*(1), 32–56.

Horton, J. J. and L. B. Chilton (2010). The labor economics of paid crowdsourcing. *Proceedings of the 11th ACM conference on Electronic Commerce*, 209–218.

Horton, J. J., D. G. Rand, and R. J. Zeckhauser (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics 14*(3), 399–425.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite population. *Journal of American Statistical Association 47*(260), 663–685.

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

Hurt, L. S., C. Ronsmans, and S. L. Thomas (2006). The effect of number of births on women's mortality: Systematic review of the evidence for women who have completed their childbearing. *Population Studies 60*(1), 55–71.

Iacovou, M. (2001). Fertility and female labour supply. *ISER Working Paper 2001-19*.

Ichino, A. and R. Winter-Ebmer (1999). Lower and upper bounds of returns to schooling: An exercise in IV estimation with different instruments. *European Economic Review 43*(4-6), 889–901.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309–334.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Imai, K. and T. Yamamoto (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis 21*(2), 141–171.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Ipeirotis, P. G. (2010). Analyzing the Amazon Mechanical Turk marketplace. *ACM XRDS 17*(2), 16–21.

Jacobsen, J. P., J. W. Pearce III, and J. L. Rosenbloom (1999). The effect of childbearing on married women's labor supply and earnings: Using twin births as a natural experiment. *Journal of Human Resources 34*(3), 449–474.

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

Kane, T. J. and C. E. Rouse (1995). Labor-market returns to two- and four-year college. *American Economic Review 85*(3), 600–614.

Kast, F., S. Meier, and D. Pomeranz (2018). Saving more in groups: Field experimental evidence from Chile. *Journal of Development Economics 133*, 275–294.

Kaur, S., M. Kremer, and S. Mullainathan (2010). Self-control and the development of work arrangements. *American Economic Review 102*(2), 624–628.

Keysar, B., S. L. Hayakawa, and S. G. An (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science 23*(6), 661–668.

Krupka, E. and R. A. Weber (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology 30*(3), 307–320.

Kuziemko, I., M. I. Norton, E. Saez, and S. Stantcheva (2015). How elastic are preferences for redistribution? Evidence from randomized survey experiments. *American Economic Review 105*(4), 1478–1508.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics 112*(2), 443–478.

Lavy, V. and A. Zablotsky (2015). Women's schooling and fertility under low female labor force participation: Evidence from mobility restrictions in Israel. *Journal of Public Economics 124*, 105–121.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics 4*(3), 165–224.

Leigh, A. and C. Ryan (2008). Estimating returns to education using different natural experiment techniques. *Economics of Education Review 27*(2), 149–160.

Lenney, E. (1977). Women's self-confidence in achievement settings. *Psychological Bulletin 84*(1), 1–13.

León, A. (2004). The effect of education on fertility: Evidence from compulsory schooling laws. Technical report.

Lochner, L. and E. Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review 94*(1), 155–189.

Lundborg, P., E. Plug, and A. W. Rasmussen (2017). Can women have children and a career? IV evidence from IVF treatments. *American Economic Review 107*(6), 1611–37.

Lusardi, A. and O. S. Mitchell (2016). Older women's labor market attachment, retirement planning, and household debt. Technical report, National Bureau of Economic Research.

Lyall, J., G. Blair, and K. Imai (2013). Explaining support for combatants during wartime: A survey experiment in Afghanistan. *American Political Science Review 107*(4), 679–705.

Mas, A. and E. Moretti (2009). Peers at work. *American Economic Review 99*(1), 112–145.

Mason, W. and S. Suri (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods 44*, 1–23.

Mason, W. and D. J. Watts (2009). Financial incentives and the "performance of crowds". *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 77–85.

McCrary, J. and H. Royer (2011). The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. *American Economic Review 101*(4), 158–195.

Meghir, C. and M. Palme (2005). Educational reform, ability, and family background. *American Economic Review 95*(1), 414–424.

Meier, S. and C. D. Sprenger (2010). Present-biased preferences and credit card borrowing. *American Economic Journal: Applied Economics 2*(1), 193–210.

Meier, S. and C. D. Sprenger (2012). Time discounting predicts creditworthiness. *Psychological Science 23*(1), 56–58.

Mobius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2011). Managing self-confidence: Theory and experimental evidence. Technical report, National Bureau of Economic Research.

Monstad, K., C. Propper, and K. G. Salvanes (2008). Education and fertility: Evidence from a natural experiment. *Scandinavian Journal of Economics 110*(4), 827–852.

Montalvo, J. (2011). Voting after the bombing: A natural experiment on the effect of terrorist attacks on democratic elections. *Review of Economics and Statistics 93*(4), 1146–1154.

Nevitte, N., A. Blais, E. Gidengil, and R. Nadeau (2009). Socio-economic status and non-voting. In H.-D. Klingemann (Ed.), *The Comparative Study of Electoral Systems*, pp. 85–108. Oxford: Oxford University Press.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics 122*(3), 1067–1101.

O'Donoghue, T. and M. Rabin (1999). Doing it now or later. *American Economic Review 89*(1), 103–124.

O'Donoghue, T. and M. Rabin (2001). Choice and procrastination. *Quarterly Journal of Economics 116*(1), 121–160.

OECD (2008). Gender and sustainable development: Maximising the economic, social and environmental role of women.

OECD (2015). Recommendation of the Council on Ageing and Employment Policies.

Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review 96*(1), 152–175.

Oreopoulos, P. (2007). Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics 91*, 2213–2229.

Paolacci, G., J. Chandler, and P. G. Ipeirotis (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making 5*(5), 411–419.

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

Phelps, E. and R. Pollak (1968). On the second-best national saving and game-equilibrium growth. *Review of Economic Studies 35*(2), 185–199.

Pischke, J.-S. and T. von Wachter (2008). Zero returns to compulsory schooling in Germany: Evidence and interpretation. *Review of Economics and Statistics 90*(3), 592–598.

Population Reference Bureau (2018). 2018 World Population Data Sheet.

Powdthavee, N., W. Lekfuangfu, and M. Wooden (2015). What's the good of education on our overall quality of life? A simultaneous equation model of education and life satisfaction for Australia. *Journal of Behavioral and Experimental Economics 54*, 10–21.

Powdthavee, N., W. N. Lekfuangfu, and M. Wooden (2013). The marginal income effect of education on happiness: Estimating the direct and indirect effects of compulsory schooling on well-being in Australia. *IZA Discussion Paper No. 7365*.

Rabin, M. and G. Weizsäcker (2009). Narrow bracketing and dominated choices. *American Economic Review 99*(4), 1508–1543.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology 299*, 172–179.

Riddell, W. C. and X. Song (2011). The impact of education on unemployment incidence and re-employment success: Evidence from the U.S. labour market. *Labour Economics 18*(4), 453–463.

Roberts, S. G., J. Winters, and K. Chen (2015). Future tense and economic decisions: Controlling for cultural evolution. *PloS ONE 10*(7), e0132145.

Rohner, D., M. Thoenig, and F. Zilibotti (2013). Seeds of distrust: Conflict in Uganda. *Journal of Economic Growth 18*(3), 217–252.

Rosenzweig, M. and K. Wolpin (1980). Testing the quantity-quality fertility model: The use of twins as a natural experiment. *Econometrica 48*(1), 227–240.

Rosenzweig, M. and K. Wolpin (2000). Natural "natural experiments" in economics. *Journal of Economic Literature 38*, 827–874.

Rozenas, A., S. Schutte, and Y. Zhukov (2017). The political legacy of violence: The long-term impact of Stalin's repression in Ukraine. *Journal of Politics 79*(4), 1147–1161.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Samuelson, P. A. (1937). A note on measurement of utility. *Review of Economic Studies 4*(2), 155–161.

Sapir, E. (1949). *Selected Writings in Language: Culture and Personality*. University of California Press.

Schneeweis, N., V. Skirbekk, and R. Winter-Ebmer (2014). Does education improve cognitive performance four decades after school completion? *Demography 51*, 619–643.

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science 18*(5), 429–434.

Silles, M. A. (2011). The effect of schooling on teenage childbearing: Evidence using changes in compulsory education laws. *Journal of Population Economics 24*(2), 761–777.

Smith, J. P. (2009). The impact of childhood health on adult labor market outcomes. *American Economic Review 91*(3), 478–489.

Strotz, R. H. (1955-1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies 23*(3), 165–180.

Sutter, M., S. Angerer, D. Glätzle-Rützler, and P. Lergetporer (2018). Language group differences in time preferences: Evidence from primary school children in a bilingual city. *European Economic Review 106*, 21–34.

Sutter, M., M. G. Kocher, D. Glätzle-Rützler, and S. T. Trautmann (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review 103*(1), 510–531.

Tchetgen Tchetgen, E. J. and I. Shpitser (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics 40*(3), 1816–1845.

Thieroff, R. (2000). On the areal distribution of tense-aspect categories in Europe. In Östen Dahl

(Ed.), *Tense and Aspect in the Languages of Europe*, pp. 309–328. Berlin: Mouton de Gruyter.

Tice, D. and R. Baumeister (1997). Longitudinal study of procrastination, performance, stress, and health: The costs and benefits of dawdling. *Psychological Science 8*(6), 454–458.

Trostel, P., I. Walkerb, and P. Woolley (2002). Estimates of the economic return to schooling for 28 countries. *Labour Economics 9*(1), 1–16.

van Veldhuizen, R., H. Oosterbeek, and J. Sonnemans (2018). Peers at work: Evidence from the lab. *PLoS ONE 13*(2), e0192038.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology 20*(1), 18–26.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology 21*(4), 540–551.

VanderWeele, T. J. and Y. Chiba (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology Biostatistics and Public Health 11*(2), 1–16.

Voors, M., E. Nillesen, P. Verwimp, E. Bulte, R. Lensink, and D. V. Soest (2012). Violent conflict and behavior: A field experiment in Burundi. *American Economic Review 102*(2), 941–964.

Weber, M. (1930). *The Protestant ethic and the spirit of capitalism.* Routledge.

Whorf, B. L. (1956). *Language, Thought and Reality: Selected Writing of Benjamin Lee Whorf.*

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin 90*(2), 245–271.

Wilson, A. (1995). The Donbas between Ukraine and Russia: The use of history in political disputes. *Journal of Contemporary History 30*(2), 265–289.

Wood, J. V. and K. L. Taylor (1991). Serving self-relevant goals through social comparison. In J. Suls and T. A. Wills (Eds.), *Social comparison: Contemporary theory and research*, pp. 23–49. Lawrence Erlbaum Associates, Inc.

Zajonc, R. B. (1965). Social facilitation. *Science 149*(3681), 269–274.

Zhukov, Y. (2016). Trading hard hats for combat helmets: The economics of rebellion in eastern Ukraine. *Journal of Comparative Economics 44*(1), 1–15.