Essays in Empirical Health Economics

Inaugural-Dissertation
zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2018

vorgelegt von

Nadine Evi Geiger

Referent: Prof. Dr. Joachim Winter
Korreferent: Prof. Thomas Crossley, PhD
Promotionsabschlussberatung: 11. Juli 2018

Tag der mündlichen Prüfung: 05.07.2018

Namen der Berichterstatter: Joachim Winter, Thomas Crossley, Derya Uysal

Acknowledgments

First and foremost, I would like to thank my supervisor, Joachim Winter, who always provided me with a lot of academic freedom and offered invaluable advice during critical moments. Without his guidance and support this thesis would not have been possible. I am also grateful to my second supervisor Tom Crossley, who generously hosted me during my visit at the University of Essex and also provided great feedback. I would also like to thank Derya Uysal for agreeing to serve on my committee.

I greatly benefited from being part of the Evidence-Based-Economics graduate program. It was a fantastic experience and I would like to thank the speakers of the program, Joachim Winter and Florian Englmaier for their effort. The EBE-program would not have been the same without Julia Zimmermann, who I owe a very big thank you for her patience and support in administrative issues.

Moreover, I would like to thank all the people who have accompanied me during this journey. Sebastian Wichert thank you for being my co-author and a supportive colleague. I am also indebted to Christoph, Corinna, Michael and Tobias for being great colleagues at the Seminar für empirische Wirtschaftsforschung. Katharina, Natalie, Raphael, Sarah you made this PhD journey worthwhile.

A special thank you goes to Katharina for her fantastic help with the layout.
Last but not least. I would like thank my family for their constant support.

Contents

List of Figures

List of Tables

Preface

Knowledge about the causes of health and disease is now greater than at any time in human history: We understand that chronic diseases, which account for about 70% of deaths globally, are linked to certain aspects of our lifestyle (WHO 2017). It is an established fact that smoking causes lung cancer (US Department of Health and Human Services et al. 2014). And there is no ambiguity in the evidence relating physical inactivity and certain dietary behavior to obesity, high blood pressure, type 2 diabetes, coronary heart disease and cancer (Forouzanfar et al. 2016; Lee et al. 2012). Still, 1.1 billion people smoked tobacco in 2015 (WHO 2018), and the incidence of obesity is increasing globally (Forouzanfar et al. 2016).

From a traditional economists' viewpoint, these behaviors stem from rational decisions based on individual preferences: Individuals maximize their utility by choosing an optimal level of positive and negative health investments and therefore bad health constitutes a conscious choice (Grossman 1972). In this framework, there is only a limited scope for policy interventions (Cawley and Ruhm 2011).

The notion of individual health as a purely self-inflicted consequence of individual behavioral choices contrasts with an economic and epidemiological literature highlighting the social dimension of health. While this literature acknowledges the importance of individual behavior as determinant of individual health, it is also concerned with the role of *culture*, *availability* and *affordability* in shaping these behaviors (Marmot and Wilkinson 2005, Chapter 1). For example, it has been shown that health behavior varies across subgroups of the population, with high socio-economic status being the greatest predictor for healthy behavior (Cawley and Ruhm 2011).

The question whether individuals engage in unhealthy behavior only because of individual optimizing behavior or also because of external circumstances, has important normative implications. The wide-spread popularization of smartphones provides companies, policy makers and health insurers with the technical capabilities to observe and monitor health behaviors. For example in early 2016 one major German statutory sickness fund launched an app, enabling its insurees to document physical activity and gain a reward of up to 180 euros a year.[1] Neglecting external factors influencing the ability to engage in healthy behavior, such an app may simply be viewed as a useful tool to incentivize physical activity. However, for example single parents with low-paying jobs might just not be able to fit requirements for gaining a reward into their daily schedule. And other individuals may find it hard to comply because of cognitive limitations. Consequently, a more negative view is that such apps provide a tool for redistribution within the public social security system to those who fit into the category of healthy-wealthy-wise (Currie 2009).

Moreover, external environments and living circumstances appear to drive health inequalities beyond their role in shaping health behaviors and providing access to health care (Marmot 2015). For example, the Whitehall Study of British civil servants has documented a robust gradient in the relationship of occupational rank and health (Marmot et al. 1991). And even within countries, life expectancy varies by location of residency (Chetty et al. 2016; Lleras-Muney forthcoming). While individuals can choose not to smoke, they cannot easily escape air pollution, which might give them lung cancer anyway. In their analysis of an increase in mortality for low educated white individuals in the US, Case and Deaton (2017) put forward the explanation of *cumulative disadvantage*. They argue that labor market conditions have been worsening from one generation of low educated individuals to the next. Together with a loss in family structures, this loss of opportunity triggers a process of long-term decline. In light of these findings it is essential to raise awareness that "*Health is not simply a matter of personal responsibility*" (Lleras-Muney forthcoming; Marmot 2015). However, much of the evidence documenting a relationship between circumstances beyond individual control and health is based on correlations. In order to convince policy makers that their actions can influence individual health - even when a measure is not directly targeted at the health system - it is important to show that these findings also hold up when institutional settings vary exogenuously.

In the first two chapters of this dissertation I exploit exogenous variation created by

[1] https://plus.aok.de/inhalt/aok-bonus-app/

modern German history to study how an institutional framework can affect health. My focus is on those periods in an individual's life cycle, where she is especially susceptible and vulnerable to health shocks, namely the perinatal period and old ages. Chapter 1 exploits the onset of World War II as a natural experiment to learn about the short term consequences of the onset of the war for newborn health and perinatal infant mortality. The German population was initially not subject to very extreme war-related conditions. However, the institutional framework people were living in changed suddenly. The economy was transformed into a war-time economy and large scale drafting of doctors put a strain on the public health system. In Chapter 2 I study contemporaneous health differences between East and West Germans arising as a consequence of the German separation. Both chapters exemplify that political actions which are not targeted at the health system, can have consequences for individual health. Mothers did not choose to give birth in Munich some months into World War II rather than some months prior to the war, yet they were more likely to see their infant die within the first week after birth. Likewise, East Germans born prior to 1949 are of worse health than their West German counterparts, even though they did - in most cases - not actively opt for a life in the East.

In addition to a credible identification strategy, empirical research on the determinants of health requires suitable data. Only very few countries, such as Sweden, Norway and Denmark, provide researchers with access to administrative data sets linking individual health records to detailed information on labor market outcomes, education and additional life experiences (Farbmacher et al. 2016; Gustavsson et al. 2012; Nilsen et al. 2012). Moreover, even detailed registry data, does lack measures of subjective health, health-related life quality, or cognitive and non-cognitive ability. Therefore, much of empirical research relies on survey data. Surveys are a convenient tool to obtain representative data for any population. Furthermore they are not restricted to certain types of outcomes and longitudinal surveys even allow to observe individual-level changes over time. On the other hand, the process of data collection inevitably leaves its mark on survey data. As a result, survey responses may not always exactly correspond to a - possibly hypothetical - true value but contain survey error (Alwin 2007). A well-known source of measurement error in face-to-face surveys is the interviewer. Interviewers have been shown to influence responses by the sheer presence of their demographic characteristics, by their interviewing style or by their willingness and ability to offer clarifications to respondents (Groves 2004). The third chapter of this dissertation investigates an additional channel for interviewer-induced measurement error: In an attempt to scale down the impact of interviewers, the administration of a cognitive test may be shifted partially or fully to a technical device. I show, however, that interviewers do not use the technical device for all respondents in my setting, creating unintended variation. These findings highlight that applied researchers working with survey data on cognitive ability should think carefully about potential sources of error.

Each chapter of this dissertation is self-contained. A common theme is the focus on empirical analysis and the use of modern microeconometric methods.

Chapter 1, titled *Birth in times of war - An investigation of health, mortality and social class using historical clinical records* is based on joint work with Sebastian Wichert. This chapter is motivated by a growing literature exploiting World War II as a natural experiment to study the impact of an adverse early life environment on later life outcomes. It has been documented that individuals who were affected by World War II in early life, are more likely to suffer from diabetes and depression (Atella et al. 2016; Kesternich et al. 2014) and show modified behavior at old ages (Kesternich et al. 2015). However, little is known about the short term effects of World War II. We seek to bridge this gap. Specifically, we estimate the short-term impact of the onset of World War II on newborn health and perinatal mortality. We focus on the first two years of WWII, a period when military operations took place outside of Germany and there was no nutritional shortage. We collected an entirely new data set of historical clinical birth records from the largest birth hospital in Munich, Germany. Our unique data contain around 10,000 births and miscarriages which took place in the hospital between December 1937 and September 1941. Our findings reveal no change in perinatal health at the onset of the war but a large and robust increase in perinatal mortality. This mortality effect can mainly be attributed to live born children who die before leaving the hospital. Infants from all social classes are more likely to die after the onset of the war and low birth weight infants are disproportionally affected. The mortality effect is greatest during the first months of the war. This is consistent with the interpretation that the onset of WWII acted as shock to individuals and the public health system, initially leading to a

jump in perinatal mortality and then gradually fading out. In our discussion of potential mechanisms we focus on maternal stress and a decline in medical quality caused by sudden conscription of doctors. We argue that the latter channel is more important in our setting, as the increase in mortality is largest where medical quality should matter.
Data on birth outcomes during the onset of World War II is not readily available. Therefore, this project involved a significant amount of preparatory work to obtain a our data set. After the birth records had been cleaned from mold and relocated to the university archive, we hired student assistants to digitize the entries. Several challenges occurred during this process. The entries in the birth records use a traditional form of German handwriting that required special reading skills on the side of the student assistants. Moreover, information on one birth is spread out across several documents and matching this information is not always straightforward. We devoted a high effort to obtaining a high-quality data set and double checked all seemingly inconsistent entries.

Chapter 2 is titled: *Does the wall still exist? Health differences between East and West Germans*. In this work I exploit the German separation and reunification as a natural experiment to learn about the long term effects of living under a Socialist regime on individual health. Identification rests on the assumption that East and West Germans would not systematically differ in the absence of a separation. Previous literature has supported this claim (Alesina and Fuchs-Schündeln 2007; Görges and Beblo 2015).
I document health differences between East and West Germans across time, age and cohorts to account for the fact that it is not straightforward to define a single "GDR" effect of interest. Moreover, even in 2018 living conditions in East and West Germany vary across several dimensions. Given these persistent differences, it seems unlikely that today's gap in health status can be attributed to experiences prior to reunification alone. In order to obtain an estimate of the long-term effect of having lived in East Germany net of contemporary input into the health production functions like income and unemployment, I apply the mediation analysis framework outlined in Acharya et al. (2016). I estimate the controlled direct effect, a well defined quantity corresponding to the treatment effect in an experiment where both the treatment and and post-treatment are manipulated by the experimenter.
Using data from the German socio-economic panel, I document a strong and persistent

gap in health among individuals whom I observe at older ages. East Germans born 1949 or earlier have been diagnosed with more chronic diseases, have lower mental health and physical health related life quality and rate their health significantly worse. While removing the influence of contemporary factors reduces the magnitude of estimates of health inequalities between East and West, a significant gap remains for older individuals. I cannot fully disentangle age and cohort effects. Nevertheless, I document evidence consistent with the interpretation of an acceleration of the aging process of East Germans between the ages of 40 and 60. Furthermore, I argue, that earlier cohorts were hit harder than younger cohorts by the shortcomings of the GDR health system and experienced more blatant repression during the 1950s.

The third chapter, *Does the laptop always help? Non-compliance and interviewer effects in cognitive tests*, deals with one possible source of survey error in the administration of a cognitive test. Cognitive tests in surveys allow researchers to study cognitive decline associated with aging (McArdle et al. 2007; Whitley et al. 2016), analyze determinants of economic decision making (Smith et al. 2010) and relate labor market outcomes to cognitive ability (Heckman et al. 2006; Heineck and Anger 2010). However, administration of a cognitive test within surveys is a challenging task. Test scores of cognitive tests do not only reflect cognitive ability but also contextual circumstances. Furthermore, heterogeneity in interviewing styles and interviewers' ability are also likely to induce measurement error. One way to scale down the impact of interviewers, is to shift administration of a cognitive test partly or fully to a technical device, such as the interviewer's laptop. This chapter studies the case of the word recall test in the third wave of Understanding Society - The UK Household Longitudinal Study. In the word recall test respondents hear a list of ten words and are subsequently asked to recall as many words as possible. According to the study protocol, the words should be read by the laptop of the interviewer. However, for about 20% of respondents, the interviewer deviated from the default procedure and read the words herself. Respondents who heard the words from the interviewer, perform on average worse than other respondents. Moreover, interviewer intra-class correlations are elevated when the test is administered without the laptop.

I aim to answer three questions. I begin by asking which determinants drive imperfect compliance with the study protocol. Here, I am interested in respondents' characteristics predicting deviation from the default mode as well as heterogeneity across interviewers. Next, I turn to the question of what drives the difference in performance between the two modes. Respondents were not randomized to modes in our setting. Therefore, differences in performance may either constitute mode effects, i.e. the test score a particular respondent achieves depends on the mode of administration, or they may stem from selection effects, i.e. cognitive ability is not evenly distributed across the two groups of respondents. I exploit the existence of test scores from additional cognitive tests, that were administered to all respondents in the same mode, to understand whether the two groups of respondents differ in cognitive ability. Finally, I seek to answer whether administration via the computer successfully reduces interviewer effects in our setting. Different interviewers deviate from the default procedure for different respondents. This mechanism can contribute to a disparity in the interviewer intra-class correlation between the two groups, similar to non-response error variance (Brunton-Smith et al. 2012; West and Olson 2010).

My results show that hearing the words from the interviewer is associated with individual characteristics such as age or hearing problems. Moreover, the propensity to read the words varies greatly across interviewers. Selection effects appear to be the main driver of performance differences in our setting. Those respondents who hear the words from the interviewer, perform significantly worse also in other cognitive tests. Finally, the differences in interviewer intra-class correlations between the two modes are greater in the word recall test than in all other cognitive tests. Therefore we conclude that the use of the laptop does indeed seem to reduce interviewer effects in the word recall test. These findings suggest that the use of laptops in administration of the word recall test is preferable, despite the problem of possible non-compliance.

Chapter 1

Birth in Times of War - An investigation of health, mortality and social class using historical clinical records [*]

1.1 Introduction

Early childhood and the time in utero may be one of the most critical time periods in life (Almond and Currie 2011). In order to establish a causal effect of adverse early-life environment on later life outcomes, a growing literature exploits historical shocks like natural disasters, recessions, famines and wars. The by far greatest shock that has affected living cohorts in Western Europe is World War II (WWII). Individuals exposed to WWII in utero or early-life have been shown to have higher morbidity and mortality rates, worse socio-economic outcomes and even a modified behavior at older ages (see e.g. Atella et al. 2016; Jürges 2013; Kesternich et al. 2014, 2015; Van den Berg et al. 2016). These findings are based on samples of the surviving population. If individuals who survive infancy during the war do systematically differ from survivors of other cohorts, estimates of long term effects may be biased (see e.g. Lindeboom and Van Ewijk 2015; Van Ewijk and Lindeboom 2016). As historical individual level data on birth outcomes are hardly available,[1] it is unclear whether the negative effects of WWII remained latent until later life or were already present at time of birth.

[*] This chapter is based on joint work with Sebastian Wichert and an earlier version is also included in Sebastian Wichert's thesis available at *https://edoc.ub.uni-muenchen.de/21785/1/Wichert_Sebastian.pdf*.

[1] A rare exception is the "Dutch Famine Birth Cohort Study". See Lumey et al. (2011) for an overview.

The aim of this research project is to estimate the short term effects of the onset of WWII on perinatal health and mortality of infants. To explore, how war induced changes in perinatal infant mortality are related to individual characteristics associated with outcomes later in life, we estimate heterogeneous treatment effects by social group and infant health. Furthermore we investigate several mechanisms through which the onset of WWII may affect a newborn's health. We collected an entirely new data set of historical clinical birth records from the largest birth hospital in Munich, Germany. Our unique data contain around 10,000 births and miscarriages which took place in hospital between December 1937 and September 1941. Besides a rich set of demographic variables, our data set contains detailed socio-economic information. In our empirical strategy we exploit the unexpected onset of WWII as natural experiment.

Even 60 years after the end of WWII, its consequences continue to shape individual life outcomes. Kesternich et al. (2014) analyze retrospective life data and document that individuals exposed to WWII during childhood are more likely to suffer from diabetes and depression at old ages. Atella et al. (2016) investigate the impact of WWII on health in an Italian context. They can link stress in early life caused by exposure to intense conflicts to depression, while exposure to famine appears to increase the probability of diabetes in later life. A number of research projects exploit WWII to study the long term consequences of hunger in early life. For example Van den Berg et al. (2016) provide causal evidence that hunger leads to a decrease in adult height and Kesternich et al. (2015) show that individual behavior can serve as a pathway between early life shocks and later life health. Similarly, the small literature drawing on historical birth records to study the short term impact of WWII on health at birth mainly focuses on the role of nutritional shortage during gestation. Stein et al. (2004) find those individuals affected by the Dutch Hunger Winter 1944/1945 during the third trimester to have decreased birth weight and birth size. No effect is found for individuals exposed during earlier stages of pregnancy. Using data similar to ours, Floris et al. (2016) study how birth weight evolves over the course of WWI in one Swiss hospital. In their setting food rationing during the end of the war leads to a decrease in birth weight for children from medium SES families. By contrast, high SES families can compensate price shocks and low SES families benefit from public interventions. Our work is also related to a strand of literature investigating the impact of shocks in utero and maternal stress using modern data.

This literature exploits a variety of shocks, for example natural disasters (Currie and Rossin-Slater 2013; Torche 2011), terrorist attacks (Quintana-Domeque and Ródenas-Serrano 2017) or mass layoffs (Carlson 2015). Most of these studies can document a small decrease birth weight following an external shock. An exception is Currie and Rossin-Slater (2013) who do not find a change in birth weight, but show that stress in utero affects more extreme health outcomes.

Finally, in our discussion of potential mechanisms, we connect to the literature evaluating the effects of physician supply on health outcomes. Drawing on historical data, Liebert and Mäder (2016) exploit the sudden expulsion of Jewish doctors in Nazi Germany as natural experiment. They find a decrease in regional physician coverage to have substantial detrimental effects on infant mortality.

While we do not find any sizable effects of the onset of the war on health measured as birth weight or asphyxia, we can document a strong, robust increase perinatal infant mortality. This mortality effect can mainly be attributed to live born children who die in hospital prior to being discharged. Perinatal mortality increases for all social classes and disproportionally for very low birth weight infants. Previous literature relating WWII to health outcomes often focuses on extreme effects of the war like bombings, hunger, combat and dispossession. Similarly to Lindeboom and Van Ewijk (2015) and Van Ewijk and Lindeboom (2016), we study less extreme war-related events. We focus on the first two years of WWII, a period when military operations took place outside of Germany and there was no nutritional shortage. Our main contribution is to document an effect of WWII on perinatal child mortality even in the absence of extreme conditions. The onset of WWII acted as shock to individuals and the public health system, which initially led to a jump in perinatal mortality and then gradually faded out. This interpretation is consistent with historical evidence, showing that the onset of the war caused turmoil in the health system and disrupted daily life.

Two mechanisms are potentially driving our results. Firstly, high maternal stress levels may contribute to an increase in infant mortality, as the onset of a war comes a long with great uncertainty and many husbands were drafted. Secondly, a sudden shortage of doctors can lead to a decrease in medical quality. With the onset of WWII, large scale conscription reduced the number of doctors considerably and put the hospital under strain. We find the mortality effects to be stronger, where medical quality should matter. Therefore we conclude the decline in medical quality to be the more important channel.

Our results have important implications for the literature on long term effects of WWII. We document a disproportional increase in mortality for very low birth weight infants, suggesting that studies using samples of the surviving population provide a lower bound for the true effect.

The remainder of the paper proceeds as follows: Section 1.2 provides more detailed information on the historical background. Section 1.3 describes our data, the way we constructed our variables and presents first descriptive analyses. We explain our empirical strategy in section 1.4 and present our results in section 1.5. Section 1.6 concludes.

1.2 Historical and institutional setting

1.2.1 General historical background

Events leading to WWII

When Hitler and the Nazi Party seized power in 1933, the transformation from a weak democracy to an autocratic dictatorship began immediately. Within months, public institutions, local and regional authorities, judicature and even private clubs were brought under the control of the Nazi party. Non Aryan Germans were dismissed from jobs in the civil service and whoever publicly raised criticism became subject to brutal repression (Evans 2004, pp. 498-509). Against the terms of the treaty of Versailles the Nazis also launched the rearmament of the German military. In 1935 a military law made all male Germans between 18 and 45 liable to military duty. Nevertheless, neither the German public nor other European powers were aware of the imminent threat of a war. When Hitler began with the restoration and expansion of Germany, he did so using massive political pressure on foreign governments instead of using military force. Between 1935 and 1938 three former German territories, separated after WWI, were reintegrated into Germany (Territory of the Saar Basin by referendum, Rhineland and Memel Territory by occupation, Austria by voluntary annexation). The first military aggression took place in 1938 when Germany occupied the Sudeten German territories in Czechoslovakia. The essential powers in Europe - Great Britain, France, Italy - tolerated this aggression to appease Hitler and to avoid a new war in Europe. Even when Hitler violated previous agreements again in 1939 by occupying the rest of Czechoslovakia, they did not intervene in any military way.

After these successes Hitler and the Nazi state were celebrated by the majority of the German population, who perceived Germany to be a world power again. The general public hoped that wars could be avoided in the future as well - either because Hitler had already achieved his goals or because his political measures were sufficient to do so (see Frei 2013, p. 150).

World War Two

WWII began with the invasion of Poland on September 1st, 1939. For the first time the German military experienced resistance , and Poland's guarantor powers - France and Great Britain - declared war on Germany. This had been unexpected by the German public, to whom it was clear quickly that this conflict would be different from any other conflict since 1918. There was a great feeling of uncertainty and no euphoria among the population, since most people had experienced the negative consequences of the previous war. Prior to 1942, military operations (i.e. air strikes or combat) mainly took place outside of Germany (see Permooser 1997). Therefore the German population was initially not subject to direct effects of the war like hunger and bombings. Nevertheless, the onset of WWII marked a distinct break in the daily routine. Firstly, conscription affected a great number of men who were subsequently absent from their families and workplaces. At the end of 1939 around 4.2 million men out of a male population of 33.8[2] million were serving the military, another 3.5 million men were drafted in 1940 (Overmanns 2009, p. 217).[3] Men were drafted based on their year of birth and previous military experience without social class dependent privileges or exceptions (Absolon 1960, pp. 4, 152-153).[4] Secondly, in order to prioritize production for military purposes, the economy was transformed into a wartime economy. Three days before Germany invaded Poland, the regime announced the introduction of ration stamps for food and other commodities like fabric, leather and soap. The local popula-

[2] German Reich as of 1937.

[3] Poland was already defeated (with minor German military losses) in October 1939 and lots of soldiers returned on furlough. However, the atmosphere in Germany remained tense as there was a constant threat that soldiers, who had just returned, would be sent to war again.

[4] Only certain conscripts were (temporarily) exempt if their specific occupation duty was classified - again on a case-by-case basis - as indispensable for "homeland defence".

tion in Munich responded to the introduction of ration stamps with a rush to the shops and officials were not well prepared to manage the new circumstances.[5] While there is no evidence suggesting that the population was affected by serious hardship during the first two years of the war, daily life became more complicated. Long queues in front of shops were common especially in the first weeks of the war and commodities like furniture and bedding eventually became objects of speculation. There was no general shortage of food.[6] However, food quality declined and availability of certain categories of food varied. Pregnant women received preferential treatment. Unlike the general population they were allocated whole milk and when coal was in short supply in February 1940, pregnant women were eligible for extra rations. Records of the hospital our data come from, do not indicate any problems with the catering of patients or shortage of fuel.

The German health system entered the war ill-prepared. No comprehensive concept existed on how to operate medical services for the civil population. Instead the military was given full priority. The army made frequent use of its authority to dispose all resources of the civil health system. Besides confiscations of local hospitals, large scale drafting of physicians lead to conflicts between the military and the civil sector. Already in fall of 1939 one third of all available physicians were in military service. In order to mitigate the shortage of physicians, the state granted final year medical students their approbations prematurely. Turmoil in the health system was greatest during the first weeks of the war[7], while the situation remained tense throughout (Christians 2013, pp. 237-244; Süß 2003, pp. 181-212).[8]

Fertility and childbirth under Nazi rule

Childbirth was no longer considered a private matter in Nazi Germany. Between 1900 and 1933 the number of yearly births in Germany had fallen by more than 50% (Sensch 2006), an unacceptable state for a regime adhering to a pro-natalist ideology. However, as the Nazis' world view was based on eugenics, their goal was not to increase every-

[5] Confidential quarterly reports by the Economic Department give a detailed account of the Economic situation in Munich (Stadtarchiv München 1939-1940).

[6] Daily food rations were sufficient until the end of 1944 (see Jürges 2013; Kesternich et al. 2015).

[7] Even high ranking Nazi officials had to acknowledge this tense situation (König 1939, pp. 385-386; KVD Bayern 1939, p. 387).

[8] A notable exception was the constant supply of pharmaceuticals, which was secured during the first years of WWII due to large production capacities (Süß 2003, p. 197).

body's fertility. The regime used brutal repression to prevent reproduction among those considered to deteriorate the gene pool (Fallwell 2013). In order to boost birthrates among healthy "Aryan" Germans, the Nazis combined family propaganda, a ban of voluntary abortion[9] and material incentives.[10] Indeed, the absolute number of births was increasing in the years prior to WWII.

Even the choice of location of delivery became infused with political agenda. The Nazi regime was heavily opposed to the increasing trend towards hospital birth. While the concept of women giving birth at home within their family members fitted in perfectly with the Nazi ideology, home births also spared the resources of the health system. Efforts to propagate home births climaxed in the so called "midwife edict" of September 1939 (RMI 1939). This edict requested hospitals to reject pregnant women without medical or social indication for hospital births. The hospital our data come from was a teaching hospital and therefore exempt from this rule. Due to decisive resistance of the association of gynecologists the "midwife edict" was modified in 1940, granting women a choice over the location of delivery (Zander and Goetz 1986). Official statistics indicate that the proportion of hospital births in Germany was growing during the Nazi era despite all otherwise attempts. In 1935, 25% of live births took place within a hospital compared to 38% in 1940 (Statistisches Reichsamt 1933-1940).[11]

1.2.2 The hospital

The hospital *Frauenklinik Maistrasse* is the oldest and one of the largest gynecological hospitals in Munich. It was founded as a state-run university hospital in 1884, succeeding the municipal birth house. In its first years the hospital mainly served lower-class and often single mothers. Women of higher social status traditionally gave birth at home. However, after moving into its current venue in 1916, the *Frauenklinik Maistrasse* became one of the leading gynecological hospitals in Germany and attracted patients among all social classes. The hospital was divided into a general and a private ward.

[9] In the late 1920's Germany was given of the most liberal abortion policy in the developed world (Usborne 2011).

[10] For example, eligible newly wed couples received marriage loans, whose repayment was reduced with each child born.

[11] Before 1935 official statistics only counted the number of births within maternity clinics. In urban areas the proportion of hospital births was even higher.

Most patients were admitted to the general ward and their treatment was completely covered by public health insurance. The private ward enabled the hospital to extract rents from more affluent, often privately insured patients. These patients received special attention by the senior staff.[12]

Deliveries were supervised by both doctors and midwives, but only doctors carried out surgeries and medical procedures. With the onset of WWII the conscription of physicians heavily affected the daily routine. The director of the hospital frequently complained in letters to the state administration and applied for exemptions from military service for many of his doctors. For example, in a letter from December 1939 he stated that already seven of his doctors were serving the military and several more had received draft calls. Much of the workload was shifted to recent graduates and unpaid trainees. In the Nazi era, the hospital carried out large numbers of forced abortions and sterilizations on women who allegedly suffered from hereditary diseases.[13]

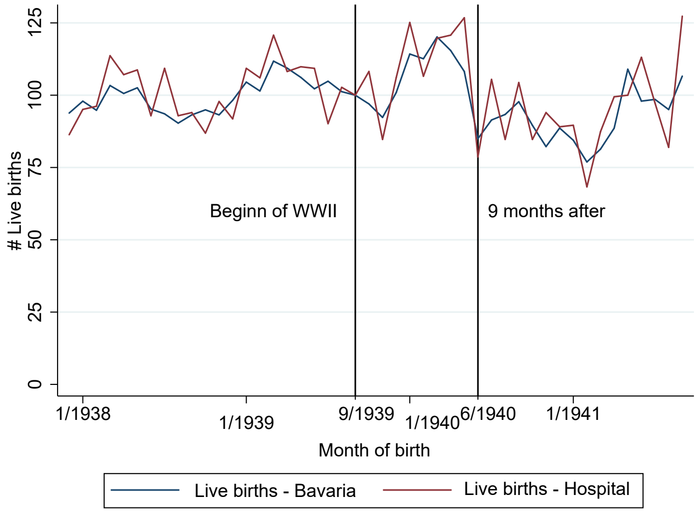
Two groups of births are likely to be oversampled in our data: births of mothers with very low socio-economic status and pathological births. Home birth was no option for women living under crowded or unsanitary conditions. Often these women would seek admittance to the hospital weeks before delivery, where they acted as teaching material for medical students and midwives in training. Women in risk of a pathological birth were referred to hospital by midwives and gynecologists. Still, as hospital births had become quite common especially in big cities by 1937, our sample is broad enough to draw conclusions also for other groups. Around half of our observations equal at least a status of a skilled worker and almost 60% of women entered the hospital without any pre-existing risk factors. Between 1938 and 1940 around 17% of all Munich live births took place in the Frauenklinik Maistrasse (see Table A.2 in the Appendix).

Figure 1.1 shows the monthly trend in the number of live births for our hospital and the whole state of Bavaria, normalized for September 1939. Both trends match quite well and no structural breaks (e.g. at the begin of the war or due to the "midwife edict") point to any differential selection into our hospital.

[12] The hospital was only allowed to charge a publicly regulated daily rate for patients in the general hospital with no extra fees for treatments. In private ward, on the other hand, there were extra fees for treatment on top of a higher daily rate.

[13] Most such records state the women suffered from "hereditary feeble-mindedness". Since the 1990's the hospital has endeavored to shed light on its role during the Nazi era (Stauber 2012).

Figure 1.1: *Number of live births in Bavaria and hospital*



Notes: Number of live births in Bavaria and our hospital by month of birth, with the number of births in September 1939 being normalized to 100.
Source: Bayerisches Statistisches Landesamt 1937-1942

1.3 Data

1.3.1 Sample selection and variables

Sample selection

We digitized the universe of entries in the hospital's birth records from December 1937 to September 1941 (see Appendix A.1). The 10,325 observations consist of live- and stillbirths, miscarriages and a small number of other conditions.[14] Other conditions comprise women who came to the hospital post birth, women receiving treatment during pregnancy, medically induced interruptions as well as forced abortions and sterilizations. We do not consider these 196 observations in our analysis.

In our definition of live births, stillbirths and miscarriages, we maintain the categorization found in the clinical records. A law of 1935 required midwives and physicians to report all miscarriages to the authorities who were wary of illegal abortions.[15] In our birth records around 1,200 observations are marked as miscarriage. These mostly lack information on the child such as weight, length and sex. Miscarriages mostly took place outside the hospital and women only went to the hospital to seek treatment afterwards.

[14] A twin birth results in two observations.

[15] *Vierte Verordnung zur Ausführung es Gesetzes zur Verhütung erbkranken Nachwuchses.* Vom 18. Juli 1935. In: RGB1 I Nr. 82, 25. Juli 1935.

Patterns of selection into the hospital are very likely to vary between women who intend to give birth and women who are treated after a miscarriage. Therefore we exclude miscarriages from our main analysis.[16]

Outcome and control variables

Our primary outcomes are perinatal infant mortality, measured as whether an infant left the hospital alive, and birth weight. Birth weight is an overall measure of health at birth (McIntire et al. 1999), while also being a predictor of future life outcomes, for example educational attainment and adult height (Behrman and Rosenzweig 2004). Currie and Rossin-Slater (2013) find that birth weight is not affected by exposure to stress in utero, while there is an effect for more extreme measures of newborn health. Therefore we also analyse asphyxia and maturity. Asphyxia is caused by deprivation from oxygen during the process of birth. It often results in the death of the infant and can cause long term damage to surviving infants. Maturity is an indicator whether the birth takes place at full term. It is assessed by the appearance of the infant.[17]

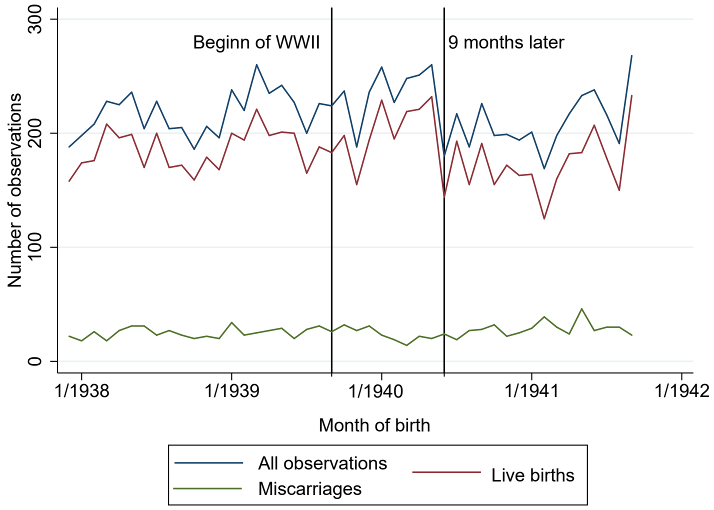
Our control variables include characteristics of the mother, namely age, the number of previous pregnancies and most importantly a measure of social status which is derived from the occupational information in the birth records. We categorize this occupational information according to HISCLASS, a validated measure of historical social classes. Each occupation is assigned one out of 12 social classes defined as "a set of individuals with the same life chances" (Van Leeuwen and Maas 2011, p. 18). In our empirical analysis we rely on the previous literature and use a compressed 7-class version of HISCLASS (Abramitzky et al. 2011; Schumacher and Lorenzetti 2005).[18] For each observation, the birth record contains either the occupation of the father or the occupation of the mother. If the occupation of the mother is given, the entry uses the female version of the occupation in German language. Otherwise the male version is used, mostly with a suffix like -wife, -daughter or -widow. We classify women accordingly as "working",

[16] Entries marked as stillbirth, on the other hand, almost always include characteristics of the child but do not generally contain a gestational age. Partly the definitions of stillbirth and miscarriage seem to overlap since weight and gestational age of "miscarriages" exceeds 1,000 grams and the fifth month in individual cases, while stillbirths" encompass a few infants with a birth weight below 1,000 grams.

[17] To assess maturity, midwives checked the colour of skin, body hair, ear conch and the appearance of genitals.

[18] This simplifies the interpretation of regression coefficients, attenuates possible coding errors and increases sample size within classes. A detailed description of the occupational coding can be found in Appendix A.2.

***Figure 1.2:** Timeline of observations in hospital*



Notes: Number of all observations, live births and miscarriages by month of birth.

"wife" or "single". Note that this approach assumes that the categories are mutually exclusive, while in reality a married women may also work. Further control variables include the sex of the infant, multiple births and the fetal position. Fetal malpositions and malpresentations are among of the most frequent reasons for complications at birth. As these can be diagnosed prior to birth easily, we expect births with an abnormal fetal position to be overrepresented in our data. Several factors, such as tumors, maternal anatomy or high parity are associated with fetal position in full term births (MacKenzie 2006). Still it is unclear why the onset of the war should causally affect the composition of fetal positions in the population. Consequently we think of fetal position as a proxy for the risk a birth can be associated with ex ante.

Descriptive statistics

Figure 1.2 displays the number of total observations, the number of live births and the number of miscarriages over our period of observation. The graph shows a distinctive drop in the number of births in June 1940 - nine months after the begin of the war, when many men were drafted for the invasion of Poland. Similarly another drop occurred in February 1941, 9 months after the begin of the invasion of France. In mid 1940 many of the German soldiers were granted furlough, leading to an increased number of births towards the end of the observation period.

Table 1.1 shows that 96% of the births in our sample are live births. In 93.5% of all

Table 1.1: *Descriptive statistics - Births*

General characteristics	N	Mean	SD	Min	Max
Birth after 9/1939	8828	0.543	0.498	0	1
General ward	8828	0.931	0.253	0	1
Length of stay	8769	12.704	11.934	0	379
Live birth	8828	0.960	0.195	0	1
Infant leaves hospital alive	8828	0.936	0.246	0	1
Regular fetal position	8688	0.919	0.273	0	1
Mother	N	Mean	SD	Min	Max
Age of mother	8828	27.921	6.211	14	50
Parity	8826	2.208	1.804	1	19
Status is wife	8828	0.651	0.477	0	1
Status is own job	8828	0.310	0.462	0	1
Status is single, divorced or widowed	8828	0.031	0.173	0	1
Social status	N	Mean	SD	Min	Max
Higher managers & professionals	8500	0.069	0.253	0	1
Lower managers & professionals, cleric	8500	0.194	0.396	0	1
Foremen & skilled workers	8500	0.225	0.418	0	1
Farmers	8500	0.072	0.259	0	1
Lower skilled workers	8500	0.133	0.340	0	1
Unskilled workers	8500	0.281	0.450	0	1
Farm workers	8500	0.025	0.157	0	1
Infant	N	Mean	SD	Min	Max
Male	8822	0.527	0.499	0	1
Birth weight	8820	3218.620	601.065	280	5510
Length of infant	8815	49.998	3.108	19	61
No. of infants	8828	1.027	0.164	1	3
Asphyxia	6784	0.023	0.148	0	1

Notes: Descriptive statistics of births in sample (excluding miscarriages).

births the infant left the hospital alive,[19] implying that in addition to the 4% stillborn children, 2.5% of infants died in hospital after birth. Most births (93%) took place in the general ward. The mothers in our sample are on average 28 years old and experience their second pregnancy, 30% of the women in our sample report an own occupation. Unreported analyses show that lower classes are overrepresented among these working women.

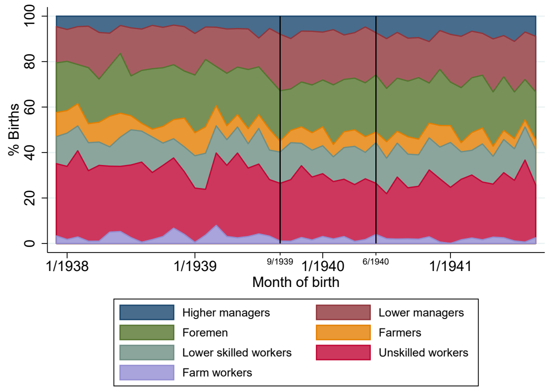
[19] The median newborn stayed in hospital for 9 days after birth.

Table 1.2: *Mean comparison - Births*

General characteristics	Mean before war	Mean after war	Diff	SD	p	N before war	N after war
General ward	0.952	0.914	-0.0374***	0.005	0.000	4035	4793
Length of stay	12.560	12.824	0.2639	0.256	0.303	3979	4790
Live birth	0.966	0.956	-0.0098*	0.004	0.019	4035	4793
Infant leaves hospital alive	0.949	0.924	-0.0247***	0.005	0.000	4035	4793
Regular fetal position	0.920	0.918	-0.0020	0.006	0.728	3954	4734
Mother	Mean before war	Mean after war	Diff	SD	p	N before war	N after war
Age of mother	27.845	27.985	0.1406	0.133	0.289	4035	4793
Parity	2.188	2.224	0.0356	0.039	0.356	4035	4791
Status is wife	0.614	0.682	0.0675***	0.010	0.000	4035	4793
Status is own job	0.339	0.285	-0.0533***	0.010	0.000	4035	4793
Status is single, divorced or widowed	0.037	0.026	-0.0108**	0.004	0.003	4035	4793
Social status	Mean before war	Mean after war	Diff	SD	p	N before war	N after war
Higher managers & professionals	0.055	0.080	0.0253***	0.006	0.000	3878	4622
Lower managers & professionals, cleric	0.174	0.212	0.0375***	0.009	0.000	3878	4622
Foremen & skilled workers	0.226	0.224	-0.0020	0.009	0.826	3878	4622
Farmers	0.084	0.062	-0.0222***	0.006	0.000	3878	4622
Lower skilled workers	0.123	0.141	0.0178*	0.007	0.016	3878	4622
Unskilled workers	0.305	0.261	-0.0434***	0.010	0.000	3878	4622
Farm workers	0.032	0.019	-0.0130***	0.003	0.000	3878	4622
Infant	Mean before war	Mean after war	Diff	SD	p	N before war	N after war
Male	0.525	0.529	0.0042	0.011	0.696	4033	4789
Birth weight	3227.907	3210.802	-17.1054	12.847	0.183	4031	4789
Length of infant	50.198	49.830	-0.3674***	0.066	0.000	4030	4785
No. of infants	1.028	1.026	-0.0019	0.004	0.583	4035	4793
Asphyxia	0.021	0.023	0.0028	0.004	0.483	1991	4793

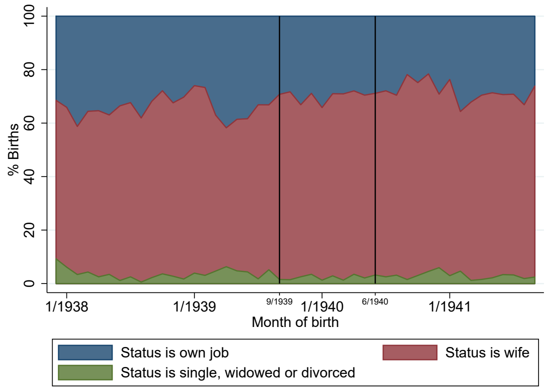
Notes: T-tests on the equality of means by war (excluding miscarriages). Significance levels: ***p < 0.01, ** p < 0.05, and * p < 0.1.

***Figure 1.3:** Composition in terms of social classes over time*



Notes: Proportion of mothers by social class by month of birth.

***Figure 1.4:** Composition in terms of marital and working status over time*

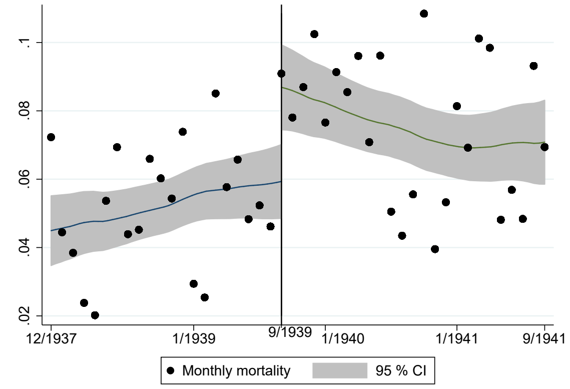


Notes: Proportion of mothers by marital and working status by month of birth.

We conduct simple t-tests to check for changes from the prewar and to the war period (see Table 1.2). There is no difference in terms of age and parity of mother, as well as maturity and weight of the infant. The proportion of regular fetal positions does also not change significantly. Since the proportion of regular fetal positions in the population is unlikely to be affected by the war, this suggests that women at risk of a complicated birth were not sent to the hospital more frequently during the war. On the other hand, the composition of mothers in terms of social status, labor force participation and marital status does show some changes. This highlights the importance of controlling for socio-economic characteristics. When examining how the socio-economic composition evolves over time, we find no abrupt break occurs with the begin of the war (see Figures 1.3 and 1.4).

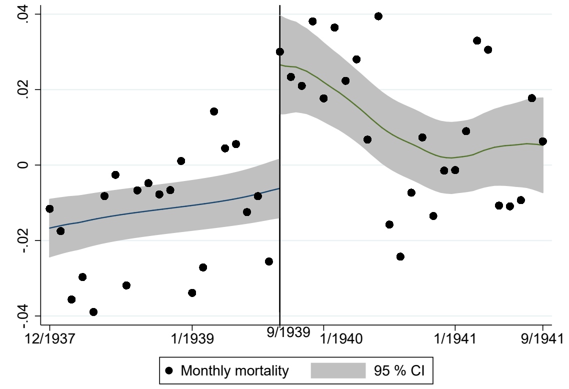
We also test whether the war had an impact on length of stay in hospital measured in days after birth. The probability of observing a mortality event increases mechanically, when mother and infants remain in the hospital for a longer period. However, both before and during the war mothers and infants stayed on average in the hospital for almost 13 days. Finally we look at perinatal mortality. We find the unadjusted perinatal mortality rate to be significantly higher during the war. Descriptive statistics and mean comparisons for miscarriages can be found in Table A.3 and A.4 in the Appendix. Women who suffer a miscarriage are on average older and have more previous pregnancies than women who give birth.

***Figure 1.5:** Raw perinatal mortality by month of birth - All births*



Notes: Perinatal death rates (monthly averaged) and local linear regressions with a ROT bandwidth and an Epanechnikov kernel separately for the pre-war and the war period.

***Figure 1.6:** Adjusted perinatal mortality by month - All births*



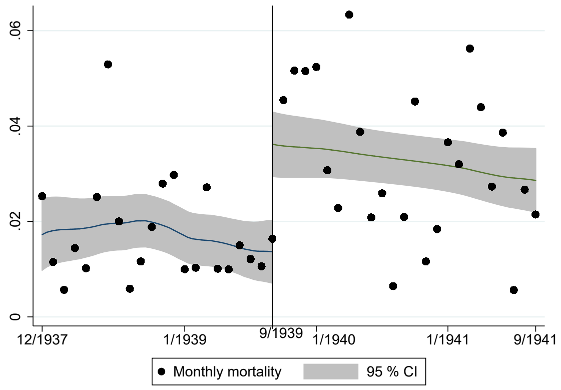
Notes: Regression residuals (monthly averaged) from regressions of perinatal mortality on social status, mother's age, parity, primipara, twinning status, infant's gender, marital status, a dummy for general ward, normal fetal position and working status.

1.3.2 Graphical analysis

We begin our analysis by documenting the effect of WWII on perinatal mortality and health graphically. The monthly trend of perinatal infant mortality is presented in Figure 1.5. The dots denote the raw monthly mortality rate. We fit local linear regressions separately for the pre-war and the war period. The graph documents a significant jump in perinatal mortality in September 1939. During the following months average perinatal mortality decreases gradually, but remains above pre-war levels. In a next step we adjust for observable characteristics. Figure 1.6 displays the monthly averages of residuals obtained from regressions of perinatal mortality on all maternal characteristics given in Table 1.1, infant gender and a dummy for regular fetal position. The jump at the threshold provided by the onset of the war remains significant. The decline in the mortality rate during the war period is slightly more pronounced compared to the graph without adjustment and the mortality rate in 1941 is no longer significantly greater than in the months preceding the war.

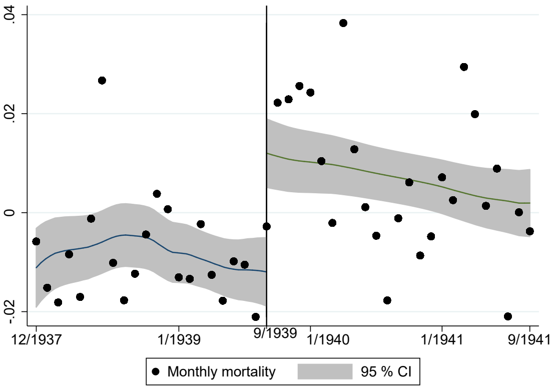
To explore whether the overall increase in perinatal mortality rate is driven by stillborn infants or by live born infants who die in hospital after birth, we repeat the analysis for live births in Figure 1.7 and Figure 1.8. Again we see a significant jump in September 1939 followed by a linear decline in mortality. This suggest that a large part of the overall mortality effect is driven by live born children.

***Figure 1.7:** Raw perinatal mortality by month - Live births*



Notes: Perinatal death rates (monthly averaged) and local linear regressions with a ROT bandwidth and an Epanechnikov kernel separately for the pre-war and the war period for live births.

***Figure 1.8:** Adjusted perinatal mortality by month - Live births*

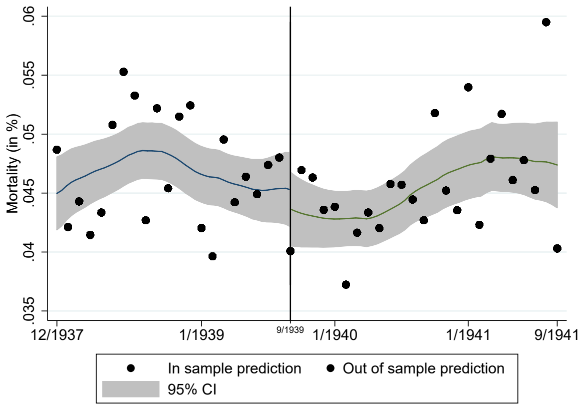


Notes: Regression residuals (monthly averaged) from regressions of perinatal mortality on social status, mother's age, parity, primipara, twinning status, infant's gender, marital status, a dummy for general ward, normal fetal position and working status for live births.

If conditions become worse permanently because of the war, one would expect the effect to stay constant or even accumulate. Our graphical results point to another interpretation. The onset of WWII might have provided a one time shock, which initially led to a jump in perinatal mortality and then gradually faded out. This explanation is consistent with the evidence presented in Section 1.2.1. The onset of the war was unexpected by the general public and affected the daily routine of individuals. Furthermore, a shift of resources towards the military caused turmoil in the unprepared health sector. Yet, prior to 1942 living conditions were not as severe as that it was impossible for individuals and organizations to adapt.

Given the duration of pregnancy, it is unlikely that the composition of mothers changes abruptly around our threshold. Still, we cannot rule out that mothers who give birth during the war are different from mothers who gave birth prior to the war. Therefore we investigate whether changes in observable characteristics can explain the increase in infant mortality. We regress perinatal infant mortality on our control variables using only observations from the pre-war period. We then use the estimated coefficients to predict perinatal infant mortality for the whole sample. If women who give birth during the war, are simply more risky in terms of obvervable characteristics, we would also expect to see an increase in predicted mortality after the onset of the war. The resulting timeline of predicted infant mortality is displayed in Figure 1.9. We do not find any sig-

Figure 1.9: *Predicted mortality*



Notes: Predicted mortality (monthly averages) from regressions of perinatal mortality on social status, mother's age, parity, primipara, twinning status, infant's gender, marital status, a dummy for general ward, normal fetal position and working status.

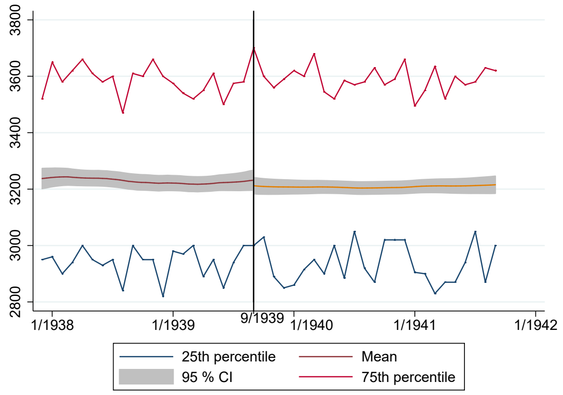
nificant change around the threshold. In fact, predicted mortality is at its lowest level in the last quarter of 1939, the time period right after the onset of WWII. On the other hand we see an increase in predicted mortality after the first quarter of 1940, while actual mortality is decreasing during this time period.

Finally, we turn to measures of perinatal health. Features of the distribution of birth weight are presented in Figure 1.10. Average birth weight stays almost constant during our whole observation period. Rather than on the average birth, war might have an impact on more extreme cases. We add lines of the 25th and 75th percentiles of monthly birth weight to our plot to investigate trends for children with higher or lower birth weight. Again, we do not see any trend. Similarly, kernel estimates of the density of birth weight do not indicate that any part of the distribution of birth weight was affected by the war (see Figure 1.11). Graphs for asphyxia and maturity are given in Figures A.1 and A.2 in the Appendix.

1.4 Empirical strategy

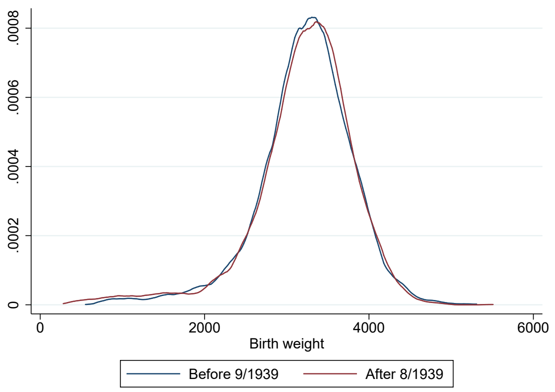
The aim of this work is to estimate the effect of the onset of WWII on perinatal health and mortality of infants. In our identification strategy we exploit the onset of WWII as a natural experiment. There is no evidence that anticipation of a coming war affected fertility patterns before September 1939 (see Section 1.2.1). Hence we argue that the onset of the war constitutes an unexpected shock for women already pregnant

Figure 1.10: *Birth weight by month of birth*



Notes: 25th percentile, mean and 25th percentile of birth weight.

Figure 1.11: *Birth weight distribution*



Notes: Kernel density estimate of birth weight by war status.

in September 1939. After September 1939 fertility decisions may be affected by the war. Therefore we conduct our analysis using both our whole observation period (1/1938-9/1941) and a restricted observation period (12/1937- 5/1940). All full term births that occurred during the restricted observation period were conceived before the onset of the war. However, given that our data do not contain a reliable measure of gestational age, we cannot exclude preterm births conceived during the war period from the restricted sample. Preterm births are associated with a higher risk of perinatal mortality. While preterm birth itself can be a consequence of war, and therefore part of the effect we want to capture with our war dummy, our results will overestimate the true effect on mortality if women with an ex ante high risk of a preterm birth increase their fertility relative to other women during the war. Although we cannot generally rule out such concerns, we argue that an increased share of premature births should be reflected in an on average lower birth weight. Our descriptive analysis of trends in birth weight in Section 1.3.2 does not indicate any change. Additionally we run all our regressions also on a sample restricted to live births, assuming that the share of preterm births is lower among live births.[20]

Our baseline results are obtained estimating the following equation:

$$y_i = \alpha + \beta \texttt{war}_i + \kappa C_i + u_i \tag{1.1}$$

[20] As explained in section 1.3.1 we generally exclude miscarriages.

y_i is the outcome (infant mortality, birth weight, maturity, asphyxia), $\texttt{war}$ is an indicator whether birth took place after the begin of WWII (i.e. in or after 9/1939), and C_i is a set of control variables. Specifically, we control for maternal age, number of pregnancy, a dummy for first pregnancy, (birth of) multiples, infant's sex, whether the mother is married, or working, a dummy for regular fetal position and a dummy for general ward. The coefficient β captures the mean difference between the treatment and the control group conditional on observable characteristics.

In Section 1.3.2 we present graphical evidence that the onset of the war rather than the war as permanent condition constitutes the shock actually driving our results. Therefore, as a next step, we include a time trend and its interaction with the treatment dummy in our regression equation:

$$y_i = \alpha + \delta \texttt{war}_i + \lambda_0 \phi(\tilde{t_i}) + \lambda_1 \phi(\tilde{t_i}) * \texttt{war}_i + \kappa C_i + \pi_i + u_i \tag{1.2}$$

$\tilde{t_i}$ denotes the time trend centered around the onset of the war. In the reported regressions we use a quadratic time trend, such that $\lambda_0 \phi(\tilde{t_i}) = \lambda_{01}\tilde{t_i} + \lambda_{02}\tilde{t_i}^2$.[21] π_i captures seasonality effects. The coefficient δ captures the jump in mortality at the threshold.

As shown in Figures 1.5 to 1.8 the time trend of infant mortality differs between the prewar and the war period. We also saw some differences the composition of treatment and control group in terms of social groups and the war might also change the structural relationship between socio-economic class, observed characteristics and outcomes. For example the war might have increased the mortality risk disproportionally for working mothers. To answer the question, by how much the war increases mortality for those who actually give birth during the war, we additionally estimate an "Average Treatment Effect on the Treated" (ATET) using regression adjustment.[22] This approach is equivalent to estimating Equation 1.2 separately for the treatment and the control group and then taking the difference in predicted outcomes under both sets of estimated coefficients for the treatment group. The ATET is constructed as follows:

$$\gamma^{\texttt{ATET}}_{\texttt{war}} = (\hat{\theta_0}^{\texttt{war}} - \hat{\theta_0}^{\texttt{nowar}}) + \frac{1}{N^{\texttt{war}}} \sum_{\texttt{i in war}} X_i(\hat{\theta}^{\texttt{war}} - \hat{\theta}^{\texttt{nowar}}) \tag{1.3}$$

[21] We also used a linear time trend and obtained similar results.

[22] For an explanation of regression adjustment see for example Uysal (2015).

X_i denotes all controls and the time trend. $\hat{\theta}^{\texttt{war}}$ and $\hat{\theta}^{\texttt{nowar}}$ are the estimated coefficients from the prewar and the war regression. $\gamma^{\texttt{ATET}}_{\texttt{war}}$ measures the average difference between the predicted effect for the treatment group and the predicted treatment effect for the treatment group if the treatment group had given birth before the war.

1.5 Results

1.5.1 Effect of war on perinatal health

Table 1.3, 1.6 and 1.8 present the effect of war on three measures of perinatal health, *birth weight*, *asphyxia* and *infant maturity*. Panel A shows regression estimates using the full sample (i.e. all births excluding miscarriages), while Panel B restricts the sample to live births. Results in columns (1)-(4) are based on the entire observation period from 12/1937-9/1941, whereas columns (5)-(8) use only births likely to be conceived before the onset of WWII. We cluster all standard errors at birth level to adjust for twin births. ATETs estimated for the same outcome variables using regression adjustment are reported in separate tables below (see Tables 1.4, 1.7 and 1.9 respectively). For neither sample we find any effect of the onset of the war on birth weight. The estimated coefficients are small in size and insignificant in all but two specifications. This is in line with the descriptive analysis presented in Section 1.3.2 above.

As intrauterine growth takes place during the whole course of pregnancy, the war

Table 1.3: *Effect of war - Birth weight*

Panel A	All observations				Born before 6/1940			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth after 9/1939	-17.1	-21.3*	-19.8	-28.8	-21.5	-27.3*	-8.74	3.34
	(13.3)	(12.3)	(37.5)	(38.7)	(17.4)	(16.1)	(46.5)	(49.4)
Observations	8820	8361	8361	8361	5942	5624	5624	5624
Panel B	Only live births				Only live births born before 6/1940			
Birth after 9/1939	4.98	-8.24	-3.28	-18.0	-7.40	-18.0	18.8	27.5
	(12.2)	(11.5)	(35.5)	(36.2)	(16.2)	(15.2)	(42.5)	(44.9)
Observations	8472	8069	8069	8069	5717	5433	5433	5433
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Trend	No	No	Yes	Yes	No	No	Yes	Yes
Seasonality	No	No	No	Yes	No	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Controls include social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position and general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

Table 1.4: *Effect of war - Birth weight - Regression adjustment*

	All observations		Observations before 6/1940	
	(1) All births	(2) Live births	(3) All births	(4) Live births
ATET				
Born after 9/1939	-81.5	-43.4	-42.8	-25.9
	(149.4)	(142.4)	(60.5)	(58.1)
Observations	8361	8069	5624	5433

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender, dummy variables for regular fetal position and general ward and a quadratic time trend fitted on each side of the threshold separately.

might manifest itself in lower birth weight only with a delay rather than the day after the war started. Furthermore, if we view the onset of the war as a shock, the impact of this shock may be related to the stage of pregnancy at which it occurred. Therefore we split the treatment variable into four categories, depending on whether the onset of the war occurred during late pregnancy (infants born 9-11 1939), during middle pregnancy (infants born 12/1939-2/1940), during early pregnancy (infants born 3-5 1940) or before the pregnancy even started.

Again our results do not provide evidence for an effect of the onset of WWII on birth

Table 1.5: *Effect of war by time of birth - Birth weight*

Panel A	All observations		Live births	
Born 9-11/1339	10.9	-9.68	31.5	7.39
	(28.8)	(27.2)	(25.7)	(25.2)
Born 12/1939-2/1940	-30.2	-42.0*	-26.3	-40.9*
	(26.4)	(24.5)	(24.7)	(23.3)
Born 3-5/1940	-39.6	-32.2	-21.0	-18.9
	(26.4)	(23.0)	(24.5)	(21.8)
Born after 5/1940	-14.2	-16.1	13.2	-1.08
	(15.4)	(14.1)	(13.9)	(13.0)
Observations	8820	8361	8472	8069
Controls	No	Yes	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position and general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

weight (Table 1.5). In specifications with controls, we see a small significant decrease in birth weight in case of births for which the start of the war fell into the second trimester. However, the level of statistical significance is only at 10% and moreover, we cannot use a reliable measure of gestation in these regressions. Therefore, we are not confident to conclude that the shock provided by the onset of the war reduced birth weight for pregnancies affected in the second semester.

Asphyxia was only consistently recorded after November 1938. Therefore we use a smaller sample when estimating the effects for asphyxia presented in Table 1.6. As in the case of birth weight we find a zero effect.

Results for infant maturity are mixed (see Table 1.8 and 1.9). There is no significant

Table 1.6: *Effect of war - Asphyxia*

Panel A	All observations				Born before 6/1940			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth after 9/1939	0.0028 (0.0039)	0.0034 (0.0040)	0.00064 (0.014)	-0.0061 (0.016)	-0.0018 (0.0044)	-0.00089 (0.0047)	0.012 (0.015)	0.0094 (0.020)
Observations	6784	6440	6440	6440	3906	3703	3703	3703
Panel B	Only live births				Only live births born before 6/1940			
Birth after 9/1939	0.0039 (0.0039)	0.0045 (0.0041)	0.000035 (0.014)	-0.0076 (0.016)	-0.00069 (0.0045)	0.00040 (0.0047)	0.0094 (0.015)	0.0080 (0.019)
Observations	6495	6196	6196	6196	3740	3560	3560	3560
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Trend	No	No	Yes	Yes	No	No	Yes	Yes
Seasonality	No	No	No	Yes	No	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Controls include social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position and general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

Table 1.7: *Effect of war - Asphyxia - Regression adjustment*

	All observations		Observations before 6/1940	
	(1) All births	(2) Live births	(3) All births	(4) Live births
ATET Born after 9/1939	0.072 (0.16)	0.045 (0.17)	0.015 (0.045)	0.0098 (0.046)
Observations	6440	6196	3703	3560

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender, dummy variables for regular fetal position and general ward and a quadratic time trend fitted on each side of the threshold separately.

difference in conditional and unconditional means between the treatment and the control sample. However we find evidence of a drop in the proportion of mature infants at the onset of the war. This may be the result of a higher number of pre-term births during the first months of the war. The estimates for the ATET in Table 1.9 are larger than the estimated regression coefficients.

Table 1.8: *Effect of war - Maturity*

Panel A	All observations				Born before 6/1940			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth after 9/1939	0.00044	-0.0047	-0.058***	-0.054**	0.00048	-0.0062	-0.061**	-0.044
	(0.0076)	(0.0075)	(0.021)	(0.022)	(0.0098)	(0.0097)	(0.025)	(0.027)
Observations	8814	8350	8350	8350	5937	5614	5614	5614
Panel B	Only live births				Only live births born before 6/1940			
Birth after 9/1939	0.0075	-0.00017	-0.051**	-0.051**	0.0055	-0.0020	-0.053**	-0.039
	(0.0073)	(0.0073)	(0.021)	(0.021)	(0.0095)	(0.0095)	(0.024)	(0.025)
Observations	8463	8058	8058	8058	5709	5423	5423	5423
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Trend	No	No	Yes	Yes	No	No	Yes	Yes
Seasonality	No	No	No	Yes	No	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Controls include social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position and general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

Table 1.9: *Effect of war - Maturity - Regression adjustment*

	All observations		Observations before 6/1940	
	(1) All births	(2) Live births	(3) All births	(4) Live births
ATET				
Born after 9/1939	-0.23**	-0.23***	-0.11***	-0.11***
	(0.090)	(0.087)	(0.035)	(0.034)
Observations	8350	8058	5614	5423

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender, dummy variables for regular fetal position and general ward and a quadratic time trend fitted on each side of the threshold separately.

1.5.2 Effect of war on perinatal mortality

We use the same specifications as in the previous subsection to estimate linear probability models for the effect of war on perinatal mortality. The results are presented in Table 1.10 and Table 1.11.

***Table 1.10:** Effect of war - Mortality*

Panel A	All observations				Born before 6/1940			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth after 9/1939	0.025***	0.025***	0.047***	0.048***	0.035***	0.038***	0.035*	0.031
	(0.0053)	(0.0052)	(0.016)	(0.016)	(0.0075)	(0.0073)	(0.020)	(0.021)
Observations	8828	8363	8363	8363	5950	5626	5626	5626
Panel B	Only live births				Only live births born before 6/1940			
Birth after 9/1939	0.016***	0.016***	0.040***	0.040***	0.024***	0.026***	0.030**	0.026**
	(0.0035)	(0.0036)	(0.0099)	(0.010)	(0.0053)	(0.0055)	(0.012)	(0.012)
Observations	8477	8071	8071	8071	5722	5435	5435	5435
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Trend	No	No	Yes	Yes	No	No	Yes	Yes
Seasonality	No	No	No	Yes	No	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Controls include social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position and general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

***Table 1.11:** Effect of war - Mortality - Regression adjustment*

	All observations		Observations before 6/1940	
	(1)	(2)	(3)	(4)
	All births	Live births	All births	Live births
ATET				
Born after 9/1939	0.049	0.086***	0.040*	0.053***
	(0.057)	(0.033)	(0.023)	(0.013)
Observations	8363	8071	5626	5435

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender, dummy variables for regular fetal position and general ward and a quadratic time trend fitted on each side of the threshold separately.

Overall, perinatal infant mortality increases significantly after the onset of WWII. Panel A presents results when the sample is not restricted to live births. Deaths in Panel A of Table 1.10 are therefore made up of stillborn children as well live born children who die in hospital after birth. While 5% of births do not result in a living infant leaving the hospital in the pre-war sample, this number increases to 7.5% in the war sample. Once we do not compare mean differences but the jump at the threshold in Column (3) and Column (4), the effect becomes even stronger. If we restrict the sample and drop all births which took place after May 1940, we see a larger difference in the means but a smaller jump. The ATET is larger than the regression coefficients in size but only significant in the restricted sample.[23] Altogether these results support our interpretation that the onset of the war provided a shock which faded out gradually.

Effect of war by social class

We investigate, whether the effect of war on mortality is heterogeneous with respect to social class. Parental social status is highly predictive of future live outcomes. If the war affects the composition of the population through the channel of selected mortality, this will be reflected in the live outcome of affected cohorts. The results displayed in Table 1.12 are based on specifications, where we omit the overall war dummy. Instead we report the estimated coefficients of interaction terms between the war dummy and the class-indicator for all social classes.

The onset of the war has a non negative effect on mortality for all social groups. Higher professionals and managers - which constitute our highest social class - do suffer from the war, but also do lower skilled workers. There does not seem to be a gradient with respect to social class. Unskilled workers as well as Foremen & skilled workers appear to be most severely affected.

Effect of war by birth weight

Just like social class, birth weight is highly correlated with later life outcomes. If low birth weight infants are more likely to die as a consequence of the war, negative effects of war on later live outcomes will be underestimated in studies based on surviving individuals. In order to explore heterogeneity by birth weight, we split our sample at 2,000 grams, 2,500 grams and 3,000 grams. Table 1.13 displays the estimated treatment effects for all four groups. We find a clear gradient with respect to birth weight in the effect of the war. In any of the specifications, the magnitude of the estimated coefficient of the interaction term between the war dummy and the birth weight-group dummy, the effect increases when birth weight decreases. However, as the number of low birth weight infants is relatively small, we lack the statistical power to detect a significant reduction in mortality for children whose birth weight is below the common low birth

[23] The ATET contrasts predicted outcomes for the group of births that took place during the war with the hypothetical predicted outcomes based on estimated coefficients from the pre-war sample. We saw in Figure 1.9 that births in the first months of the war have a slightly lower predicted mortality risk than pre-war observations, while births later in the war do not. Since mortality rates are higher mainly at the beginning of the war, it is not surprising that we do not find an significant ATET for the whole sample.

Table 1.12: *Effect of war by social class - Mortality*

Panel A	All observations			Born before 6/1940		
	(1)	(2)	(3)	(4)	(5)	(6)
War * Higher managers & professionals	0.0096	0.019	0.042**	0.048*	0.046**	0.040
	(0.015)	(0.015)	(0.021)	(0.026)	(0.023)	(0.029)
War * Lower managers & professionals, cleric	0.032***	0.031***	0.053***	0.023	0.025*	0.019
	(0.012)	(0.011)	(0.018)	(0.015)	(0.014)	(0.024)
War * Foremen & skilled workers	0.027**	0.028***	0.051***	0.055***	0.056***	0.050**
	(0.011)	(0.010)	(0.018)	(0.017)	(0.015)	(0.024)
War * Farmers	0.060**	0.055**	0.077***	0.058*	0.052*	0.046
	(0.024)	(0.023)	(0.028)	(0.031)	(0.029)	(0.036)
War * Lower skilled workers	0.0060	-0.0048	0.018	0.0053	0.0031	-0.0031
	(0.014)	(0.013)	(0.020)	(0.019)	(0.018)	(0.026)
War * Unskilled workers	0.032***	0.029***	0.051***	0.045***	0.044***	0.038
	(0.011)	(0.010)	(0.019)	(0.016)	(0.015)	(0.025)
War * Farm workers	0.034	0.0075	0.031	0.058	0.051	0.046
	(0.039)	(0.036)	(0.039)	(0.068)	(0.064)	(0.067)
Observations	8500	8363	8363	5729	5626	5626
Panel B	Live births			Live births born before 6/1940		
War * Higher managers & professionals	0.0076	0.010	0.034**	0.018	0.018	0.018
	(0.011)	(0.013)	(0.016)	(0.018)	(0.018)	(0.020)
War * Lower managers & professionals, cleric	0.018**	0.018**	0.041***	0.017	0.018*	0.019
	(0.0082)	(0.0081)	(0.012)	(0.011)	(0.010)	(0.016)
War * Foremen & skilled workers	0.019**	0.022***	0.046***	0.037***	0.040***	0.040***
	(0.0074)	(0.0072)	(0.012)	(0.012)	(0.012)	(0.016)
War * Farmers	0.039**	0.031*	0.055***	0.031	0.018	0.018
	(0.016)	(0.016)	(0.018)	(0.021)	(0.017)	(0.020)
War * Lower skilled workers	0.0065	0.0055	0.029**	0.012	0.013	0.013
	(0.0091)	(0.0090)	(0.013)	(0.014)	(0.013)	(0.017)
War * Unskilled workers	0.014**	0.014**	0.038***	0.029**	0.029***	0.029**
	(0.0069)	(0.0068)	(0.012)	(0.011)	(0.011)	(0.015)
War * Farm workers	0.016	0.0093	0.034	0.052	0.046	0.048
	(0.025)	(0.023)	(0.026)	(0.059)	(0.053)	(0.055)
Observations	8164	8071	8071	5512	5435	5435
Controls	No	Yes	Yes	No	Yes	Yes
Trend	No	No	Yes	No	No	Yes
Seasonality	No	No	Yes	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position and general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

weight threshold at 2,500 grams but above 2,000 grams. For very low birth weight children with less than 2,000 grams at birth, the effect is largest. The probability to leave the hospital alive decreases by more than 10 percentage points.[24] Also children born between 2,500 and 3,000 grams are affected to a larger extent than the group of children above 3,000 grams.

[24] A surprisingly large number of infants below 2,000 grams survives. We checked the most extreme cases in the birth records carefully but found no sign of misreporting. In one case we found a letter stating that a child born at around 1,300 grams had left the hospital and was doing well.

Table 1.13: *Effect of war by birth weight - Mortality*

Panel A	All observations			Born before 6/1940		
	(1)	(2)	(3)	(4)	(5)	(6)
War*Birth weight below 2000 grams	0.13**	0.14**	0.15**	0.18***	0.19***	0.18***
	(0.055)	(0.058)	(0.059)	(0.062)	(0.065)	(0.066)
War*Birth weight 2000-2499 grams	0.054	0.059	0.068*	0.049	0.054	0.043
	(0.036)	(0.036)	(0.038)	(0.048)	(0.048)	(0.051)
War*Birth weight 2500-3999 grams	0.039***	0.036***	0.046***	0.045***	0.046***	0.035
	(0.010)	(0.010)	(0.016)	(0.015)	(0.015)	(0.022)
War*Birth weight 3000 grams and above	0.0049	0.0087**	0.019	0.012**	0.017***	0.0065
	(0.0040)	(0.0040)	(0.013)	(0.0057)	(0.0058)	(0.017)
Observations	8820	8361	8361	5942	5624	5624
Panel B	Live births			Live births born before 6/1940		
War*Birth weight below 2000 grams	0.14*	0.15*	0.16**	0.25***	0.24***	0.24***
	(0.075)	(0.077)	(0.077)	(0.086)	(0.088)	(0.088)
War*Birth weight 2000-2499 grams	0.022	0.025	0.041	0.050	0.045	0.048
	(0.028)	(0.028)	(0.029)	(0.041)	(0.041)	(0.040)
War*Birth weight 2500-3999 grams	0.022***	0.020***	0.036***	0.035***	0.034***	0.036***
	(0.0068)	(0.0069)	(0.010)	(0.011)	(0.011)	(0.013)
War*Birth weight 3000 grams and above	0.0083***	0.0088***	0.025***	0.0080***	0.0098***	0.012
	(0.0021)	(0.0023)	(0.0080)	(0.0030)	(0.0033)	(0.0096)
Observations	8472	8069	8069	5717	5433	5433
Controls	No	Yes	Yes	No	Yes	Yes
Trend	No	No	Yes	No	No	Yes
Seasonality	No	No	Yes	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position and general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

1.5.3 Robustness

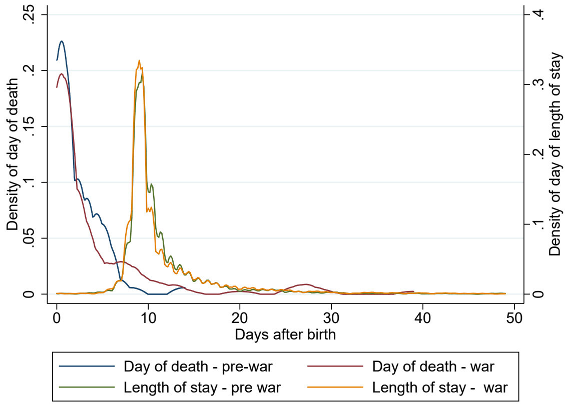
Length of stay

While the conventional definition of neonatal mortality includes deaths up to 28 days after birth (WHO 2006), we only observe newborns until they leave the hospital. As long as the day of discharge and the treatment are independent, our definition of infant mortality will not pose a thread to identification. Figure 1.12 shows the distribution of the length of stay in hospital after birth and length of life in days for live born children separately for the pre-war and the war period.[25]

First we notice that there is hardly any difference in the distribution of the length of stay in hospital after birth in our treatment and control group. Most observations stay in hospital for around 9-10 days after birth and only 1.5% of live born children are discharged before completing the first week of life. Neonatal deaths on the other hand

[25] To facilitate legibility, we exclude a small number of observations who stayed in hospital for more than 50 days.

Figure 1.12: *Length of stay and day of death*



Notes: Distribution of length of stay in hospital and length of life in days for live born children.

mostly occur within the first four days after birth. Since mothers received postnatal care in hospital, the death of an infant does not automatically lead to a discharge of the mother. As a robustness check we estimate the regression models used for analysis with a modified versions of infant mortality. We define an infant to have died if the death occurred either in the first 5 days (see Table 1.14) or the first 7 days (see Table 1.15) after birth. In these specifications we exclude all observations which left the hospital before that specific day. Although the coefficients become smaller in size, we still see a significant effect of the onset of the war on perinatal infant mortality.

Temperature

In the first two months of 1940, Munich was hit by a particularly low temperatures (Stadtarchiv München 1939-1940). In order to rule out, that the effect we measure is in fact a shock caused by low temperatures, we include the average monthly temperatures in Munich as additional control variables. The results are presented in Table 1.16. The estimated coefficients hardly change compared to the baseline estimates. This suggests that temperature does not confound our baseline estimates.

Structural break

In our empirical specification we estimate infant mortality as a function of maternal characteristics and time variables. In the regression adjustment we allow this function to differ between the pre-war and the war period. To investigate whether such a struc-

Table 1.14: *Effect of war - Mortality - Death within 5 days*

Panel A	All observations				Born before 6/1940			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth after 9/1939	0.019***	0.019***	0.036**	0.036**	0.025***	0.026***	0.038*	0.033
	(0.0050)	(0.0050)	(0.015)	(0.015)	(0.0069)	(0.0067)	(0.020)	(0.021)
Observations	8762	8305	8305	8305	5907	5589	5589	5589
Panel B	Only live births				Only live births born before 6/1940			
Birth after 9/1939	0.0085***	0.0086***	0.026***	0.024***	0.011**	0.012***	0.031***	0.024**
	(0.0031)	(0.0032)	(0.0090)	(0.0090)	(0.0043)	(0.0044)	(0.011)	(0.011)
Observations	8426	8023	8023	8023	5689	5404	5404	5404
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Trend	No	No	Yes	Yes	No	No	Yes	Yes
Seasonality	No	No	No	Yes	No	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, parity, primipara, twinning status, infant's gender, marital status, a dummy for general ward and working status; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth. Only observations staying in the hospital at least five days.

Table 1.15: *Effect of war - Mortality - Death within 7 days*

Panel A	All observations				Born before 6/1940			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth after 9/1939	0.020***	0.020***	0.039**	0.038**	0.028***	0.030***	0.034*	0.029
	(0.0050)	(0.0050)	(0.015)	(0.016)	(0.0070)	(0.0069)	(0.020)	(0.021)
Observations	8669	8219	8219	8219	5836	5523	5523	5523
Panel B	Only live births				Only live births born before 6/1940			
Birth after 9/1939	0.0093***	0.0097***	0.029***	0.026***	0.014***	0.015***	0.029**	0.021*
	(0.0032)	(0.0033)	(0.0093)	(0.0094)	(0.0046)	(0.0047)	(0.011)	(0.012)
Observations	8343	7944	7944	7944	5625	5344	5344	5344
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Trend	No	No	Yes	Yes	No	No	Yes	Yes
Seasonality	No	No	No	Yes	No	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, parity, primipara, twinning status, infant's gender, marital status, a dummy for general ward and working status; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth. Only cases staying in the hospital at least seven days.

Table 1.16: *Effect of war - Mortality - Robustness check: Monthly temperature*

Panel A	All observations				Born before 6/1940			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth after 9/1939	0.025*** (0.0053)	0.025*** (0.0052)	0.044*** (0.016)	0.047*** (0.016)	0.035*** (0.0075)	0.038*** (0.0075)	0.035* (0.020)	0.031 (0.021)
Observations	8828	8363	8363	8363	5950	5626	5626	5626
Panel B	Only live births				Only live births born before 6/1940			
Birth after 9/1939	0.016*** (0.0035)	0.016*** (0.0036)	0.038*** (0.010)	0.040*** (0.010)	0.024*** (0.0053)	0.026*** (0.0056)	0.030** (0.012)	0.026** (0.012)
Observations	8477	8071	8071	8071	5722	5435	5435	5435
Temperature	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Trend	No	No	Yes	Yes	No	No	Yes	Yes
Seasonality	No	No	No	Yes	No	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position, general ward and the average temperature in Munich for the current month; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

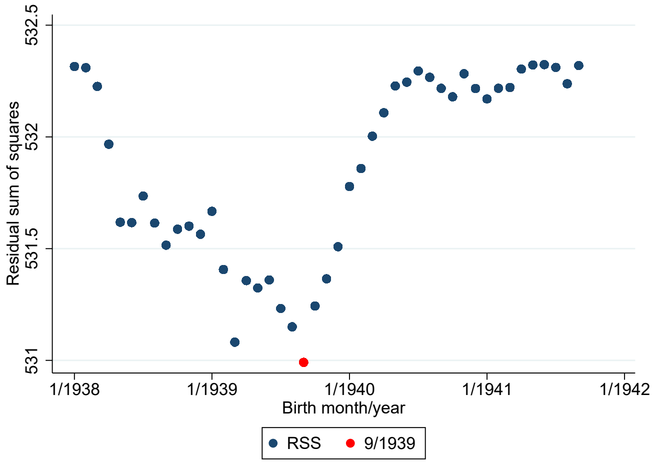
tural break actually took place in September 1939, we investigate whether allowing for a structural break in September 1939 leads to a better fit than allowing for a structural break in any other month. For each month between January 1938 and September 1941, we estimate Equation 1.1 (without the war dummy) separately to both sides of the respective month. We calculate the total residual sum of squares, that is the sum of residuals sum of squares of models from either side of the threshold. If no structural break occurred during our period of observation, the total residual sum of squares would not exhibit any systematic pattern (Hansen 2001). However, Figure 1.13 depicts a clear trend. The residual sum of squares decreases when shifting the separating month from January 1938 to September 1939 to reach a minimum in September 1939. When shifting the separating month further into the war period, the residual sum of squares increases again. This indicates that the begin of WWII indeed marked a breakpoint, changing the relationship between maternal characteristics and infant mortality.

1.5.4 Mechanisms

In our setting, we can rule out direct effects of the war like hunger, bombing or displacement.[26] Furthermore, archival records of the hospital do not indicate any problems with the catering of patients or any shortage of fuel or pharmaceuticals. In order to explain

[26] The severe food crisis only started towards the end of WWII and there was not yet any military action in Munich (see section 1.2.1.

Figure 1.13: *Structural break analysis - Mortality*

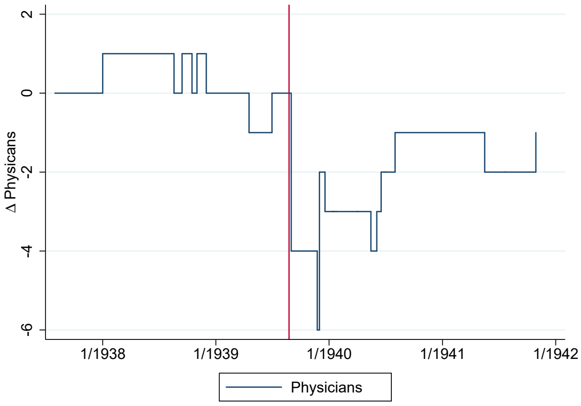


Notes: Residual sum of squares for infant mortality. We estimate the regression model $\texttt{Death}_i = \texttt{Controls}_i\beta + \epsilon_i$ separately for all births prior to month m and births in month m or later. m is shifted from January 1938 to September 1941. RSS denotes the combined sum of residual sum of squares.

the increase in perinatal mortality, we focus on two potential channels already present in fall of 1939 - maternal stress and a decline in the quality of medical care.

Firstly, for the local population the onset of the war came along with changes in the daily routine: the economy was transformed into a planned war-time economy and and conscription took a large number of men away from their families. All these factors are likely to contribute to a feeling of uncertainty and to elevate stress levels among pregnant women. Uncertainty and maternal stress during pregnancy have been shown to affect a newborn's health negatively in the short-run (see Bozzoli and Quintana-Domeque 2014; Carlson 2015; Currie and Rossin-Slater 2013), and might also drive mortality in our setting. Secondly, the onset of the war did not only put a strain on individuals, but the conscription of experienced physicians led to staff shortage in the hospital. This phenomena was not restricted to the hospital *Frauenklinik Maistrasse*, but shared by family practices and hospitals in Munich and elsewhere (Christians 2013, p. 243; Miller 1964, p. 29; Eckart et al. 2006, pp. 26,868). To our knowledge, all physicians working at the hospital *Frauenklinik Maistrasse* in August 1939 were male and therefore potentially liable to military duty. In frequent letters to the state administration, the director of the hospital raised alarm. He warned that, the hospital routine threatened to break down and proper patient care was in jeopardy. To aggravate the situation, conscription foremost targeted experienced doctors. Local hospitals were supposed to fill vacant positions temporarily with inexperienced graduates and - often unpaid - trainees. Fierce

***Figure 1.14:** Staff changes in hospital*



Notes: Number of physicians from December 1937 to September 1941 normalized to zero at the onset of the war. In- and outflows are reconstructed from letters between the hospital administration and several official government bodies archived at the Archives of the LMU and the Archives of the Bavarian State.

competition among local hospitals about these replacements, led to a high turnover in staff. Especially when the first physicians left in 1939, it came as an unexpected shock for the hospital, while it could prepare for further drafts. On the other hand, there is no indication for a shortage of midwives or nurses in letters or archival records.[27]

While clinical records and a large number of letters and official documents have been preserved in the hospital or state archives, staff records have not. Therefore it is impossible for us to reconstruct the in- and outflow of physicians exactly. In order to give an approximate picture of the staff situation over time, we combine information from letters and documents found across various hospital, university and state archives. These documents typically do not mention the overall number of physicians, but tell the date of conscription. Figure 1.14 shows our reconstructed timeline of how the number of physicians evolves over time. We normalize the stock of doctors at the beginning of the war to zero. Right after the onset of, the number of physicians drops by four. After several weeks, the hospital is able to find replacements and the number of physicians increases again. After mid 1941 the number of physicians working at the hospital falls below the number of physicians in 1939 by one or two. As replacements were typically less experienced than the actual physicians, these numbers do not necessarily reflect the quality of medical care.

[27] In the hospital nursing care was exclusively provided by nuns.

We cannot fully quantify these mechanisms. However, in the following we argue that our results are mainly driven by a decline in medical quality. We show that the mortality effect is stronger, where medical quality should matter and furthermore we document a change in provision of certain medical procedures.

Unlike birth weight which is measured at birth and miscarriage, survival of life born children is partly under the control of the medical personnel. If live born children are disproportionally affected by the war, this will hint to a decline in medical care. Maternal stress on the other hand should lead to an increase in stillbirths and miscarriages. Panel B of Table 1.10 conducts the regression analysis for the sample of live born children. Given the low baseline mortality of live born children before the war of 1.8 % the effect of the war is surprisingly large. Between 9/1939 and 5/1940 the mortality of live born children almost doubles compared to the prewar period. Again we find that the jump around the threshold is larger than the differences in means.

The proportion of stillborn children also increases after the onset of WWII (see Panel A of Table 1.17). However, the effect is less than the increase in mortality of live born children and not robust to the inclusion of a time trend and seasonality effects. Therefore our overall effects seem to be driven by children who die in hospital after birth. While we did exclude miscarriages from the main analysis, we estimate the effect of war on the probability of miscarriage in Panel B of Table 1.17. We do not find any evidence that the onset of WWII lead to an increased number of miscarriages.

Table 1.17: *Effect of war - Mortality - Non-livebirths*

Panel A: Stillbirth	Births				Births before 6/1940			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Birth after 9/1939	0.0098**	0.013***	0.012	0.012	0.012**	0.015***	0.014	0.013
	(0.0042)	(0.0043)	(0.014)	(0.014)	(0.0057)	(0.0058)	(0.018)	(0.019)
Observations	8828	8499	8499	8499	5950	5728	5728	5728
Panel B: Miscarriage	All observations				Observations before 6/1940			
Birth after 9/1939	0.0074	0.0036	-0.018	-0.023	-0.015*	-0.015*	0.018	0.013
	(0.0065)	(0.0066)	(0.021)	(0.021)	(0.0081)	(0.0081)	(0.025)	(0.027)
Observations	10022	9617	9617	9617	6689	6416	6416	6416
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Trend	No	No	Yes	Yes	No	No	Yes	Yes
Seasonality	No	No	No	Yes	No	No	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; All regressions include the following controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status and a dummy variable for general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by a set of month dummies.

A shortage of physicians is likely to shift work from physicians to female midwives. Whereas midwives are able to supervise normal deliveries, only physicians can carry out surgeries like caesarean sections. We test whether women who should receive a caesarean section by modern standards are less likely to receive a caesarean section during the first months of the war. We construct a measure of whether a women has an indication for caesarean section based on a guideline described in Mylonas and Friese (2015).[28] As shown in columns (1)-(2) of Table 1.18, the proportion of women with an indication for caesarean section does not change with the onset of the war. However, women with an indication are less likely to actually have a section performed. Instead we see the performance of another procedure. Symphysiotomy is an operation to widen the pelvis that can be carried out by non-specialist doctors and experienced midwives (see Monjok et al. 2012). It was frequently used in the 19th century, when caesarean section was a high risk for mothers. Due to negative consequences for maternal health today's WHO guidelines recommend the use of symphysiotomy only, when safe caesarean sectio is not available (WHO 2003). This result shows that the hospital replaced procedures in need of an experienced surgeon by simpler procedures. We also investigate how the use of medical procedures changes in less severe cases. We look at episiotomy, a simple procedure to prevent perineal tear. While perineal tear can be painful for the mother, it is not a live threatening condition. Columns (7)-(8) of Table 1.18 show a small decrease in the use of this procedure, but this is not reflected in a higher incidence of perineal tear.

[28] We assume a women to have an indication if one of the following conditions is present: Non regular fetal position, eclampsia, placenta previa, disproportion of pelvis and child, uterine rupture. We do not include the condition umbilical cord prolapse, since none of the cases with umbilical cord prolapse is treated with caesarean section in our sample.

Table 1.18: *Effect of war - Medical procedures - Births before 6/1940*

	Indication		Caesarean sectio		Symphysiotomy		Episiotomy		Perineal tear	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Birth after 9/1939	-0.00044	-0.013	0.00021	0.017*	-0.0014**	0.0093	-0.0054	-0.031**	-0.0096	-0.014
	(0.0040)	(0.013)	(0.0013)	(0.0090)	(0.00066)	(0.0075)	(0.0042)	(0.015)	(0.0088)	(0.025)
Indication for caesarean			0.22***	0.22***	0.093***	0.093***				
			(0.041)	(0.041)	(0.029)	(0.029)				
War * Indication			-0.046*	-0.046*	0.045*	0.044*				
			(0.028)	(0.028)	(0.024)	(0.024)				
Observations	5626	5626	5626	5626	5626	5626	5626	5626	5626	5626
Trend + Seasonality	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: (Clustered) Standard errors in parentheses; Significance levels: $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$; All regressions include the follwing controls: Social status, mother's age, marital status, working status, parity, primipara, twinning status, infant's gender and dummy variables for regular fetal position and general ward; Trend denotes a quadratic time trend fitted on each side of the threshold separately; Seasonality is captured by quarter of birth.

1.6 Conclusion

In this work, we investigate the effects of the onset of WWII on health at birth and perinatal mortality. We use a unique data set of historical birth records from Munich's largest birth hospital covering the period 1937-1941. Exploiting the onset of WWII as natural experiment, we show that the onset of the war had no effect on health at birth measured by birth weight, maturity and asphyxia. However, we document an increase in perinatal mortality. This effect is strongest at the beginning of the war and fades out gradually. Additional analyzes reveal that perinatal mortality increases after the begin of the war for all social classes and especially for newborns below 2000 grams.

Since the data cover only the onset of WWII, we can rule out direct effects of the war, like hunger, bombings or flight in our setting. We discuss two potential mechanisms to explain the increase in mortality. On the one hand increased uncertainty and conscription of husbands are likely to increase stress levels of pregnant women and may therefore lead to this mortality increase. On the other hand according to letters from the head physician the conscription of experienced physicians led to severe staff shortage and later to a decrease in the quality of medical care due to the replacement with untrained medical students. To evaluate the importance of each mechanism we investigated whether war affected the proportion of women with an indication for a cesarean section, which was not the case. However women with an indication, are less likely to actually have a sectio performed. Instead the probability of other less complicated birth procedures increases which can be performed by auxiliary medical staff, but which are less safe. In combination all these results point to the deterioration of the quality of medical inputs (i.e. doctors) as the main driver of our results.

Chapter 2

Does the wall still exist? Health differences between East and West Germans

2.1 Introduction

Health differs across countries and populations. Among poor countries a higher national income often comes along with better health. However, the notion that richer countries are also healthier countries does not hold up any longer once national income exceeds a certain level (Deaton 2003; Marmot 2015). For example, life expectancy of Americans falls below the life expectancy of Japanese by more than four years, despite the US having a 40% higher GDP per capita than Japan. Also Mediterranean countries like Greece and Italy surpass the US in terms of life expectancy.[1] Cutler et al. (2006) reject the idea of income being the main driver of health inequalities between rich and poor countries. Instead they emphasize "...*institutional ability and political willingness to implement known technologies, ...*"

Disentangling the role of institutional capability from other factors which contribute to inequalities in health across countries is not an easy task. Institutions and culture are highly interdependent (Alesina and Giuliano 2015). And cultural norms provide a very powerful explanation as to why differences in health exist. So has "*being culturally Japanese*" been found to protect Japanese migrants to California from heart disease (Marmot 2015, Chapter 1) and the health benefits of a traditional Mediterranean diet have made it into conventional wisdom.

[1] GDP 2015, USD, constant prices, 2010 PPPs, data from OECD (2017); Life expectancy 2015, data from WHO (2016, Annex B).

In this study I explore the long lasting effect of institutions on health in a setting where cultural differences are either minor or caused by the institutions. Specifically, I exploit the German separation and reunification as a natural experiment to learn about the long term effects of living under a Socialist regime on individual health. Making use of rich microdata from a large German panel study I investigate how the German separation continues to contribute to health inequalities more than two decades after reunification. Identification rests on the assumption that East and West Germany would not systematically differ in the absence of a separation. Previous literature has supported this claim, for example Alesina and Fuchs-Schündeln (2007) show that East- and West Germany were similar in terms of income in the period between WWI and WWII.

My results show that individuals who lived under a socialist regime in East Germany prior to 1989, continue to be disadvantaged in terms of health more than two decades after reunification. Health inequalities between East and West Germans are mostly driven by individuals who I observe at older ages. For this group I document a robust East-West gap that is particularly persistent for health measures that deteriorate as individuals age.

While individuals from East and West Germany have been subject to the same political system and health system since 1990, living conditions between the two parts of the country yet have to fully converge. By 2018 East Germans still have lower incomes than West Germans and they are more likely to be unemployed. Besides, East Germans differ in personality form their West German counterparts (Friehe et al. 2015) and displayed different patterns of food consumption after reunification (Dragone and Ziebarth 2017). Given these persistent differences, it seems unlikely that today's gap in health status can be attributed to experiences prior to reunification alone. In order to obtain estimates of disparities between East and West net of contemporary factors influencing health, I apply the mediation analysis framework outlined in Acharya et al. (2016). I estimate the controlled direct effect of having lived in the East prior to 1989, keeping the level of present day determinants of health, like income, unemployment, health behavior and locus of control constant. While removing the influence of contemporary factors reduces the magnitude of estimates of health inequalities between East and West, a significant gap remains for older individuals.

During the 41 years of separation, East and West Germany varied along a wide range of dimensions likely to influence health. For example, the health system of the former East emphasized the role of prevention, but was heavily underfunded and lacked modern equipment. Furthermore, the East was far more affected by environmental pollution and living under a repressive political regime was likely to increase stress levels.

This study is not the first to point out differences in health between East and West Germans. Demographers have studied the fact that differences in mortality between East and West Germans peaked at the time of reunification and have been declining since (Gjonça et al. 2000; Nolte et al. 2000). By 2008 East Germans had gained up to six years of additional life expectancy as a consequence of the reunification (Vogt 2013). Nolte et al. (2002) credit advances in medical care for these improvements. The increasing availability of nursing care in East Germany may also be a major factor behind the convergence in mortality (Luy 2004).

Whilst East Germans have been able to catch up in terms of life expectancy, medical studies continue to report discrepancies between East and West. East German regions have been found to exhibit a higher incidence of hypertension (Diederichs and Neuhauser 2014; Neuhauser et al. 2017) and diabetes (Heidemann et al. 2017; Kroll and Lampert 2010; Schipf et al. 2012), as well as higher mortality from ischemic heart disease (Müller-Nordhorn et al. 2004) and a higher frequency of limitations to normal activities due to health problems (Lippe et al. 2017). On the other hand, the prevalence of asthma (Steppuhn et al. 2017) and allergies of individuals born prior to reunification (Krämer et al. 2015) is higher in West Germany and East German men report fewer diagnoses of depression (Thom et al. 2017). Economists have also contributed to this literature. Eibich and Ziebarth (2014) conduct a spatial analysis of health in Germany and find counties located in East Germany to have worse average health. The authors equate this effect to an age effect of up to 5 years for a 40-year old.

I extend these studies in several ways. As living under a socialist regime might not affect all aspects of health in the same way, I consider four dimensions of health as outcomes: Self assessed health, the mental health and physical health summary scores of the SF12-questionnaire and a summary measure of of self-reported medical diagnoses. Secondly I use residency prior to reunification instead of current residency to classify individuals into East and West Germans. This procedure makes my results less sensitive to the effects of East-West migration.

Moreover, I document health differences between East and West Germans across time, age and cohorts. Any gap in health between East and West Germans is not necessarily constant over time. As the separation moves further into the past, initial differences might become weaker or fade away completely. Furthermore health shocks are not uniformly distributed over the life cycle. Therefore disparities may depend on the age at which health is measured. Finally life experiences vary across cohorts and heterogeneity in life experiences might translate into heterogeneous East-West gaps in health.

I also contribute to the understanding of the long lasting legacy of living under the Socialist regime in East Germany on health by using a formal mediation analysis framework, designed to eliminate the influence of post-treatment confounders from the estimated effect. When assessing the importance of post-treatment mechanisms, applied empirical work often confines to including hypothesized channels as additional control variables in the model. This approach can induce serious bias (Acharya et al. 2016). I estimate the controlled-direct effect instead, a well defined quantity corresponding to the treatment effect in an experiment where both the treatment and and post-treatment are manipulated by the experimenter.

Overall, my results suggest a strong and robust health disadvantage of East Germans born prior to the separation. While I cannot fully disentangle age and cohort effects, I document evidence consistent with the interpretation of an acceleration of the ageing process of East Germans between the ages of 40 and 60. I also argue, that those earlier cohorts were hit harder than younger cohorts by the shortcomings of the GDR health system. The rest of the paper proceeds as follows. In Section 2.2 I give information on the institutional background. Section 2.3 presents the data. In Section 2.4 I document differences in health between East and West Germans across time, age and birth cohorts. In Section 2.5 I estimate controlled direct effects and Section 2.6 concludes.

2.2 Institutional background

2.2.1 The German separation and reunification as natural experiment

The German separation followed World War II. After being defeated in May 1945 Germany was partitioned into four zones of occupation lead by the US, France, the UK and the Soviet Union. Soon, conflicting interests between the Soviet Union and its former allies started to emerge. Consequently the year 1949 saw the creation of two German states. The Federal Republic of Germany (FRG, West Germany) in the West was modeled after other Western democracies and integrated into the capitalist economic system of the Western hemisphere. The German Democratic Republic (GDR, East Germany) on the other hand was a socialist state closely aligned with the Soviet Union.

Politically the GDR was characterized by the dominance of a single party. The Socialist Unity Party had full control over all aspects of the country.[2] Especially during the 1950s and 1960 a majority of the population disapproved of the regime (Weber 2011) as people witnessed repression by a *Stalinist regime* (Bouvier 1999) and the forced transition into a socialist planned economy. East Germans opted for migration to West Germany in large numbers. This development came to an end when the regime erected the Berlin Wall in 1961. The wall was part of a rigid border regime making attempts to leave the country a dangerous endeavor. During the late 1960s and 1970s the situation somewhat consolidated. People arranged themselves with a system that traded political conformity with career opportunities and access to goods. Furthermore the State Security Service (Stasi) - which at one time employed one in 60 adults as unofficial collaborators - infiltrated large parts of the country, making it risky to voice dissent even in private settings.

In terms of Economic success and standard of living the GDR clearly lagged behind the FRG. Weekly working hours in the East exceeded those in the West by more than five hours, while household incomes reached barely half of Western levels (Richter 2009).

In 1989 dissatisfaction among the population turned into massive protests which eventually resulted in the fall of the Berlin Wall in 1989 and the German reunification in 1990. East Germany was fully integrated into the political, economic and institutional

[2] The Socialist Unity party controlled not only the executive and legislative branch of government but also the justice system and media. It granted access to higher education and held a monopoly on public opinion (Richter 2009).

system of the FRG.

A growing literature in economics argues that the German separation and reunification provide a valid natural experiment. This conclusion commonly rests on pre-war comparisons of East and West Germany. Alesina and Fuchs-Schündeln (2007) show that East- and West Germany were similar in terms of income in the period between WWI and WWII and already earlier a comparable share of the workforce was employed in industry, agriculture and commerce in both areas. Analyzing Prussian data and statistics from the yearbook of the German statistical office (1936), Görges and Beblo (2015) detect no differences between East and West German districts with respect to school enrollment, literacy as well as marriage and absolute fertility patterns. While industrialization happened faster in the West, the authors document partial convergence until 1933. An additional argument for the validity of the German separation as natural experiment is that the exact line of the inner German border was the outcome of bargaining among the US, the UK and the Soviet Union and unrelated to regional characteristics (Alesina and Fuchs-Schündeln 2007; Friehe et al. 2015).

The separation of Germany lasted for 41 years. Whereas long enough to leave a lasting impact, this time-span only covers a certain period of an individuals life-cycle allowing researchers to exploit heterogeneity across cohorts. Fuchs-Schündeln and Schündeln (2005) use the German reunification to validate the consumption life-cycle model. They show that observed saving rates in East and West Germany are consistent with a consumption life-cycle model that incorporates a precautionary saving motive. Fuchs-Schündeln and Masella (2016) study the effect of socialist education on labor-market outcomes. Exploiting heterogeneity among birth cohorts induced by a cut-off in school entry, they find a additional year of socialist education to reduce the probability of obtaining a college degree. Görges and Beblo (2015) investigate the nurture effect of political regimes on gendered work preferences. They show that among individuals who spent adolescence in separated Germany, gendered work preferences converge in the East while they diverge in the West. Friehe et al. (2015) analyze the impact of living in East Germany on personality traits and conclude that length of exposure to the system of GDR is an important determinant of this relationship.

2.2.2 Health and institutions

Since the seminal paper by Grossman (1972) economists have treated health as an investment good that depreciates with age (Humphreys et al. 2014). Individuals maximize their individual utility by allocating a certain share of their time and resources to medical and non-medical investments in health. The health production function determines how these investments, together with the past stock of health and other individual or environmental factors, translate into current health status. I consider a health production function of the following form:

$$H_t = f(H_{t-1}, M_t, I_t, P_t, S_t, L_t) \tag{2.1}$$

where H_{t-1} denotes the past stock of health, M denotes a vector of medical inputs, I is non-medical input i.e. sport and a healthy diet, P stands for environmental pollution, S denotes factors that determine the socioeconomic status like income and unemployment and L stands for non-cognitive skills or personality traits.

Under this framework two mechanisms will cause the post-reunification health of East and West Germans to differ. Firstly, differences in pre-reunification inputs continue to shape present day health via the lagged stock of health and secondly present day inputs may vary between East and West Germans as a consequence of the separation. Prior to reunification differences in the institutional settings of East and West Germany lead to different inputs into the health production function. Consider for example medical inputs M. Both the GDR and the FRG had universal health coverage such that all citizens had access to health care without significant out-of pocket expenditures. However, especially during the later years, the health system of the GDR was heavily underfunded and lacked modern equipment. Shortcomings in medical capacity in East Germany have been associated with higher infant mortality and undertreatment of hypertension (Busse and Riesberg 2004).

Secondly, environmental regulations were largely absent in the GDR. Especially the area around Leipzig was affected by heavy air pollution and the situation only improved after 1990 (Luechinger 2009).

Finally, East Germans who did not conform to the state were the target of outright political repression. Between 1949 and 1989 about 250,000 individuals were imprisoned for political reasons (Weber 2011). Political imprisonment in East Germany has been associated with long-term impairments in mental health and increased levels of anger (Bauer et al. 1993; Schützwohl and Maercker 2000). An increased prevalence of psychological disorders has been reported among individuals subject to less severe repressions (Spitzer et al. 2007). On the other hand the distribution of wealth and income was more equal in the East than in the West. Higher equality is often associated with better health although there seems to be no direct link (Deaton 2003).

Even today East and West Germany have not yet fully converged along many dimensions, which translates into differences in contemporary inputs into the health production function. More than 25 years after reunification the East of Germany has not managed to achieve the same level of economic prosperity as the West. While the nature of the relationship between income and health is not yet fully understood, income and health within countries are strongly correlated (Deaton 2003; Kawachi et al. 2010). If income has a economically significant causal effect on health, contemporary economic differences between East and West will explain some part of the gap in health. Moreover, the economy of the GDR suffered from low productivity and following the reunification many plants were shut down. As a result unemployment in East Germany remains high, again providing a potential channel explaining health disparities between East and West.

The theory has been put forward that good health is connected to a feeling of being in control (see for example Marmot 2015, Chapter 1). In East Germany individuals handed a major part of control over their lives over to the state which decided over educational opportunities, distribution of housing and even the allocation of goods such as cars or holidays. The turmoil of the reunification may also have contributed to a feeling of loss of control. East Germans have in fact been shown to possess a lower locus of control than West Germans (Friehe et al. 2015).

Lastly health behaviors have been shown to be a significant factor when explaining health differentials. Dragone and Ziebarth (2017) show that East and West Germans displayed different patterns of food consumption after reunification and Eastern Germans gained more weight.

2.3 Data, sample and variables

2.3.1 Data and sample

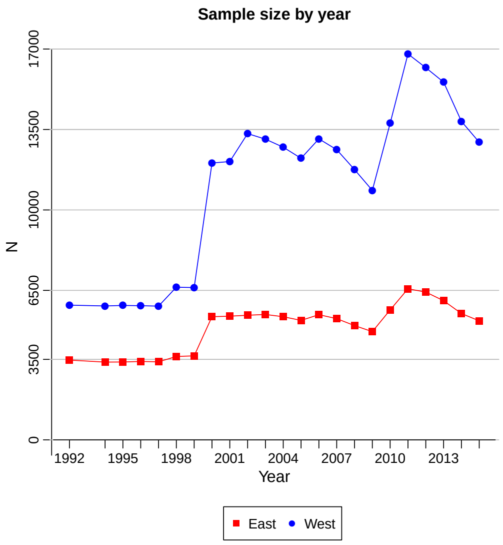
My data stem from the German Socioeconomic Panel (SOEP). The SOEP is a longitudinal survey with the aim of providing a representative picture of all private households in Germany (Gerstorf and Schupp 2016; Wagner et al. 2007). Individuals and private households are followed annually. In order to obtain representative results I use cross sectional weights as provided by the SOEP throughout my analysis (Kroh 2009). The SOEP collects information on an individual's current living situation as well as past experiences along a wide range of dimensions. These include education, income and labour market status, personality traits and health.

Importantly, the SOEP also inquires residency shortly before the fall of the wall in 1989 from each respondent. Rather than current residency or ever having lived in the GDR, I use this information to define the treatment status. The variable `East` takes the value one if the respondent lived in the GDR - including East Berlin - in 1989 and is zero if she lived in West Germany. Subsequently I will refer to individuals living in the GDR in 1989 as East Germans and to those living in West Germany as West Germans. All respondents for whom this information is missing or who did not live in Germany in 1989 are excluded from the sample. This treatment definition also implies exclusion of all individuals born after 1989. Furthermore I restrict the sample to respondents born not earlier than 1925. Since the FRG and the GDR were differentially affected by migration from third countries, I drop all respondents born outside Germany.

In total the sample includes 367,609 observations with non-missing outcome variables from 43,149 individuals. 12,102 of those individuals lived in East Germany in 1989 and 31,047 lived in West Germany. Figure 2.1 shows how the sample size varies over the observation period. During the 1990s the sample includes roughly 3500 East German individuals and 5800 individuals from West Germany. This number increases significantly due to a 2000 refreshment sample and declines subsequently.

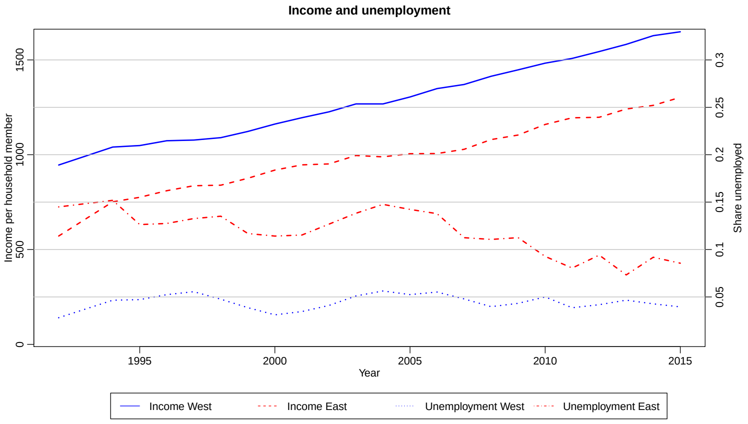
The basic characteristics of the sample are displayed in Table 2.1. West Germans are on average older than East Germans, while the proportion of sexes is roughly equal. East Germans have on average higher educated mothers, are less likely to have obtained

Figure 2.1: *Sample size*



Notes: Source: SOEP 1992-2015; cross sectional weights; only individuals born in Germany 1925-1989 with non missing outcome information. Number of individuals from East and West Germany by survey year.

Figure 2.2: *Income and unemployment*



Notes: Net income per household member and unemployment share for East Germans and West Germans in the weighted sample.

Table 2.1: *Descriptive statistics*

Individual level

	East	West	Diff	p-value	N East	N West
Sex	0.5022	0.5053	-0.0031	0.6581	12102	31047
Year of birth	1958.6814	1957.3693	1.3122	0.0000	12102	31047
Upper secondary education	0.2079	0.2856	-0.0778	0.0000	12031	30835
Intermediate secondary education	0.4835	0.2571	0.2264	0.0000	12031	30835
College degree	0.2199	0.1885	0.0315	0.0000	11997	30804
Apprenticeship	0.6899	0.6586	0.0312	0.0000	11997	30804
Number of siblings	1.8304	1.9741	-0.1437	0.0000	10045	26355
Mother: Year of birth	1932.5072	1929.4902	3.017	0.0000	11342	29287
Father: Year of birth	1929.5597	1926.0833	3.4765	0.0000	10967	28808
Mother: Upper secondary schooling	0.0767	0.0589	0.0179	0.0000	10944	28425
Mother: Immediate secondary education	0.2722	0.1591	0.1131	0.0000	10944	28425
Father: Upper secondary schooling	0.1112	0.126	-0.0147	0.0016	10468	27941
Father: Immediate secondary education	0.2381	0.1273	0.1108	0.0000	10468	27941

Observation level

	East	West	Diff	p-value	N East	N West
Monthly net income per household member	1001.7635	1300.4946	-298.7311	0	112223	255386
Unemployed	0.0968	0.0358	0.061	0.0000	112223	255385
In Education	0.0206	0.0235	-0.0029	0.0000	112223	255386
Willingness to take risk	4.6034	4.4101	0.1933	0.0000	103171	236657
Widowed	0.0667	0.0699	-0.0033	0.0129	111601	253581
Divorced or separated	0.121	0.1058	0.0153	0.0000	111601	253581
Married	0.5452	0.5703	-0.0251	0.0000	111601	253581
Number of individuals in household	2.4044	2.48	-0.0756	0.0000	112223	255386
Number of children	0.533	0.5819	-0.0489	0.0000	105286	243461
Reached statuatory pension age	0.1821	0.1984	-0.0163	0.0000	103598	241131
Healthy diet not important	0.5229	0.4914	0.0315	0.0000	29996	71273
Body-Mass-Index	26.244	25.9874	0.2565	0.0000	35815	85956
Low locus of control	0.0868	0.0336	0.0533	0.0006	14470	34503

Notes: Source: SOEP 1992-2015; Crosssectional weights; Only individuals born in Germany 1925-1989 with non missing outcome information.

upper secondary education but more likely to own a college degree.[3] Particularly large differences exist in terms of net household income per household member and unemployment. While the gap in income remains constant across the observation period, the gap in unemployment narrows decisively after 2005 (see Figure 2.2).[4] Furthermore

[3] In the GDR it was not uncommon for students to enter special engineering schools after obtaining an intermediate secondary schooling degree. After reunification graduates could apply for recognition of the engineering school degree as college degree. For example in the federal state of Saxony application was granted in the majority of cases (see https://www.medienservice.sachsen.de/medien/news/198827). Therefore the proportion of individuals with an college degree is higher than the proportion with upper secondary schooling in the East German sample.

[4] Figure 2.2 depicts the share of respondents in the sample who are unemployed. These numbers do not correspond to the official unemployment rates for East and West Germany for several reasons: Firstly I include all individuals below the age of 65 in the denominator rather than excluding indi-

East Germans have fewer children, are less likely to be married, have a higher BMI, rate a healthy diet less important and have a lower locus of control.

2.3.2 Outcome variables

In this study I consider four dimensions of health. A subjective measure of health is given by self-assessed health (SAH). In the GSOEP respondents are asked to rate their health on a 5-point Likert scale. I code this outcome in such a way that higher values of SAH correspond to a better health status. This variable is available in the SOEP from 1992 onwards.[5] Previous literature has raised concerns that measurement error in SAH might be endogenous (T. Crossley and Kennedy 2002). Nevertheless SAH has been shown to be a good predictor of future morbidity and mortality, and it is one of the most widely used measures of health (Jylhä 2009).

Secondly, I use the SF12-questionnaire which has been included in the SOEP every second wave since 2002. The SF12-questionnaire consists of 12 items on health related life quality. These cover limitations in daily life due to problems in mental or physical health, the presence of physical pain and the emotional state in the past four weeks (see for example SOEP 2013). The results are summarized in two subscales - the mental health summary score (MCS) and the physical health summary score (PCS). Each score has a mean of 50 and a standard deviation of 10 in the population. While the physical summary score decreases as individuals age, mental health does not detoriate in the same way (Andersen et al. 2007). Despite relying on a relatively short number of items, the summary scores from the SF12-questionnaire are highly correlated with those from the more extensive SF36 questionnaire (Gandek et al. 1998).

Finally I analyze reported medical diagnoses available for a range of diseases. Since 2009 the SOEP has been asking respondents biannually whether they have ever been diagnosed one of the following diseases: "Diabetes", "Asthma", "Heart Disease", "Cancer", "Stroke", "Migrane", "High blood pressure", "Depressive disorder", "Dementia". My measure "Number of diagnoses" is the total number of diseases on this list which an individual has been diagnosed with.

viduals who stay out of the labor force voluntarily. Secondly the sample does not include migrants and thirdly I group East and West Germans by residency in 1989 rather than current residency.

[5] With the exception of 1993 this variable is available from each wave since 1992.

Simple t-tests reveal significant differences in health between the two populations (see Table 2.2). When pooling all survey years and not adjusting for additional factors, individuals from East Germany have lower mental health and physical health summary scores and rate their health worse than individuals from West Germany. While the average number of diagnoses is slightly higher in the East German population, the difference is not significant.

Table 2.2: *Outcome variables*

	East	West	Diff	p-value	N East	N West
Mental health summary score (SF12)	49.2868	50.0402	-0.7534	0	35206	84644
Physical health summary score (SF12)	48.398	49.0308	-0.6327	0	35206	84645
SAH	3.3417	3.3671	-0.0254	0	112205	255312
Number of diagnoses	0.7647	0.7507	0.014	0.254	19608	47231

Notes: Means and t-tests by East and West on weighted sample. Self-assessed health is measured in all survey years 1992-2015 with the exception of 1993. Physical and mental health are measured biannully starting in 2002 and the number of diseases is measured biannually starting in 2009.

2.4 Health disparities

I begin the analysis by examining differences between the health status of former inhabitants of the GDR and their West German counterparts. Following a recent strand of literature, I exploit the German separation as natural experiment. Specifically, I assume the health status of individuals having lived in the GDR in 1989 would not systematically differ from those of individuals having lived in West Germany in the absence of a separation. Therefore, any gap in health we observe after reunification should be attributed to the separation itself. This effect might operate either directly, with experiences East Germans made living under a socialist regime explicitly entering the health production function, or indirectly by affecting third factors which serve as present day input into the health production function.

Even under the hypothetical scenario that I was able to observe both counterfactual outcomes - having lived in East Germany and having lived in West Germany in 1989 - for all observations in my data, it is not straightforward to define one parameter capturing the effect of having lived in the GDR on health entirely for several reasons.

Firstly individuals experienced living in East Germany at different stages in their life cycle. The oldest individuals in my sample were born before the German separation.

East and West Germans from this cohort share the experience of early live exposure to WWII and the post war period, while spending their adult years - including most of their working biography - under different regimes. The subsequent cohorts were born and reached adulthood during separation. Those individuals differ in their early live and education experiences as well as having entered two distinctive labour markets. On the other hand the youngest individuals ever having lived in the GDR made their occupational and marital decisions only after reunification. Nevertheless - unlike their West German counterparts - they were exposed to a socialist education system at a young age and raised in an environment influenced by GDR culture even after the reunification. Since one would not expect solely going to primary school in East Germany to affect health in the same way as working in the East labour market for many years, imposing a treatment effect which is homogeneous across cohorts seems very restrictive.

Secondly, differences in health status between East and West German individuals are likely to depend on the age at which health is measured. An earlier change in the latent stock of health might only become apparent older ages when health shocks occur more frequently.

Thirdly my data allow me to observe individual health outcomes for a period spanning up to 24 years. As health in East and West might converge over time, the treatment effect I measure will be sensitive to the timing of the measurement. Therefore I refrain from defining the one parameter of interest. Instead I will document health differences between East and West Germany across time, age and cohorts.

2.4.1 Empirical strategy

For the subsequent analysis, I divide my sample into three cohort groups. The oldest cohort group comprises individuals born before the German separation in 1949.[6] The middle cohort group includes individuals born between May 1945 and May 1973, while individuals born after May 1973 but before 1990 belong to the youngest cohort group. When choosing this second cut-off I follow the rationale of Fuchs-Schündeln and Masella (2016). After finishing the tenth grade students in the GDR would either continue their schooling (mostly with the goal of obtaining an university entrance

[6] I define the exact cut-off as May 1949, the month the FRG was founded.

qualification) or start an apprenticeship. However the decision of who was allowed to continue schooling was not based on academic achievement alone but also followed political criteria. Due to a cut-off rule in school entry, individuals born after May 1973 only finished the tenth grade after the fall of the wall. Therefore they did not face any political restrictions when choosing their further education, while individuals of earlier cohorts were possibly denied obtaining university entrance qualifications despite good grades. In Appendix B.1 I confirm that a data driven partitioning of the sample would lead to similar cut-offs.

In order to obtain point estimates and confidence intervals for the difference in health between East and West by cohort group and year, I estimate the regression model

$$Y_i = \beta_{0ct} + \beta_{1ct} * \text{East}_i + C_i + \epsilon_i \tag{2.2}$$

by OLS separately for each cohort group and survey year in which the outcome was measured. C_i denotes a vector of controls. The effect of interest β_{1ct} gives the difference in health between East and West Germans of cohort c in year t. Furthermore I am interested in testing whether there is any significance difference in the East-West gradient between the three cohort groups when considering all years together. Therefore I estimate the following regression model:

$$\begin{aligned} Y_{it} = &\gamma_1 * \text{Old} + \gamma_2 * \text{Middle} + \gamma_3 * \text{Young} + \gamma_4 * \text{East} * \text{Old} + \\ &\gamma_5 * \text{East} * \text{Middle} + \gamma_6 * \text{East} * \text{Young} + \delta_t + C_i + \epsilon_{it} \end{aligned} \tag{2.3}$$

Here Old is dummy variable indicating that an individuals was born 1949 or earlier, Middle is a dummy variable indicating that an individuals was born between 1949 and 1973 and the Young is an indicator for individuals born 1973-1989. δ_t denotes time fixed effects and C_i is the vector of controls. Since this specification does not include an East dummy, the coefficients of the interactions terms of the cohort group dummies and the East give the mean difference between West Germans and East Germans over all survey years. The main results using this specification come from pooled OLS regressions while results from poisson and ordered probit models - since the number of diagnoses are count data and SAH is measured on an ordinal scale - can be found in the Appendix. I cluster all standard errors at the household of origin level.[7] Controls include dummies for age in 5-year categories, a linear age trend and sex of respondent in each estimation. One might be worried that the randomization of the German population into East and West Germans was not successful in terms of balancing the family background of treatment and control group. Therefore I use maternal and paternal education (in 5 categories), year of birth of parents and the number of siblings as additional controls. However conditioning on these additional factors comes at the price of potentially introducing a bad control problem (Angrist and Pischke 2009). A control variable constitutes a bad control variable if it is an outcome of the treatment itself. While parental characteristics of individuals born prior to the separation can precede the treatment,[8] parents of younger individuals received their education and took their fertility decisions under the system of the GDR or the FRG. Therefore, when conditioning on family characteristics, results especially for the younger cohort groups should be treated with caution.

2.4.2 Results

Health differences over time

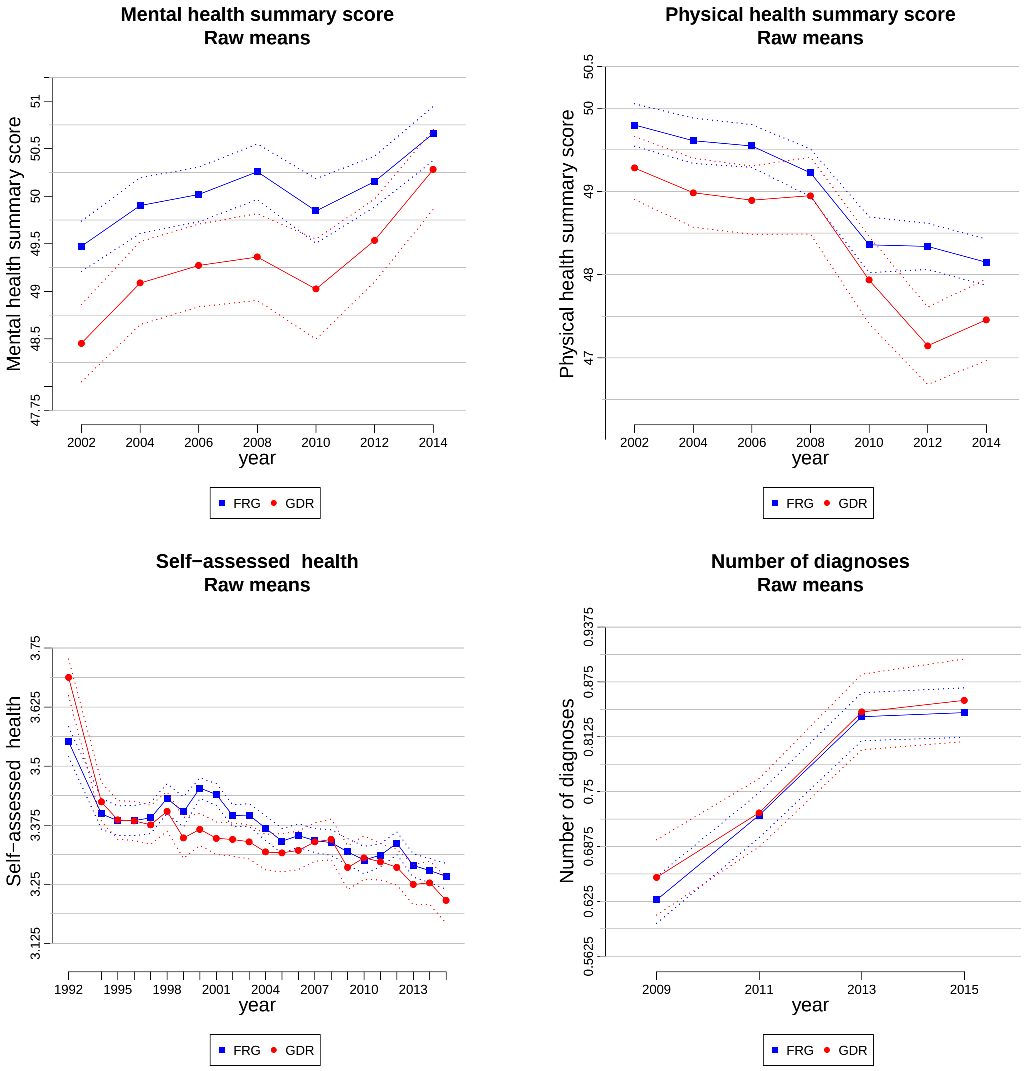
Firstly, I investigate how differences in health between East and West Germans evolve over time. Figure 2.3 shows the mean of each health outcome separately by survey year. For both East and West Germans I observe an improvement in mental health related life quality over time, but a deterioration in health measured by other outcomes. This is not surprising given the aging of the sample.

In each survey year, East Germans have a lower average mental health summary score than West Germans. The gap appears to become smaller in the final years of the observation period. Similarly, East Germans show a lower physical health related life quality throughout the observation period. There is no indication of convergence. Self-assessed health is the only health outcome observed during the years shortly after

[7] Household of origin is an identifier that includes the household id of the household the individual was first observed in. If for example an adult child moves out, he or she will be assigned a new household identifier, but keep the household of origin identifier.

[8] However, the number of siblings might be a bad control even for this group, when siblings are born (or would have been born) after the separation. Furthermore parents might have obtained an additional school degree within an adult-education program (*Zweiter Bildungsweg*).

Figure 2.3: *Raw means over time*



Notes: Health outcomes: Raw means for East Germans and West Germans over time and 95% confidence intervals. The plotted values are based on the estimated coefficients from the following regression model: $Y_{it} = \gamma_t + \delta_t * \texttt{East}_i + \epsilon_{it}$, with γ_t being time-fixed effects and δ_t additional time-fixed effects for East Germans. Standard errors are clustered at the household of origin level.

reunification. East Germans rate their health significantly better than West Germans in 1992, but they drop below West German levels quickly and remain so throughout the 1990s. In the final years of the observation period the gap becomes less clear-cut, but East Germans do not appear to fully catch up. With respect to the number of diagnoses, differences between East and West Germans are only marginal.

Regression results from estimation of Equation 2.2 (baseline controls: left panel of Figure 2.4, additionally controlling for parental background variables: right panel of Figure 2.4), suggest similar dynamics. For most years and outcomes East Germans exhibit an on average worse health status, and apart from mental health there is little evidence for a convergence over time.

Health differences by cohort

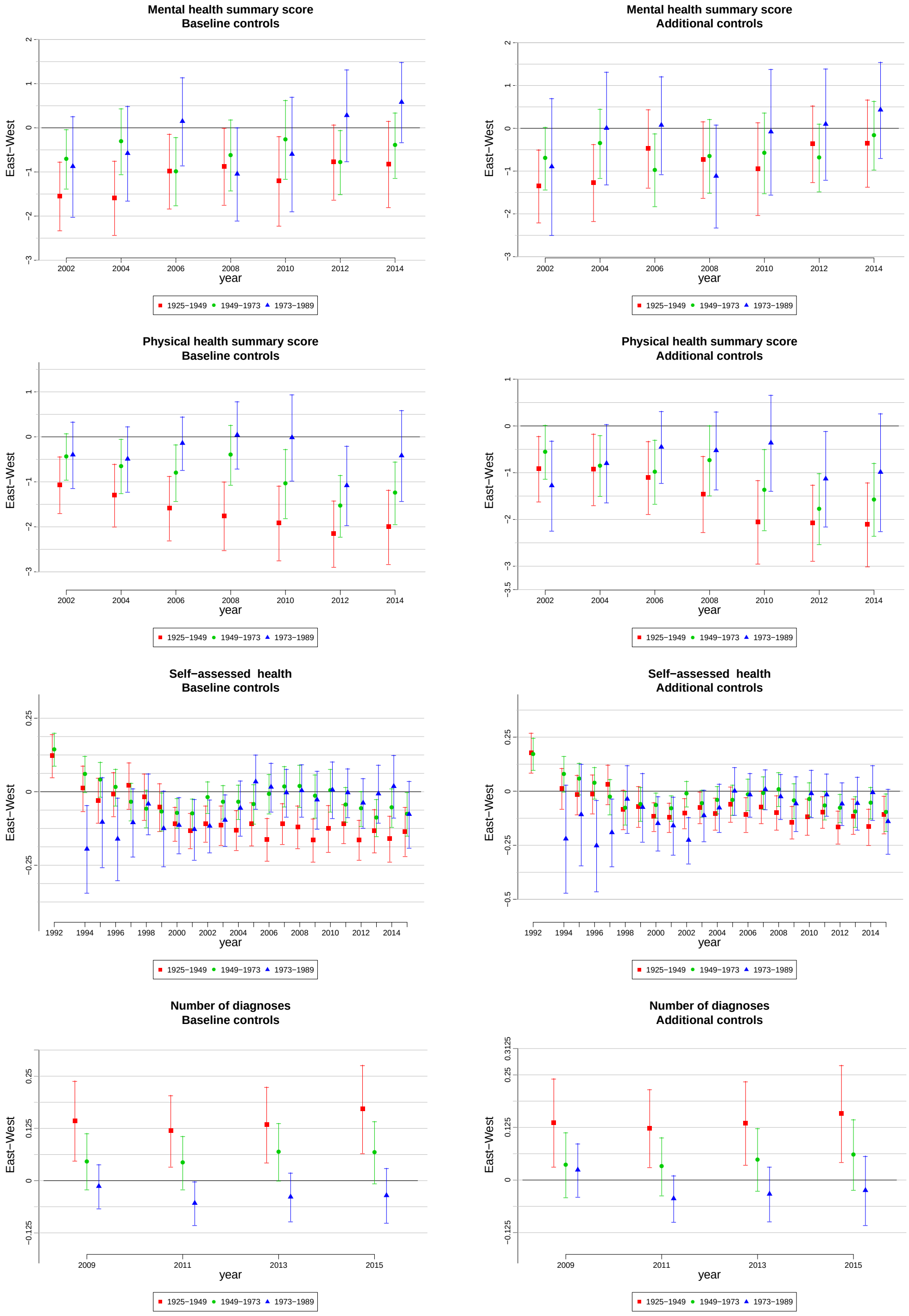
Next I turn to differences across cohort groups. Revisiting Figure 2.4, it is easy to see that health inequalities between East and West Germans are strongest for individuals born 1949 or earlier. When only controlling for baseline characteristics, the old cohort group exhibits the greatest disadvantage of East Germans for almost all outcomes and survey years.[9] In the case of the physical health summary score and the number of diagnoses, one even observes a clear ordering with respect to birth cohort. In the youngest cohort group, the difference between individuals from East and West Germany is hardly ever significant at the 95% significance level, while there is a significant and even growing gap for the oldest cohort group.

Additionally controlling for parental and family characteristics, does not alter results profoundly (see right Panel of Figure 2.4). East Germans from the old cohort group appear to be less disadvantaged in terms of mental health compared to the baseline specification. However, these results should be treated with caution as they are likely to suffer from bad control bias.

Subsequently, I pool all survey years and estimate the overall East-West gap in health for each cohort. Results are displayed in Table 2.3. Columns (1) and (2) show the estimated coefficients of the interaction term between the East-dummy and the cohort indicators of Equation (2.3). When pooling all years, I detect a significant health disad-

[9] This does not hold up for SAH in the 1990s. However, when focusing on the time since 2002 when we also observe the outcomes from the SF12-questionnaire, the pattern is consistent across outcomes.

Figure 2.4: *East-West gap by cohort groups and survey years*



Notes: Health outcomes: East-West gap by cohort groups and survey years. Plotted are the estimated coeffiecients $\hat{\beta}_{ct}$ obtained from estimating the regression model $Y_{it} = \beta_{0ct} + \beta_{1ct} * \texttt{East}_i + C_i + \epsilon_{it}$ separately for the old, middle and young cohorts and each survey year. The error bars denote the pointwise 95% confidence intervals. Baseline controls include dummies for age in 5-year categories, a linear trend in age and sex. Additional covariates include the year of birth for both parents, education of parents (5-categories) and the number of siblings. Regression are weighted using cross sectional weights provided by the SOEP. Standard errors are clustered on the household of origin level.

Table 2.3: *East-West differences by cohort group*

	All		Female		Male	
Mental health summary score						
	(1)	(2)	(3)	(4)	(5)	(6)
East*Born 1925-1945	-1.1568***	-0.8967***	-0.9251**	-0.5569	-1.4311***	-1.2875***
	(0.2914)	(0.3081)	(0.3646)	(0.4004)	(0.3614)	(0.3804)
East*Born 1945-1973	-0.5783**	-0.5285*	-1.2397***	-1.22***	0.0457	0.1384
	(0.2568)	(0.2718)	(0.3372)	(0.3494)	(0.3296)	(0.3586)
East*Born 1973-1989	-0.2862	-0.3515	-0.7323*	-0.5282	0.1486	-0.1917
	(0.3073)	(0.3469)	(0.4145)	(0.4644)	(0.4108)	(0.4746)
Observations	119850	102239	62647	53556	57203	48683
Individuals	34253	28333	17995	14943	16258	13390
p-value	0.1058	0.4763	0.592	0.3142	0.003	0.0215
Physical health summary score						
	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
East*Born 1925-1945	-1.6575***	-1.5582***	-1.3681***	-1.2448***	-2.0031***	-1.9178***
	(0.2573)	(0.2803)	(0.3404)	(0.3762)	(0.3554)	(0.3839)
East*Born 1945-1973	-0.8557***	-1.1033***	-0.7782***	-1.0763***	-0.9425***	-1.1444***
	(0.2226)	(0.2418)	(0.2973)	(0.3201)	(0.2918)	(0.3311)
East*Born 1973-1989	-0.3548	-1.062***	-0.5023*	-0.975***	-0.2299	-1.1772***
	(0.2348)	(0.2637)	(0.2888)	(0.3183)	(0.3433)	(0.3921)
Observations	119851	102240	62647	53556	57204	48684
Individuals	34253	28333	17995	14943	16258	13390
p-value	9e-04	0.3696	0.1518	0.8623	0.0015	0.2611
Number of diagnoses						
	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
East*Born 1925-1945	0.1403***	0.1373***	0.141***	0.1227**	0.1382**	0.1551***
	(0.0386)	(0.0421)	(0.0498)	(0.0544)	(0.0542)	(0.0579)
East*Born 1945-1973	0.0562*	0.0455	0.0555	0.0499	0.058	0.0427
	(0.029)	(0.0317)	(0.0377)	(0.0418)	(0.0423)	(0.0454)
East*Born 1973-1989	-0.0383*	-0.0268	-0.0341	-0.0045	-0.039	-0.0431
	(0.0202)	(0.0238)	(0.0309)	(0.0372)	(0.0258)	(0.0292)
Observations	66839	57344	35515	30491	31324	26853
Individuals	26684	22648	14258	12113	12426	10535
p-value	1e-04	0.0017	0.0077	0.136	0.0048	0.0051
Self-assessed health						
	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
East*Born 1925-1945	-0.0831***	-0.0846***	-0.0701**	-0.072**	-0.0977***	-0.0972***
	(0.0209)	(0.024)	(0.0275)	(0.0319)	(0.0279)	(0.0321)
East*Born 1945-1973	-0.0194	-0.0279	-0.0308	-0.0434*	-0.0093	-0.0135
	(0.0173)	(0.0197)	(0.0225)	(0.0258)	(0.0228)	(0.0264)
East*Born 1973-1989	-0.0388	-0.092***	-0.0375	-0.0693**	-0.04	-0.1159***
	(0.0248)	(0.0286)	(0.0269)	(0.0322)	(0.037)	(0.0427)
Observations	367517	296376	193448	157001	174069	139375
Individuals	43147	31321	22530	16581	20617	14740
p-value	0.0553	0.063	0.5242	0.7061	0.0468	0.0411
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Additional controls	No	Yes	No	Yes	No	Yes

Notes: Results from estimation of Model 2.3 using pooled OLS. Standard errors clustered on the household of origin level. Significance levels: : * p<0.10, ** p<0.05, *** p<0.01.

vantage of East Germans of the old cohort group for all outcomes. East Germans who were born prior to 1949 score on average 1.16 points (11.6% of a standard deviation) lower in the mental health summary score and 1.66 points (16.6% percent of a standard deviation) lower in the physical health summary score than West Germans from the same age group. Furthermore, they report on average 0.14 more diagnoses and rate their overall subjective health 0.08 points lower. In the middle cohort group the inequalities amount to a 0.58 points lower MCS for East Germans, a 0.86 lower PCS, 0.06 more diagnoses while the difference in self assessed health is negligible. To put these numbers into context, the gap in MCS is about 40% of the effect size Marcus (2013) find for becoming unemployed in the old cohort group and one third of the effect size of a spouse becoming unemployed for the middle cohort. For the youngest cohort the East-West gap is mostly negligible.[10]

In columns (4)-(6) I split the sample by sex. East-West differences tend to be stronger for males than for females in the old cohort. I further test the hypotheses that the East-West gap in health is equal across cohort groups. The corresponding p-values are displayed in the bottom row of each panel. In specifications where I only include baseline controls, the difference is mostly significant for the whole sample and the male sample, but not for the female sample. As controlling for parental and family characteristics reduces the magnitude of the estimated coefficients for the oldest group and increases them for the youngest cohort group, I reject the null-hypothesis of equal effects across cohorts less often when these additional controls are included. Differences across cohorts are stronger for males than for females.

In the main analysis I treat all outcome variables as continuous measures. In Appendix Table B.3 I show that results are very similar when estimating Equation 2.3 using nonlinear models for the number of diagnoses and a binary version of SAH. The differences in the predicted number of diagnoses between East Germans stemming from Poisson regression are both quantitatively and in terms of significance levels only marginally different from the estimated coefficients in the linear OLS model. I furthermore collapse SAH into a binary variable indicating *good* or *very good* health status and predict

[10] The young cohort group East Germans tend to report fewer diagnoses although the difference is not large. Among young individuals the average number of reported diagnoses is very low and East Germans in my sample suffer from asthma less often, replicating a finding from previous studies (see for example Steppuhn et al. 2017).

the difference in the probability of reporting an at least good health status between East and West Germans. I find East Germans in the oldest cohort to have an about 5% percentage points lower probability of reporting a good health status than West Germans.

Health differences by age

So far it has become apparent that health disparities between East and West Germans are strongest for individuals who I observe at older ages. This result could either be driven by cohort effects, with the experience of living in the GDR having been more detrimental to health of individuals born prior to 1949, or by an age effect with health disadvantages of East Germans manifesting themselves more strongly at high ages. For example, in the case of PCS and the number of diagnoses, a disadvantage of East Germans in the middle cohort group emerges over the course of the observation period (see Figure 2.4). This finding hints to an age effect rather than a cohort effect. While it is impossible to answer this question definitely, I will take closer look at the patterns of ageing in this section.

Figure 2.5 plots health outcomes over age. Here I adjust for sex of respondents and year effects. With the exception of mental health, we see a deterioration in health as individuals become older. This decline is stronger in the East German sample. While levels of health remain barely indistinguishable for young individuals, a gap emerges after the age of forty for SAH and the number of diagnoses - and even earlier for PCS and MCS.

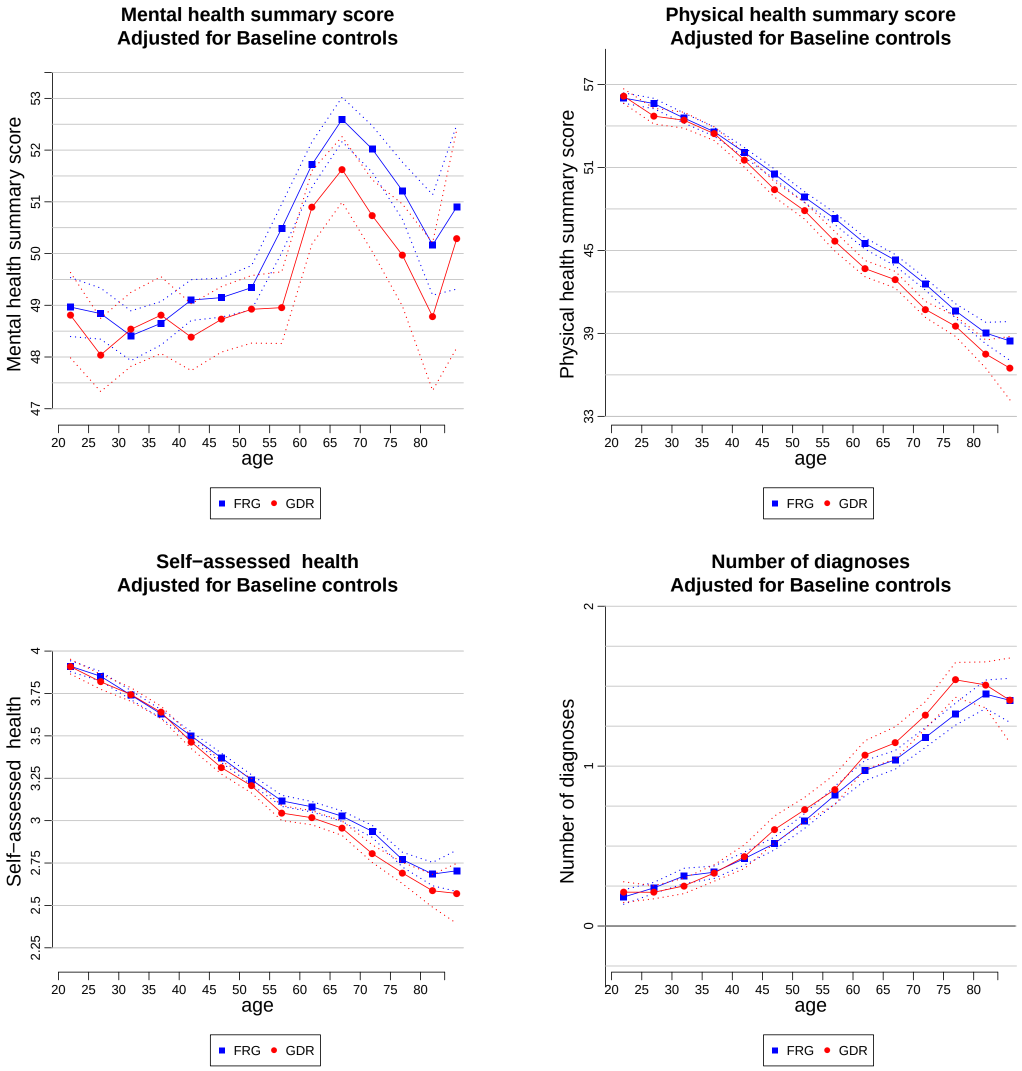
In order to obtain a more detailed picture of within cohort dynamics, I split the sample into groups of 5-year-adjacent birth cohorts. Figure 2.6 plots health differences between East and West Germans against age separately for each group of 5-year birth cohorts.[11]

Whereas this graph confirms the impression that East-West differences are in general more pronounced for individuals whom I observe at older ages, the within cohort dynamics of East-West health differences depend on the outcome. For mental health, physical health and the number of diagnoses there is a downwards sloping trend across cohorts.[12] The earlier a cohort was born the more disadvantaged are East Germans

[11] To increase readability I plot smoothed trends rather than the raw values.

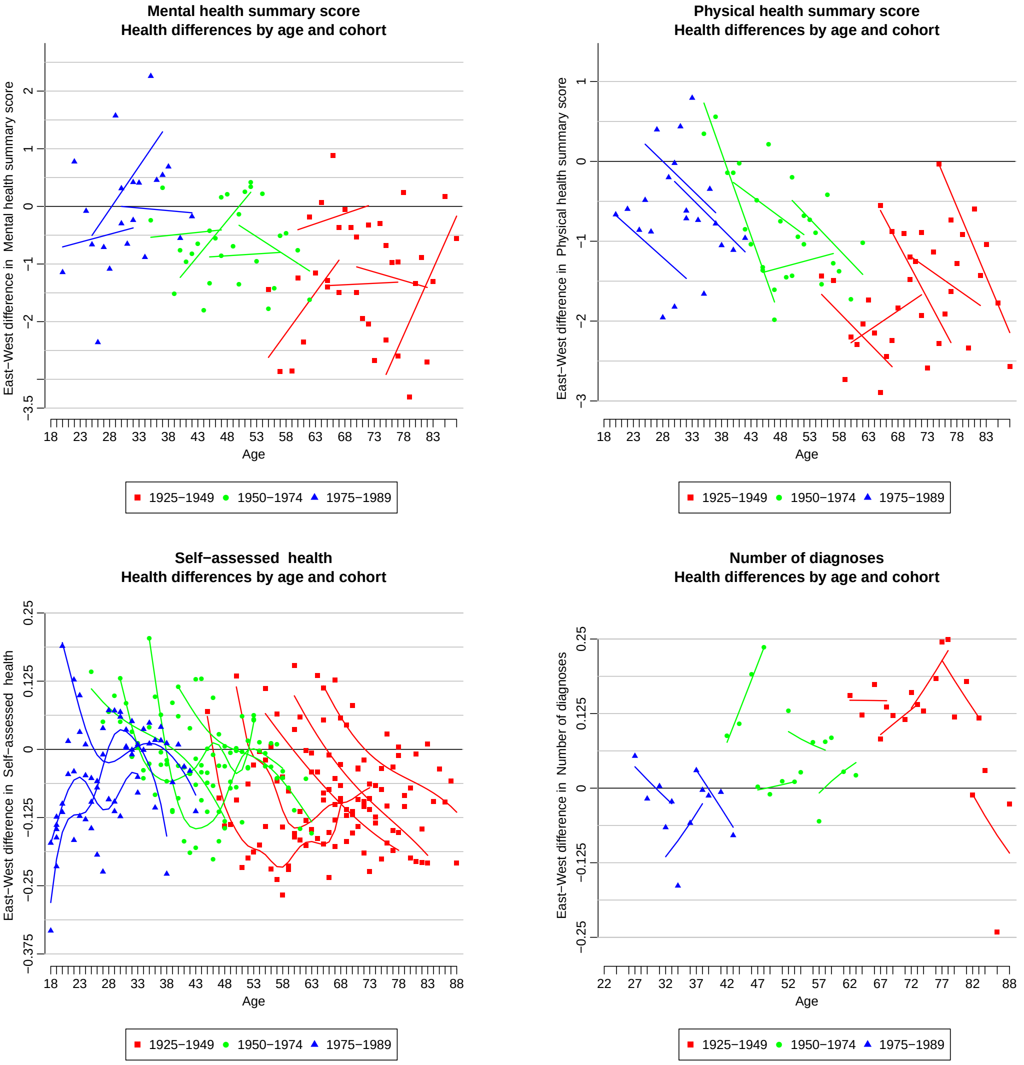
[12] This does not hold for the oldest 5-year group in terms of the number of diagnoses. However, this sample is very small and estimates are imprecise.

Figure 2.5: *Healthy by age*



Notes: Health outcomes: Adjusted means for East Germans and West Germans over age and 95% confidence intervals. The plotted values are based on the estimated coefficients from the following regression model: $Y_{ia} = \gamma_a + \delta_a * \texttt{East}_i + C_{ia} + \epsilon_{ia}$, with γ_a being age-fixed effects in 5-year bins and δ_a additional aged-fixed effects for East Germans. Controls include year-fixed effects and sex. Standard errors are clustered at the household of origin level.

***Figure 2.6:** Trends in 5-year birth cohorts*



Notes: Health outcomes: East-West differences by age and birth-cohort in 5-year bins. Each line represents the smoothed trend for a 5 adjacent birth cohort. The values are calculated running the following linear regression model separately for each 5-year-group of birth cohorts: $Y_{it} = \gamma_t + \delta_t * \texttt{East}_i + C + \epsilon_{it}$ with controls including sex and linear year of birth. I plot the mean age of each 5-year group against the health difference between East and West Germans. The smoothed trends are obtained using local linear regression (bandwidth obtained by cross validation using the AIC criterion). Regression are weighted using cross sectional weights provided by the SOEP.

compared to West Germans. For PCS the trend is also downwards sloping within most cohorts, the same holds for the number of diagnoses between ages of 40 and 75. On the other hand no such consistent pattern can be observed for MCS and the number of diagnoses. In the case of SAH no between-cohort gradient is present and most cohorts exhibit a downwards sloping trend. The fact that East Germans rated their health better than their West German counterparts in 1992 is not driven by any particular cohorts.

In the Appendix in Figure B.2 I look at the aging process within persons. I normalize health at age 20 for each person to zero and investigate how health changes in the following ten years. I repeat this procedure up to a base age of 70. For PCS the trends match those in Figure 2.5. Divergence between East and West takes mostly place in the middle thirties and late forties, suggesting that the emerging gap between East and West might be driven by aging rather than selection for this outcome. On the other hand for SAH differences between East and West tend to be very small while patterns of divergence do not completely conform to those observed in Figure 2.5 in the case of the number of diagnoses and MCS.

To summarize, my analysis reveals a significant disadvantage in health for those East Germans whom I observe at older ages. East Germans from the oldest cohort group exhibit a disadvantage in health across all outcomes and almost all survey years. Among individuals born after 1949 inequalities are less pronounced. The finding that an East-West gap builds up gradually in the middle cohort for, points to the existence of age effects in case of age sensitive measures.

2.4.3 Robustness

Migration prior to 1990

The socialist regime of East Germany is now infamous for the construction of the Berlin Wall - part of a deadly border regime preventing its citizens from leaving the country. These measures were in fact a reaction to massive outmigration. Between 1949 and 1961, about 2.7 million of originally 18.3 million inhabitants relocated to the West. Only after the Berlin wall was constructed and the coincident closure of all borders these numbers dropped significantly (see Figure 2.7). If relocation decisions are related to the health status of an individuals, East-West migration will invalidate the natural

Figure 2.7: *Number of relocations from GDR to FRG (in 1000) according to Mayer 2002*

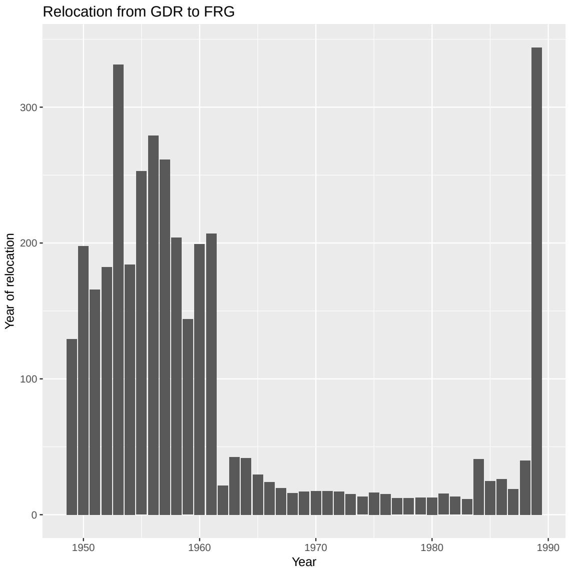
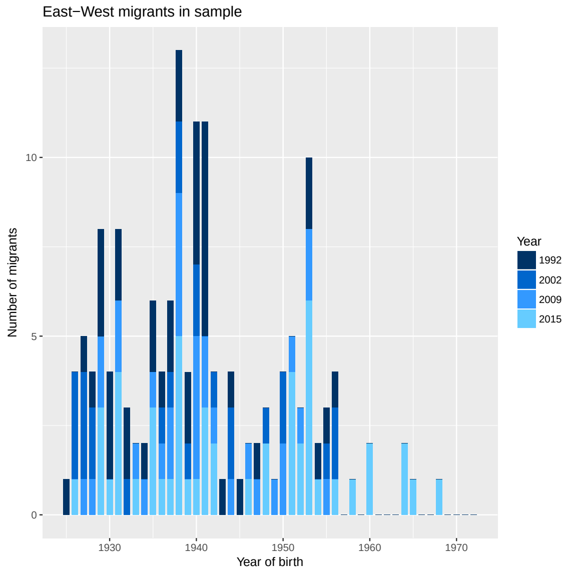


Figure 2.8: *East-West migrants in the sample by year of birth and survey year.*



experiment which I am exploiting. For most individuals in my data I observe only the location of residency in 1989. Therefore my definition of East Germans excludes individuals who previously left the East. Consequently the East-West gap I observe might in principle stem from positive selection. In this robustness check I examine the sensitivity of my results with respect to the classification of East-West migrants. For this purpose I draw on a subsample of my data, who answered a 1990-1993 module on migration experiences including migration within Germany.[13] Among the 10,320 eligible individuals having answered the first question of the migration module, I identify 154[14] respondents who migrated from the GDR to the FRG between 1949 and 1989. Figure 2.8 shows the distribution of year of birth for these migrants. All East-West migrants were born before 1972 and the vast majority belongs to the oldest cohort group. In order to archive a balanced sample I exclude the youngest cohort group from this analysis and pool the other two cohort groups. Figure 2.8 shows the number of East-West migrants for the survey years 1992, 2002, 2009 and 2015. The colors of the bars correspond to the number of migrants observed in a specific year. In 2009 only about 55% of

[13] This module requests respondents to state whether they have been living in that part of Germany where they currently reside since birth or since 1949. Individuals who have relocated are subsequently asked about the year of relocation and where they lived before.

[14] This count only includes respondents with non-missing data on other relevant outcomes.

Table 2.4: *Robustness check: East-West migrants prior to 1990*

	Physical health summary score				Mental health summary score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
East in 1989	-1.03*** (0.29)			-1.06*** (0.29)	-0.49 (0.35)			-0.49 (0.35)
East in 1989 or relocated		-1.04*** (0.29)				-0.47 (0.34)		
Share of years spent in East Germany			-1.06*** (0.29)				-0.51 (0.35)	
Relocated				-0.82 (1.17)				-0.12 (1.17)
Observations	28901	28901	28892	28901	28901	28901	28892	28901
	Number of diagnoses				**Self assessed health**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
East in 1989	0.069* (0.041)			0.074* (0.042)	-0.015 (0.020)			-0.018 (0.021)
East in 1989 or relocated		0.079* (0.042)				-0.022 (0.020)		
Share of years spent in East Germany			0.070* (0.041)				-0.017 (0.020)	
Relocated				0.14 (0.16)				-0.066 (0.068)
Observations	13177	13177	13173	13177	120772	120772	120738	120772

Notes: *East in 1989* denotes the standard dummy for East Germans. *East in 1989 or relocation* additionally includes individuals who relocated from East to West Germany prior to 1989. *Share of year spent in East Germany* denotes the share of year 1949 (or since birth) to 1989 an individual spent in East Germany. I include only individuals who answered a 1990-1993 module on migration experiences. Standard errors are clustered at the household of origin level. Significance levels: * p<0.1, ** p<0.05, ***p<0.01

the original migrants were still in the sample, by 2015 this number has further reduced to about a third.[15]

Columns (1) and (5) of Table 2.4 present the East-West gap within the subsample for whom information on East-West migration prior to 1990 exists , when I define East Germans in the same way as in the main analysis. The magnitude of the East-West gap in this highly unrepresentative sample of long-time survey participants is not in the focus of this exercise. Instead I am interested in assessing the sensitivity of the estimated coefficients with respect to the classification of migrants. I build on the assumption that a change in the classification would affect the results on the sample used for the main analysis in a similar way. In columns (2) and (6) I define all individuals who ever resided in East Germany between 1949 and 1989 as East Germans. The estimated coefficients are comparable to those in Columns (1) and (5), only for the number of diagnoses the effect size increases slightly. In columns (3) and (7) I account for the fact that individuals who relocated spent only some years in East Germany. Instead of a binary indicator I use share of years 1949 (or since birth) to 1989 an individual spent in

[15] The share of migrants does not drop as drastically as East and West Germans who answered the migration module in 1990-1993 also exit the survey.

each part of Germany. Again the estimated coefficients are of similar magnitude as the original effects. Finally migration itself might affect health, such that migrants should neither be classified as East or West Germans nor as weighted average. Therefore I include a separate dummy for migrants in columns (4) and (8). Once more the estimated difference do remain very close to the original effect.

Permutation test

Previous research has documented that geographic differences in health exist not only across countries but also within countries (see for example Chetty et al. 2016; Müller-Nordhorn et al. 2008; Walter 1992). To address the concern the East-West gap in health might reflect general regional differences rather than being a consequence of the separation, I conduct a permutation test. Constructing synthetic separations of Germany allows me to compare the size of the original East-West differences in health to differences resulting from general regional variation. I proceed by repeatedly assigning 6 German federal states to a synthetic East Germany and the remaining 10 federal states to a synthetic West Germany. I only allow for partitions where both synthetic states form a closed territory. The administrative geography of Germany admits 51 such partitions.[16] Furthermore, in order to mimic the separation of Berlin, I randomly assign half of the observations in one of the six states originally belonging to the synthetic East, to the synthetic West Germany. This procedure provides 306 synthetic East and West Germanys. I obtain synthetic treatment effects by estimating Equation 2.3 for each partition 10 times. Optimally I would assign each respondent to the state where she lived in 1989. Since this information is not available for a large fraction of observations, I assign each individual to the federal state where she is observed first. As individuals who moved from East to West Germany might be positively selected in terms of health, I repeat the analysis on a subsample excluding those who moved between East and West prior to entering the GSOEP. I compare the estimated synthetic treatment effects to the original effects shown in the first column of Table 2.3. If the absolute size of the original East-West gap is in the lower of or middle range of the distribution of (absolute) synthetic treatment effects, this would suggest that the East-West gap can be attributed to general regional variation.

[16] Excluding the true separation.

Table 2.5: *Permutation test: Probability of obtaining an effect size lower than the East-West difference*

	Full sample			
	Mental health	Physical health	Number of diagnoses	Self-assessed health
East*Born 1925-1949	0.91	1	0.96	1
East*Born 1949-1973	0.59	0.97	0.83	0.4
East*Born 1973-1989	0.47	0.67	0.99	0.55
	Non movers			
	Mental health	Physical health	Number of diagnoses	Self-assessed health
East*Born 1925-1949	0.99	1	0.97	1
East*Born 1949-1973	0.73	0.94	0.92	0.42
East*Born 1973-1989	0.68	0.64	1	0.49

Notes: Comparison of estimates displayed in the first column of Table 2.3 and synthetic East-West differences. I construct 306 synthetic East Germanys and use these synthetic treatments to reestimate Equation 2.3. This Table displays the probability of obtaining a smaller absolute effect size than the true effect size when using the full sample. E.g. East Germans from the oldest cohort report on average 0.14 diseases more than their West German counterparts. 4% of the of synthetic separations produce a larger effect. East Germans from the middle cohort report on average 0.06 diseases less than West Germans of the same group. 17% of synthetic separations produce treatment effect with a larger absolute value.

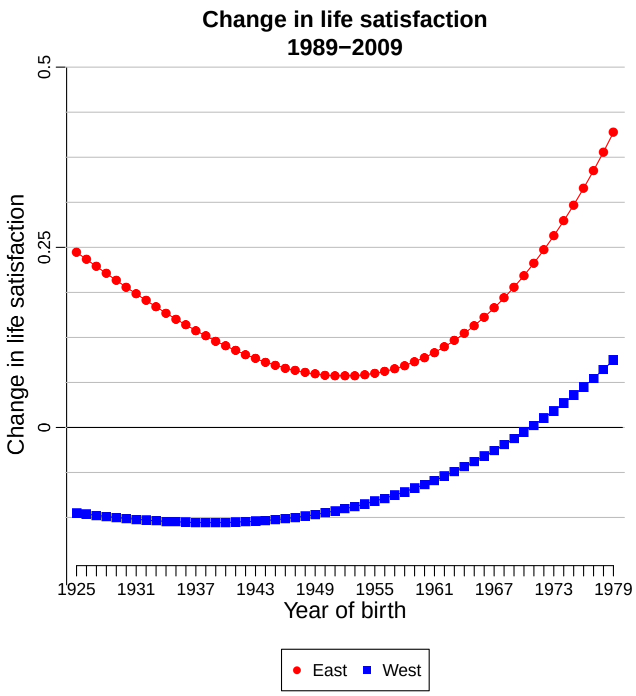
For the oldest cohort group the size original East-West gap is above the 90th percentile for all four outcomes. In the case of PCS and SAH no synthetic separation provides an estimated treatment effect that exceeds the original estimated treatment effect. This confirms the finding that the separation of Germany persists in significant health differences among old individuals (see Table 2.5). Results are less clear when considering those born after 1949. For the middle cohort group the size original East-West gap only resides in the top 10% of effect sizes in the case of the physical health summary score. For the young cohort group only the East-West difference in the number of diagnoses is greater than all other regional variation.[17] When focusing on the sample of non movers for which results are displayed in the lower panel, the probability of obtaining effect sizes in the synthetic partitions stays nearly constant or increases.

Reunification shock

The analysis above has revealed that post-reunification health inequalities between East and West Germans are strongest among individuals born prior to 1949. I argue that a health disadvantage of East Germans is a direct or indirect consequence of having lived under the Socialist regime in the GDR. However, as difficulties in adapting to a new environment increase with age, the reunification in 1990 might have provided a more

[17] East Germans of the youngest cohort group report fewer diagnoses than West Germans. Most notably they report a lower incidence of asthma.

***Figure 2.9:** Change in life satisfaction 1989-2009*



Notes: Self reported change in life satisfaction 1989-2009 as collected in the SOEP 2009. The graph shows $\alpha_{east} + f_{east}(yob)$ and $\alpha_{west} + f_{west}(yob)$ $\Delta_{ls} = \alpha_{east} + f_{east}(yob) + alpha_{west} + f_{west}(yob) + C + u$. I use B-splines of degree 3 without internal knots to approximate $f_{east}(age)$ and$f_{west}(age)$. Regression are weighted with cross sectional weights provided by the SOEP. Controls include sex of respondent.

distressful experience for individuals born prior to 1949 than for younger individuals. Therefore another possible interpretation of my results is that the sudden transition from one regime to the other - rather than exposure to the GDR itself - provided a long-lasting burden to individual health. As individuals living in East Germany in 1989 inevitably also experienced reunification, both effects are hard to disentangle. In this exercise I exploit self-reported changes in life satisfaction to investigate whether a health disadvantage of older east Germans might be driven by a reunification shock. In 2009 the SOEP asked its respondents to rate the change individual life satisfaction since 1989, the year prior to reunification.[18] I code this variable to be equal to one if life satisfaction has improved, equal to minus one if life satisfaction has reduced and equal to zero if it stayed constant. If the reunification disproportionally distressed old individuals, this should be reflected in old East Germans reporting a disproportional decline in life

[18] While the question explicitly states that 1989 is the year prior to reunification, it clearly enquirers individual life satisfaction and not satisfaction with political conditions.

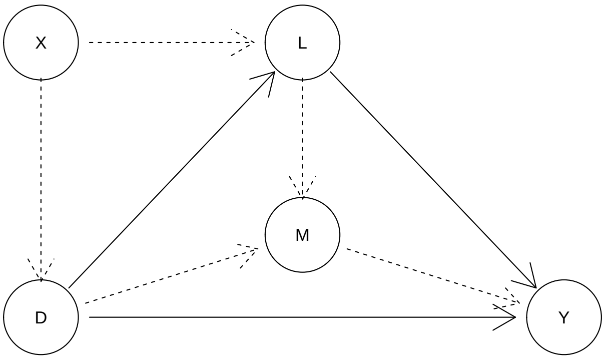
satisfaction. Figure 2.9 displays the average change in life satisfaction by birth cohort for East and West Germans. The graph does not support the idea that old East Germans suffered more than other cohort groups in the early nineties. For all birth cohorts East Germans are more satisfied with their lives now relative to West Germans. Among West Germans change in life satisfaction increases with year of birth and only individuals born in 1969 or later report on average that life satisfaction has improved. For East Germans on the other hand the relationship between average change in life satisfaction and year of birth is U-shaped. East Germans born in the early 1950s are the least satisfied with their lives now compared to 1989 and the difference between East and West Germans are smallest among birth cohorts born around 1960.

2.5 Controlled direct effect

So far, this analysis has revealed a significant disadvantage in health for East Germans born prior to 1949 and small disparities for younger cohorts. As East and West Germany continue to vary along many dimensions, the gap I documented above does not necessarily have to be attributed to experiences East Germans made while living under the GDR regime alone. Rather one would expect the post-reunification experiences of economic uncertainty and low wages as well as differences in health behavior and personality to affect individual health as well. I ask, whether there exists any direct effect of having lived in the GDR on health when holding the level of contemporary inputs into the health production function constant. Conditioning on post treatment confounders induces a bad control problem. Therefore this question cannot be addressed by simply including additional control variables in a regression of health on the East dummy. Instead I turn to a quantity that is known in the mediation analysis literature as controlled direct effect (Pearl 2001). The controlled direct effect (CDE) is the effect of changing a treatment D from level d' to d while keeping an intermediate variable M, which is itself an outcome of the treatment, at a fixed level m. As East and West Germans differ strongly in terms of income and unemployment rates even 25 years after the reunification, my focus will be on netting out these two factors. Furthermore I consider the role of locus of control and dietary behavior.

To define the CDE formally, I apply the potential outcome framework of Rubin (1974). $Y_i(d, m)$ denotes the potential outcome of individual i under treatment level d when the

Figure 2.10: *CDE in direct acyclic graph*



Notes: DAG showing the causal relationship between pretreatment confounders X, treatment D, intermediate confounders L, mediator M and outcome Y. The solid arrows denote the controlled direct effect, that is the effect of D on Y which does not operate via M (see also Acharya et al. (2016), Figure 3).

mediator M is fixed at value m. Similarly $M_i(d)$ denotes the potential value of the mediator for individual i under treatment d. Consistency requires the observed variables Y_i and M_i to correspond to the potential outcomes under the actual treatment status, such that $Y_i = Y_i(D_i, M_i)$ and $M_i = M_i(D_i)$. The CDE is defined as:

$$CDE_i(d, d', m) = Y_i(d, m) - Y_i(d', m) \tag{2.4}$$

In a directed acyclic graph (DAG) the CDE can be represented as the set of paths from T to Y which do not pass the mediator (Acharya et al. 2016). In Figure 2.10 the CDE is indicated by solid arrows. Besides D, M and Y the graph includes two additional nodes. X denotes the set of confounders which are not affected by the treatment, while L is a vector of confounders affected by the treatment but not by the mediator. Since one holds only M fixed, the CDE does contain the effect of D on Y which operates via L. Testing for the presence of a controlled direct effect corresponds to testing whether there exists any path from D to Y in the DAG after removing the edge leading from M to Y. The controlled direct is distinct from the natural direct effect (NDE).[19] While the

[19] Unlike the NDE, the CDE does not provide a straightforward composition of the ATE into a direct and an indirect effect.

mediator M is held constant at a particular value m in the CDE, one fixes the treatment level of the potential value of the mediator $M_i(a)$ when estimating the NDE.[20]

$$NDE_i(d') = Y_i(d, M(d')) - Y_i(d', M(d')) \tag{2.5}$$

Thus the natural direct effect gives the effect of the treatment when individuals are not allowed to change the value of the mediator because of the treatment. The CDE on the other hand corresponds to the treatment effect in an experiment where both the treatment and the mediator are manipulated by the experimenter.[21] Importantly for applied researchers the controlled direct is identified under weaker assumptions than the natural direct effect (VanderWeele and Vansteelandt 2009). Economists have employed the controlled direct effect to evaluate the effects of awarding vouchers for vocational training to the unemployed (Huber et al. 2017).

2.5.1 Empirical framework

In my empirical strategy I follow the framework of Acharya et al. (2016) to estimate the average controlled direct effect, $ACDE = E\left(Y_i(d,m) - Y_i(d',m)\right)$, using sequential g-estimation.[22] The ACDE is nonparametrically identified under the assumption of sequential unconfoundness and common support:

Assumption 2.1. *(Conditional independence of potential outcomes and treatment)*
$Y_i(d,m) \perp\!\!\!\perp D_i \mid X_i = x$ *for all* $t \in \mathcal{D}, m \in \mathcal{M}$ *and* $x \in \mathcal{X}$

Assumption 2.2. *(Conditional independence of potential outcomes and mediator)*
$Y_i(d,m) \perp\!\!\!\perp M_i \mid D_i = t, L_i = l, X_i = x$ *for all* $d \in \mathcal{D}, m \in \mathcal{M}, l \in \mathcal{L}$ *and* $x \in \mathcal{X}$

Assumption 2.3. *(Common support)*
Pr$(D_i = d \mid X_i = x)$ *and Pr*$(M_i = m \mid D_i = d, L_i = lX_i = x)$ *for all* $d \in \mathcal{D}, m \in \mathcal{M}, l \in \mathcal{L}$ *and* $x \in \mathcal{X}$

Assumptions 2.1 and 2.3 are the standard assumptions necessary for identification of the average treatment effect. Identification of the CDE additionally requires indepen-

[20] Or pure direct effect under the nomenclature of Robins and Greenland (1992).

[21] Pearl (2001) calls to the CDE "prescriptive" and the NDE "descriptive".

[22] The sequential g-estimator was first proposed by Goetgeluk et al. (2008) and - under the name RS2S estimator by Joffe and T. Greene (2009).

dence of potential outcomes and the mediator conditional on D, L and X (Assumption 2.2).

The effect identified under Assumptions 2.1-2.3 still depends on the conditional distribution of intermediate confounders L (Robins 1994). Therefore estimation of the CDE requires the researcher to integrate the CDE conditional on L over the distribution of L. I avoid this step by assuming the absence of any interaction between the mediator and L (Acharya et al. 2016).

Assumption 2.4. *(No intermediate interactions)*
$E\left(Y_i(d,m) - Y(d,m') \mid D_i = d, L_i = l, X_i = x\right) = E\left(Y_i(d,m) - Y(d,m') \mid D_i = d, X_i = x\right)$
for all $d \in \mathcal{D}, m \in \mathcal{M}, l \in \mathcal{L}$ *and* $x \in \mathcal{X}$

Assumption 2.4 requires the effect of the mediator to be independent from intermediate confounders. If this assumption is violated, the estimated CDE will be a weighted average of CDEs within different levels of confounders (Vansteelandt and Joffe 2014).

Sequential g-estimation proceeds in two steps. In the first stage the causal effect of the mediator is removed from the outcome using a quantity that Acharya et al. (2016) call the demediation function. In the second stage the demediated outcome is regressed on the treatment variable as well as the pretreatment confounders.

The demediation function, $\gamma(d, m, x)$, is defined as the change in the potential outcome when the mediator is switched from m to 0 while holding the treatment constant.[23]

$$\gamma(d,m,x) = E\left(Y_i(d,m) - Y_i(d,0) \mid X_i = x\right) \tag{2.6}$$

By subtracting the demediation function from Y, one obtains the expected potential outcome at treatment level d and mediator value 0 conditional on X (see Appendix B.2).

$$E\left(Y_i - \gamma(d, M_i, x) \mid D_i = d, X_i = x\right) = E\left(Y_i(d,0) \mid X_i = x\right), \tag{2.7}$$

such that the controlled direct at mediator value 0 is given by:

$$\begin{aligned} &E\left(Y_i - \gamma(d, M_i, x) \mid D_i = d, X_i = x\right) - E\left(Y_i - \gamma(d', M_i, x) \mid D_i = d', X_i = x\right) \\ =&E\left(Y_i(d,0) - Y_i(d',0) \mid X_i = X\right) \end{aligned} \tag{2.8}$$

[23] Under Assumption 2.4 this quantity does not depend on L.

Acharya et al. (2016) propose to estimate the demediation function by regressing Y on T, M, X and L. For example specifying

$$E\left(Y_i \mid D_i{=}d, M_i{=}m, X_i{=}x, L_i{=}l\right) = \delta_0{+}\delta_d D_i{+}\delta_m M_i{+}\delta_x X_i{+}\delta_l L_i \tag{2.9}$$

will result in a demediation function of:

$$E\left(Y_i \mid D_i{=}d, M_i{=}m, X_i{=}x, L_i{=}l\right) - E\left(Y_i \mid D_i{=}d, M_i{=}0, X_i{=}x, L_i{=}l\right) = \delta_m M_i \tag{2.10}$$

One may also allow for interaction between the mediator and the treatment or the mediator and the pretreatment confounders by including the respective interaction terms in the demediation function. In the second stage, the demediated outcome is regressed on the treatment variable and the pretreatment confounders. The $CDE(d,0,0)$ is then obtained as the coefficient of the treatment variable:

$$Y_i^{dem}{=}\beta_0{+}\beta_{CDE} * D_i{+}\beta_x * X_i{+}\epsilon_i, \tag{2.11}$$

where Y_i^{dem} is the demediated outcome $Y_i - \hat{\gamma}(d,m,x)$. While sequential g-estimation is easy to perform and can also be applied in the estimation of certain nonlinear models (see for example Vansteelandt 2009), this method requires the correct parametric specification of the conditional means. Nevertheless Goetgeluk et al. (2008) show that the sequential g-estimator remains unbiased even in the presence of certain types of missspecification. Similarly Huber et al. (2016) compare the performance estimators for the NDE using Monte-Carlo simulations based on empirical data and find that a parametric g-estimator slightly outperforms several more flexible semiparametric estimators.

2.5.2 Implementation

Assumption 2.1 requires potential health outcomes to be independent of whether an individual lived in East or West Germany prior to 1989. This follows from the assumption that the German separation provides a valid natural experiment - a claim previous literature has supported. On the other hand, Assumption 2.2 is not backed by any underlying experiment in my setting and therefore the mediator might be correlated with potential outcomes. While I might not be able to fully recover causal mechanisms, the results are still informative as they describe to which extend differences in contemporary factors account for observed differences in health.

In a straightforward application of the framework outlined above I obtain the demediating function separately for each cohort group, where I allow for interaction between `East` and the mediator. I estimate the following equation using pooled OLS:

$$Y_{it} = \delta_{0c} + \delta_{ec}\text{East}_i + \delta_{mc}\text{M}_{it} + \delta_{mdc}\text{M}_{it} \cdot \text{East}_i + \delta_{Xc}X_i + \delta_{Lc}L_{it} + \delta_t + \epsilon_{it}, \tag{2.12}$$

where δ_c denotes the coefficient of cohort c. The demediation function is given by: $\gamma_c(D_i, M_{it}, \hat{\delta}) = \hat{\delta}_{mc}\text{M}_{it} + \hat{\delta}_{mdc}\text{M}_{it} \cdot \text{East}_i$. In this specification identification relies on a selection on observables assumption. The demediation function will be estimated consistently if X_i and L_i include all factors which are correlated with the mediator and do have an influence on individual health. In practice this claim might be problematic. For income I expect the effects based on this demediation function to provide a lower bound. One generally assumes that income has an non-negative causal effect on health. However, as a consequence of unobserved heterogeneity the correlation between income and health is greater than the causal effect leading to an upward bias of the OLS-estimator and also affecting the demediation function. For observations with a value of the mediator greater than zero, the demediation function will be biased upwards resulting in a estimated demediated outcome smaller than the true demediated outcome - and vice versa for observations with a value of the mediator smaller than zero. The magnitude of the bias increases with the absolute value of the mediator. East Germans have on average lower incomes than West Germans. Therefore - assuming that the bias of the estimated demediation function is similar for East and West Germans - the estimated demediated outcome understates the difference between East and West Germans. Consequently the estimated direct effect will be biased towards zero.[24] In a second approach I exploit the longitudinal structure of my data in order to mitigate

[24] This argument does not necessarily hold for other mediators. For example when investigating the effect of unemployment on mental health Farré et al. (2015) find IV estimates to be larger than OLS estimates. In Marcus (2013), on the other hand, simple comparison of the treatment and control group lead to similar conclusions as the preferred specification which uses regression adjustment. In the case of health behaviors it is unclear whether reenforcing or compensating behavior prevails. Therefore the sign of the bias of the OLS-coefficient in the first stage is hard to determine.

the problem of unobserved individual heterogeneity. I apply within transformation and estimate the following equation:

$$\tilde{Y_{it}}=\delta_{mc}\tilde{\mathrm{M}}_{it}+\delta_{mdc}\tilde{\mathrm{M}}_{it}\cdot\text{East}+\delta_{Lc}\tilde{L_{it}}+\delta_t+\tilde{\epsilon}_{it}, \tag{2.13}$$

where δ_t denotes time-fixed effects and $\tilde{Y_{it}}$, $\tilde{\mathrm{M}}_{it}$ and $\tilde{L_{it}}$ denote the deviation from the individual mean in year t . Here identification stems from variation in the mediator within individuals. As `East` is constant within individuals, this variable drops out of the equation.[25] Nevertheless the treatment status is held constant implicitly. The demediation function is obtained as $\hat{\delta}^{FE}_{mc}M_{it}+\hat{\delta}^{FE}_{mdc}M_{it}\cdot\text{East}$, where $\hat{\delta}^{FE}_{mc}$ is the coefficient of M from the fixed effects estimator for cohort c. One potential drawback of this procedure is that it does not take into account that the mediator might be influenced by past health status. The fixed effect estimator cannot account for this type of dynamics (Imai and Kim 2016). Nevertheless static panel data models have been applied to estimate the causal impact of income on health (see for example Frijters et al. 2004). I estimate the CDE for the whole observation period using pooled OLS as well as on yearly cross-sections. When estimating the CDE on yearly cross-sections, I only use the respective wave to estimate Equation 2.12. I cluster all standard errors at the household of origin level (see Appendix B.3).

Definition of mediator and intermediate variables

My measure of income is based on monthly net household income as collected in the SOEP. Since households tend to share expenditures I calculate the log of net household income per household member rather than directly using individual income. This definition also alleviates the problem of unobserved dynamics as household income is influenced by an individual's past health status to a lesser extent than individual income. The procedure outlined above estimates the controlled direct effect when the mediator variable takes the value zero. Since the CDE at a household income of one Euro does not constitute a quantity of Economic interest, I normalize the log household income[26] per household member by subtracting the median of the respective survey year for the

[25] Since I only condition on non-time varying pretreatment variables X, this vector also drops out.

[26] For example Chetty et al. (2016) find a concave relationship between income and life expectancy. The use log income rather than the level of income is intended to capture such a relationship.

respective sex and five-year birth cohort. Therefore the controlled direct effect I estimate is the effect of having lived in the GDR when the income is at the median of the income distribution for the respective survey year in the respective demographic group. Unemployment is measured as being currently registered unemployed at the German Federal Employment agency. Here I obtain the controlled direct effect for individuals who are currently not registered unemployed. I exclude all individuals above the age of 65, the statutory retirement age prior to 2012.[27]

When using health behaviors as mediators, I consider level of exercise and attachment to a healthy diet. I create a binary variable that takes the value one if an individual engages in physical exercise at least once a week. Attachment to a healthy diet is captured by a dummy variable that takes the value one if an individual agrees strongly or very strongly that a healthy diet is important.[28]

Finally I am interested in the question, whether the disadvantage in health of East Germans is driven by a feeling of not being in control of one's life. To combine the seven items measuring locus of control into a single measure, I apply PCA. Economists usually assume personality traits to be time-invariant. Cobb-Clark and Schurer (2013) find that locus of control indeed appears to be relatively stable in the short and medium run for most of the population, but advise researchers to account for aging related changes. Therefore I standardize the resulting score to have a mean of zero for each year-sex-5 years birth-cohort cell. As variation in locus of control will to a large extend reflect measurement error rather than true variation, I only estimate the CDE using the cross-sectional demediation function.

I use the same set of pretreatment variables as in Section 2.4. Thus X includes dummies for year of birth in 5-year categories, linear year of birth and sex in all specifications and additionally maternal and paternal education (in 5 categories), year of birth of parents and the number of siblings in some specifications. The set of intermediate controls L comprises factors, which are itself influenced by the fact whether an individual is from East or West Germany, do affect the mediators and influence health. Most importantly

[27] From 2012 onwards the retirement age is raised by one month every year.

[28] Information on health behavior is not collected in each wave of the GSOEP. Whenever possible I use information on health behavior collected in the same year as the health outcome. Otherwise I use information on health behavior that was collected in the previous wave. When no information on health behavior exists in both the current and the previous year, I discard the observation.

L includes the level of education. In East Germany high ability individuals could be denied access to further education because political non-conformity or because of their parental background, which leads to different patterns of sorting between East and West Germany. In order to control for potential labor market status, I also include whether an individual is currently in education or has reached the statutory retirement age. I further control for marital status and the number of children,[29] the willingness to take risk as previous research has revealed different attitudes towards risky behavior between East and West German individuals (Ziebarth and Wagner 2013) and the number of individuals living in a household.

2.5.3 Results

Income

When income per household member is held constant at the median of the respective survey year, the gap in health between East and West Germans narrows. Nevertheless for the oldest cohort group there clearly exists an effect of having lived in the GDR on health that does not solely operate through income. Table 2.6 presents the estimates of the CDE for both specifications of the demediation function as well as the original East-West gap estimated on the same sample as the CDE.[30] Columns (1)-(3) show the results separately for each cohort group pooling men and women, while in columns (4)-(9) I split the sample by sex. For the oldest cohort group the estimated direct effects are significant throughout all outcomes and for both men and women. Comparing the total effect to the CDE based on the demediation function estimated using pooled-OLS, the estimated effect size decreases from 11.4% of a standard to 8.2% of a standard deviation for MCS, from 17.1% to 16.1% of a standard deviation for PCS, it roughly stays constant for the number of diagnoses and decreases from from -0.11 to -0.09 for SAH (see column (1)). With the exception of the number of diagnoses, the estimated total effect is greater for males than for females. The estimated CDEs preserve this pattern. In the middle cohort on the other hand, holding income fixed removes most differences between East and West Germans. The effect of having lived in the GDR

[29] In the GDR there existed strong legal and financial incentives to get married. Klärner (2015) find that East Germans perceive post reunification incentives as negligible.

[30] This sample is smaller than the sample used in Section 2.4 as incorporating the mediator and intermediate variables increases the number of observations with missing values.

on mental health differs between males and females. For females both the original gap as well as the estimated CDEs are significant, while for males the original gap is insignificant and the sign of the CDE turns positive. In the young cohort differences between East and West are mostly negligible.

I additionally examine how the CDE evolves over time (see Figure 2.11). The estimated CDE tends to be smaller than the original gap but follows a similar trend. Most notably a widening of the gap between East and West for PCS in the middle cohort group, which I have documented above, is visible again but less strong. Differences between the original gap and the CDE are strongest for the middle cohort and more pronounced for PCS than for other outcomes.

Unemployment

The effect of having lived in the GDR that does not operate via unemployment is of similar size as the total effect (see Table 2.7). Here the sample size for the oldest cohort group is smaller than in other analyzes since I only include individuals aged 65 or younger. For some outcomes and samples the CDE is even larger than the total effect, however the difference seems negligible in most cases.[31] Figure 2.11 depicts estimated effects over time.

As the unemployment gap between East and West Germans narrows after 2006 (see Figure 2.2) one would expect the CDE to become more similar to the total effect over time. However already prior to 2006 the CDE and the total effect differ only marginally and no changes over time are visible. The difference between the estimated CDE and the original gap stems from individuals whose value of the mediator valuable is originally not at the reference level. 95% (95.5% of West Germans and 90.5% of East Germans) of the (pooled) sample are not unemployed, so that their contribution to the estimating equation in the second stage of the CDE is exactly the same as when estimating the original gap. The remaining 5% who are unemployed do not appear to be driving the gap in health between East and West Germans. Additionally there might be spillover effects from one unemployed person in the household to other household members (see for example Marcus (2013)). Thus I also estimate the controlled direct effect for indi-

[31] An exception are men of middle and young cohort groups, where the CDE indicates a larger effect on mental health than the original effect.

Table 2.6: *Controlled direct effect: Mediator: Log of net household income per household member*

Mental health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.8185***	-0.1486	-0.1468	-0.7298*	-0.8507***	-0.4657	-0.8979**	0.5278	0.1253
	(0.3091)	(0.2588)	(0.3623)	(0.3921)	(0.3273)	(0.4856)	(0.3888)	(0.3487)	(0.4823)
CDE-FE	-0.9678***	-0.3999	-0.3185	-0.68*	-1.0938***	-0.7142	-1.3797***	0.2709	7e-04
	(0.3055)	(0.2699)	(0.3577)	(0.3862)	(0.3545)	(0.481)	(0.3953)	(0.3528)	(0.4822)
TE	-1.1366***	-0.6991***	-0.357	-0.9579**	-1.4221***	-0.766*	-1.3551***	-0.0223	0.0219
	(0.2972)	(0.2647)	(0.3486)	(0.3725)	(0.35)	(0.4661)	(0.3714)	(0.3409)	(0.471)
N	34639	54970	19831	17696	28786	10818	16943	26184	9013

Physical health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-1.6038***	-0.2731	-0.2948	-1.5185***	-0.3378	-0.3956	-1.6749***	-0.2062	-0.2369
	(0.2743)	(0.2239)	(0.2628)	(0.3661)	(0.2953)	(0.3219)	(0.3715)	(0.3036)	(0.3904)
CDE-FE	-1.7202***	-0.7619***	-0.4023	-1.4396***	-0.8499***	-0.4103	-2.1022***	-0.6741**	-0.4005
	(0.2765)	(0.2371)	(0.2694)	(0.3584)	(0.3178)	(0.3275)	(0.4215)	(0.312)	(0.4104)
TE	-1.7084***	-0.8479***	-0.4406*	-1.4863***	-0.7736**	-0.4487	-1.9827***	-0.9348***	-0.4439
	(0.263)	(0.2293)	(0.2526)	(0.3514)	(0.3037)	(0.3161)	(0.3589)	(0.303)	(0.3709)
N	34639	54970	19831	17696	28786	10818	16943	26184	9013

Number of diagnoses

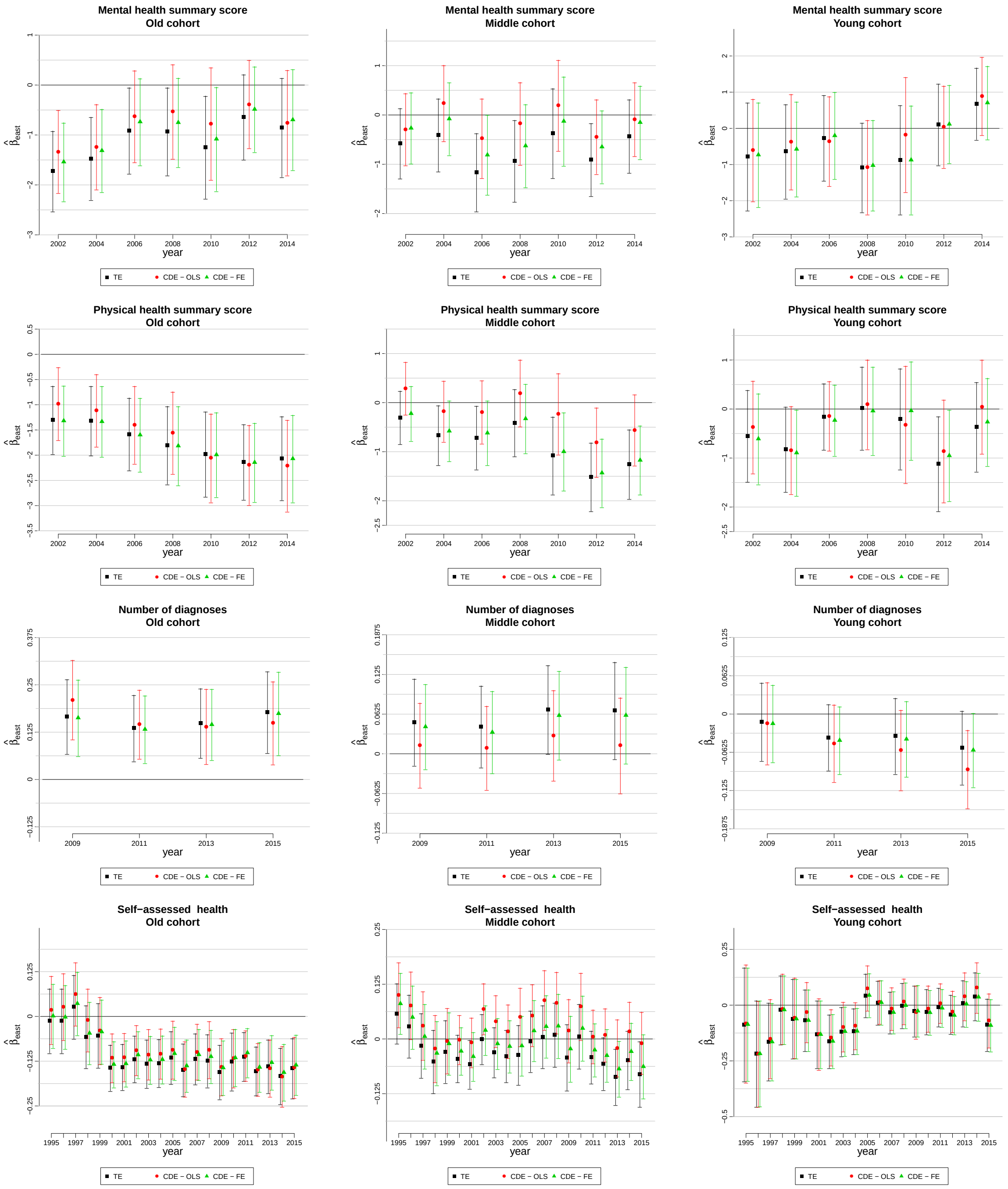
	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	0.1601***	0.0163	-0.0555**	0.1678***	0.0127	-0.0615*	0.1464**	0.0193	-0.0482*
	(0.0411)	(0.0297)	(0.022)	(0.053)	(0.0373)	(0.0345)	(0.0598)	(0.0443)	(0.0287)
CDE-FE	0.1536***	0.0494*	-0.0411*	0.1591***	0.0489	-0.0622*	0.1465**	0.0511	-0.0244
	(0.0409)	(0.0297)	(0.0218)	(0.0531)	(0.039)	(0.0347)	(0.0568)	(0.0435)	(0.028)
TE	0.1568***	0.0575*	-0.037*	0.1603***	0.0555	-0.0509	0.1508***	0.0609	-0.024
	(0.0394)	(0.0298)	(0.0211)	(0.051)	(0.0389)	(0.0332)	(0.0555)	(0.0435)	(0.0275)
N	18868	31352	12529	9746	16636	6970	9122	14716	5559

Self-assessed health

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.0905***	0.0312*	-0.0164	-0.0868***	0.008	-0.0134	-0.0906***	0.0551**	-0.0195
	(0.0232)	(0.0188)	(0.0284)	(0.0303)	(0.0243)	(0.0296)	(0.0317)	(0.0258)	(0.0454)
CDE-FE	-0.0993***	-0.0089	-0.0384	-0.0842***	-0.0438*	-0.0296	-0.1182***	0.0242	-0.0478
	(0.0232)	(0.0197)	(0.0289)	(0.0305)	(0.026)	(0.0297)	(0.031)	(0.0265)	(0.0465)
TE	-0.1102***	-0.0289	-0.0376	-0.0949***	-0.0472*	-0.031	-0.1289***	-0.012	-0.0424
	(0.023)	(0.0195)	(0.0284)	(0.0301)	(0.0253)	(0.0292)	(0.0306)	(0.0261)	(0.0454)
N	93134	156215	55260	47940	82266	31224	45194	73949	24036

Notes: Controlled direct effect: Effect of having lived in the GDR in 1989 when keeping the level of household income per household member at the median of the respective sample year. CDE-OLS denotes the estimated direct effect when the demediation function is estimated using Equation 2.12. CDE-FE denotes the estimated direct effect when the demediation function is estimated Equation 2.13. TE denotes the total effect of having lived in East Germany estimated on the same sample as CDE-OLS and CDE-FE. Controls in the second stage include dummies for age of birth in 5-year categories, a linear trend in age and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. Regressions are weighted using cross sectional weights provided by the SOEP. All standard errors are clustered at the household of orign level.

***Figure 2.11:** Controlled direct effect: Income as mediator*



Notes: Controlled direct effect: Effect of having lived in the GDR in 1989 when keeping the level of household income per household member at the median of the respective sample year. CDE-OLS denotes the estimated direct effect when the demediation function is estimated using Equation 2.12 on yearly samples. CDE-FE denotes the estimated direct effect when the demediation function is estimated Equation 2.13. TE denotes the total effect of having lived in East Germany estimated on the same sample as CDE-OLS and CDE-FE. Controls in the second stage include dummies for year of birth in 5-year categories, a linear trend in birth year and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. Regression are weighted using cross sectional weights provided by the SOEP. All standard errors are clustered at the household of origin level.

Table 2.7: *Controlled direct effect: Mediator: Unemployed*

Mental health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-1.0791**	-0.6694**	-0.4678	-1.1158*	-1.2302***	-0.7557	-1.056*	-0.1379	-0.1825
	(0.4542)	(0.2624)	(0.3612)	(0.5972)	(0.3517)	(0.4781)	(0.6418)	(0.3355)	(0.4977)
CDE-FE	-0.9547**	-0.6604**	-0.5072	-0.9989*	-1.286***	-0.7559	-0.9266	-0.0835	-0.2654
	(0.4569)	(0.2697)	(0.3684)	(0.5949)	(0.3557)	(0.4874)	(0.6639)	(0.348)	(0.5086)
TE	-1.1823***	-0.6903***	-0.3548	-1.2595**	-1.4262***	-0.7729*	-1.0927*	-0.0024	0.0338
	(0.4554)	(0.2653)	(0.3494)	(0.591)	(0.3509)	(0.466)	(0.6443)	(0.3411)	(0.4728)
N	9906	54769	19839	4945	28685	10822	4961	26084	9017

Physical health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-2.2979***	-0.753***	-0.3199	-1.9834***	-0.6842**	-0.3689	-2.6283***	-0.8149***	-0.2704
	(0.454)	(0.2234)	(0.2524)	(0.6127)	(0.3028)	(0.3198)	(0.6329)	(0.2918)	(0.3701)
CDE-FE	-2.1271***	-0.7878***	-0.423	-1.823***	-0.8194***	-0.446	-2.4832***	-0.7794**	-0.4073
	(0.4596)	(0.2331)	(0.2833)	(0.6127)	(0.3166)	(0.3257)	(0.6443)	(0.3124)	(0.4377)
TE	-2.1963***	-0.8558***	-0.4362*	-1.9054***	-0.7788**	-0.4473	-2.5139***	-0.945***	-0.4364
	(0.442)	(0.2292)	(0.2526)	(0.6019)	(0.3037)	(0.3159)	(0.5993)	(0.3032)	(0.3711)
N	9906	54769	19839	4945	28685	10822	4961	26084	9017

Number of diagnoses

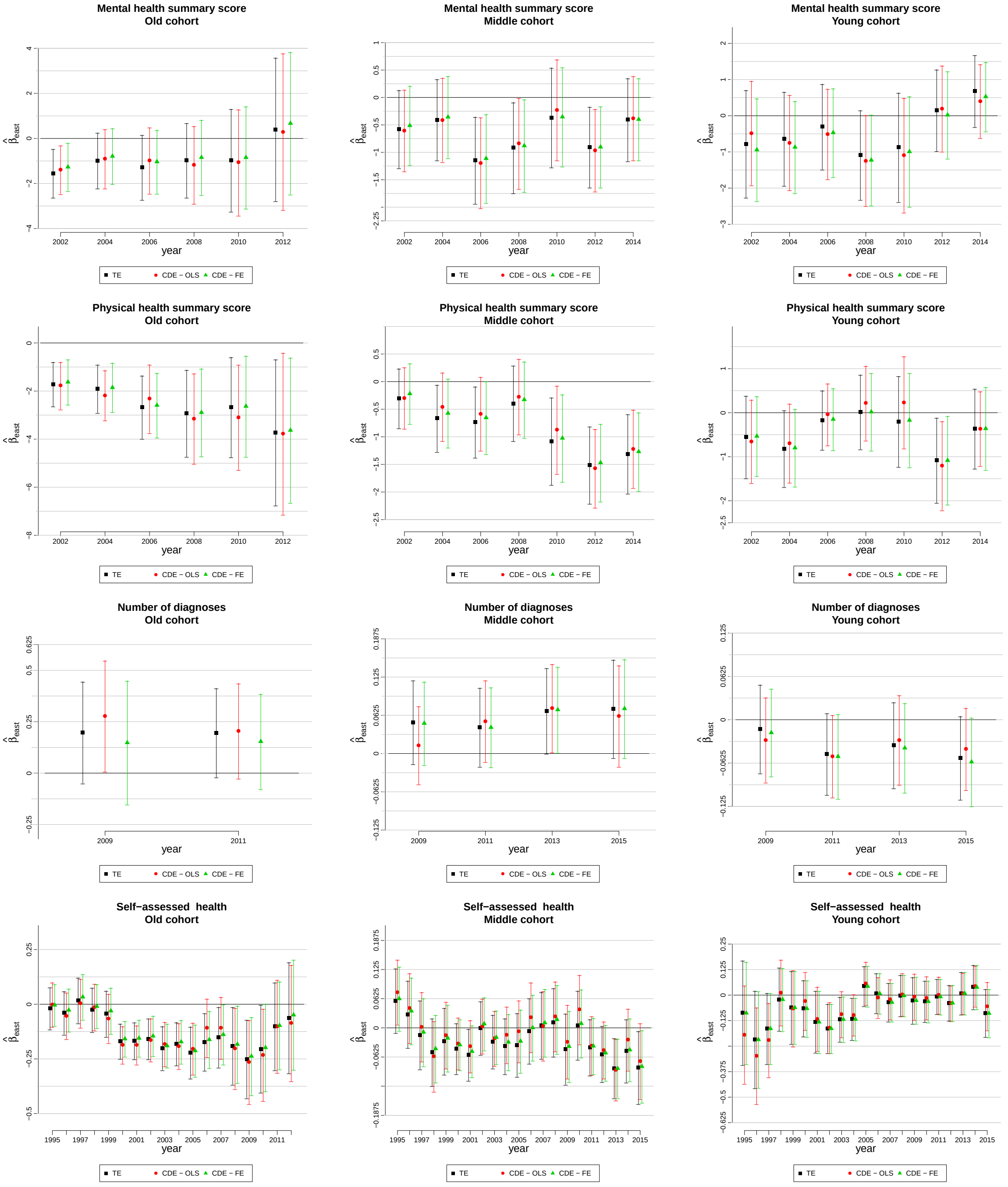
	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	0.2403**	0.0498*	-0.0398*	0.1505	0.0355	-0.0555*	0.3531**	0.0638	-0.0277
	(0.1074)	(0.0294)	(0.0216)	(0.1368)	(0.0369)	(0.0338)	(0.1559)	(0.0441)	(0.0282)
CDE-FE	0.154	0.0591*	-0.0445*	0.121	0.0542	-0.0465	0.2087	0.0656	-0.0442
	(0.1187)	(0.0303)	(0.0232)	(0.1464)	(0.0387)	(0.0335)	(0.1559)	(0.045)	(0.0324)
TE	0.2002**	0.059**	-0.0399*	0.1561	0.0563	-0.0512	0.2629*	0.0623	-0.0293
	(0.1001)	(0.03)	(0.0215)	(0.1272)	(0.0393)	(0.0332)	(0.1468)	(0.0438)	(0.0285)
N	1726	30880	12533	888	16402	6971	838	14478	5562

Self-assessed health

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.11***	-0.0151	-0.0285	-0.106**	-0.028	-0.0243	-0.1137***	-0.0026	-0.0292
	(0.0334)	(0.019)	(0.0262)	(0.0469)	(0.0249)	(0.0294)	(0.0409)	(0.0253)	(0.0406)
CDE-FE	-0.0957***	-0.0233	-0.0378	-0.0841*	-0.0431*	-0.0353	-0.1082**	-0.0057	-0.0372
	(0.0333)	(0.0196)	(0.0294)	(0.0459)	(0.0254)	(0.0302)	(0.042)	(0.0263)	(0.0473)
TE	-0.1084***	-0.029	-0.0372	-0.0976**	-0.0475*	-0.0309	-0.1196***	-0.012	-0.0418
	(0.0333)	(0.0195)	(0.0284)	(0.0455)	(0.0253)	(0.0292)	(0.0419)	(0.0262)	(0.0454)
N	34333	155544	55281	17374	81932	31234	16959	73612	24047

Notes: Controlled direct effect: Effect of having lived in the GDR in 1989 when an individual is currently not registered unemployed. The sample includes only individuals aged 65 or younger. CDE-OLS denotes the estimated direct effect when the demediation function is estimated using Equation 2.12. CDE-FE denotes the estimated direct effect when the demediation function is estimated Equation 2.13. TE denotes the total effect of having lived in East Germany estimated on the same sample as CDE-OLS and CDE-FE. Controls in the second stage include dummies for age of birth in 5-year categories, a linear trend in age and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. Regressions are weighted using cross sectional weights provided by the SOEP. All standard errors are clustered at the household of orign level.

Figure 2.12: Controlled direct effect: Unemployment as mediator



Notes: Controlled direct effect: Effect of having lived in the GDR in 1989 for an individual who is currently not registered unemployed. CDE-OLS denotes the estimated direct effect when the demediation function is estimated using Equation 2.12 on yearly samples. CDE-FE denotes the estimated direct effect when the demediation function is estimated Equation 2.13. TE denotes the total effect of having lived in East Germany estimated on the same sample as CDE-OLS and CDE-FE. Controls in the second stage include dummies for year of birth in 5-year categories, a linear trend in birth year and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. Regression are weighted using cross sectional weights provided by the SOEP. All standard errors are clustered at the household of origin level.

viduals who are neither unemployed themselves nor live with an currently unemployed household member, extending the pool of observations for whom the mediator variable is not at the reference level. The estimated CDEs reduce slightly but tend to be very close to the estimated total effect again (see Table B.4 in the Appendix).

Health behavior

When attachment to a healthy diet is held at a constant level, the estimated effect of having lived in East Germany reduces for most outcomes and subsamples (see Table 2.8). The change between the CDEs and the total effect is, however, quite heterogeneous. For the oldest cohort group the CDE is no longer significant for mental health - an effect mostly driven by the female sample. It is larger than the total effect for PCS and stays close to the total effect when looking at SAH. For the number of diagnoses the result is inconclusive. In the middle cohort group the CDEs are usually smaller than the total effect, but remain significant in the case of PCS for the whole sample and the male subsample. Finally I estimate the direct effect for individuals who do not exercise at least every week (see Table 2.9). For most outcomes and samples the estimated CDEs are not as strong as the original effect but remain significant when the original gap was significant. All in all I my findings suggest that health behaviors might be contributing to health disparities between East and West as erasing differences in attachment to a healthy diet does eliminate a large part of the gap.

Locus of control

The CDE when holding the score of locus of control at its mean does not differ substantially from the total effect in the respective subsample (see Table 2.10). Consequently I conclude that the feeling of not being in full control over one's life does not account for a major share of the disadvantage in health of East Germans. Interestingly the largest change in the coefficient takes place in the case of mental health. This hints to the explanation that a low locus of control hurts the mental dimension of health stronger than the physical dimension. However, the result is not strong enough to draw a final conclusion.

Table 2.8: *Controlled direct effect: Mediator: Healthy diet is important*

Mental health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.6463	-0.0942	0.346	-0.1324	-1.1006**	-0.0346	-1.1061**	0.5698	0.5705
	(0.4205)	(0.3513)	(0.4164)	(0.6117)	(0.522)	(0.6121)	(0.4873)	(0.4244)	(0.5335)
CDE-FE	-0.6783*	-0.6521**	-0.1583	-0.3002	-1.4372***	-0.4905	-1.0805**	0.0626	0.1715
	(0.4118)	(0.3249)	(0.4767)	(0.5802)	(0.4487)	(0.6742)	(0.4685)	(0.4144)	(0.6188)
TE	-1.0017***	-0.7337***	-0.3053	-0.7888**	-1.4298***	-0.7182	-1.2604***	-0.0819	0.0836
	(0.311)	(0.2778)	(0.3682)	(0.3914)	(0.3637)	(0.492)	(0.3845)	(0.3647)	(0.4959)
N	28523	45887	17137	14581	24103	9378	13942	21784	7759

Physical health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-2.0354***	-0.8087***	-0.3998	-1.8455***	-0.4909	-0.7866*	-2.2052***	-1.0699***	-0.2164
	(0.3764)	(0.2777)	(0.3215)	(0.5321)	(0.4104)	(0.4704)	(0.5104)	(0.3447)	(0.4319)
CDE-FE	-2.1682***	-0.8399***	-0.4641	-1.6925***	-0.4603	-0.9204*	-2.6786***	-1.144***	-0.1942
	(0.3405)	(0.2726)	(0.3315)	(0.4653)	(0.3848)	(0.4822)	(0.4577)	(0.3408)	(0.4375)
TE	-1.818***	-0.9148***	-0.465*	-1.5496***	-0.7876**	-0.4778	-2.1487***	-1.0501***	-0.4639
	(0.2756)	(0.2468)	(0.2723)	(0.3666)	(0.3243)	(0.3322)	(0.3782)	(0.3266)	(0.4081)
N	28523	45887	17137	14581	24103	9378	13942	21784	7759

Number of diagnoses

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	0.1008	0.0465	-0.0416	0.0934	0.0464	-0.0539	0.1118	0.0493	-0.0315
	(0.062)	(0.036)	(0.03)	(0.0765)	(0.051)	(0.0497)	(0.087)	(0.0491)	(0.0376)
CDE-FE	0.1795***	0.0346	-0.0301	0.1497**	-0.0024	-0.0764	0.2078***	0.067	-0.0062
	(0.0518)	(0.0354)	(0.0313)	(0.0685)	(0.0475)	(0.0509)	(0.0708)	(0.051)	(0.0377)
TE	0.1631***	0.0616*	-0.0421*	0.1551***	0.0533	-0.0529	0.1718***	0.0708	-0.0317
	(0.0417)	(0.0319)	(0.0236)	(0.0539)	(0.0421)	(0.0357)	(0.0588)	(0.0464)	(0.0312)
N	16665	27610	10593	8592	14649	5902	8073	12961	4691

Self-assessed health

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.1349***	0.0094	-0.0053	-0.1048**	-0.0028	-0.0041	-0.163***	0.012	-0.0066
	(0.0353)	(0.0281)	(0.0358)	(0.0467)	(0.0382)	(0.0446)	(0.0506)	(0.0358)	(0.0505)
CDE-FE	-0.1462***	-0.008	-0.0438	-0.1205***	0.0017	-0.0625	-0.1807***	-0.014	-0.0254
	(0.0314)	(0.0262)	(0.0364)	(0.041)	(0.0353)	(0.0464)	(0.0429)	(0.0332)	(0.0547)
TE	-0.1362***	-0.0312	-0.0235	-0.1017***	-0.0455	-0.0279	-0.1795***	-0.019	-0.0177
	(0.0256)	(0.024)	(0.034)	(0.0327)	(0.0314)	(0.0342)	(0.0356)	(0.0322)	(0.0544)
N	56572	89761	32877	29051	47284	18037	27521	42477	14840

Notes: Controlled direct effect: Effect of having lived in the GDR when a healthy diet is not important CDE-OLS denotes the estimated direct effect when the demediation function is estimated using Equation 2.12. CDE-FE denotes the estimated direct effect when the demediation function is estimated Equation 2.13. TE denotes the total effect of having lived in East Germany estimated on the same sample as CDE-OLS and CDE-FE. Controls in the second stage include dummies for age of birth in 5-year categories, a linear trend in age and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. Regressions are weighted using cross sectional weights provided by the SOEP. All standard errors are clustered at the household of orign level.

Table 2.9: Controlled direct effect: Mediator: Exercise every week

Mental health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.918**	-0.708**	-0.1603	-0.6731	-1.5145***	-0.4309	-1.2182***	-0.0038	8e-04
	(0.37)	(0.3475)	(0.5103)	(0.4675)	(0.4765)	(0.7117)	(0.4657)	(0.4353)	(0.6629)
CDE-FE	-0.7741**	-0.7819**	-0.5931	-0.4211	-1.6192***	-1.1399	-1.195***	0.0122	0.0011
	(0.3568)	(0.3195)	(0.5348)	(0.4463)	(0.4243)	(0.6989)	(0.4557)	(0.4239)	(0.739)
TE	-1.0833***	-0.8167***	-0.6871	-0.7946**	-1.5765***	-1.049*	-1.4284***	-0.0984	-0.3541
	(0.32)	(0.2994)	(0.4211)	(0.3998)	(0.3954)	(0.5898)	(0.4035)	(0.3885)	(0.5497)
N	22714	34087	11067	11677	17753	5940	11037	16334	5127

Physical health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-1.399***	-0.8607***	-0.4767	-1.1503***	-0.5819	-0.0536	-1.7602***	-1.1585***	-0.9684*
	(0.3141)	(0.2966)	(0.3784)	(0.4133)	(0.3956)	(0.4805)	(0.4393)	(0.4031)	(0.5603)
CDE-FE	-1.7055***	-0.9098***	-0.7982**	-1.3973***	-0.7865**	-0.1819	-2.0897***	-1.0882***	-1.6718***
	(0.3011)	(0.2822)	(0.3771)	(0.3909)	(0.3599)	(0.4812)	(0.4248)	(0.3985)	(0.5411)
TE	-1.8163***	-0.8914***	-0.418	-1.5653***	-0.745**	-0.4738	-2.1258***	-1.0502***	-0.4056
	(0.2855)	(0.2536)	(0.2881)	(0.3805)	(0.3324)	(0.3907)	(0.3946)	(0.3441)	(0.4073)
N	22714	34087	11067	11677	17753	5940	11037	16334	5127

Number of diagnoses

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	0.1116**	0.0728*	-0.0582**	0.118**	0.0899*	-0.0748**	0.1021	0.0593	-0.0422
	(0.046)	(0.0376)	(0.029)	(0.0597)	(0.052)	(0.037)	(0.0667)	(0.0529)	(0.0436)
CDE-FE	0.1367***	0.048	-0.0376	0.1393**	0.0644	-0.0719*	0.1251**	0.0329	0.0096
	(0.0434)	(0.0329)	(0.0307)	(0.0594)	(0.0445)	(0.0397)	(0.0614)	(0.0477)	(0.0476)
TE	0.1668***	0.058*	-0.0402*	0.1859***	0.0564	-0.0542*	0.1413**	0.0612	-0.0271
	(0.0402)	(0.03)	(0.0218)	(0.0513)	(0.0398)	(0.0327)	(0.0572)	(0.0435)	(0.0297)
N	14004	23780	9743	7220	12598	5442	6784	11182	4301

Self-assessed health

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.057**	-0.0067	-0.009	-0.0421	-0.0196	0.0171	-0.0761**	0.0027	-0.0372
	(0.0268)	(0.0232)	(0.0364)	(0.0347)	(0.0303)	(0.0378)	(0.0356)	(0.0317)	(0.0593)
CDE-FE	-0.0932***	-0.0224	-0.0266	-0.0758**	-0.0403	-0.0238	-0.1148***	-0.0085	-0.0293
	(0.0246)	(0.0214)	(0.0326)	(0.0319)	(0.0279)	(0.0368)	(0.0337)	(0.0296)	(0.0492)
TE	-0.0986***	-0.0237	-0.039	-0.0808**	-0.0413	-0.0377	-0.12***	-0.0077	-0.0396
	(0.0241)	(0.0204)	(0.0296)	(0.0315)	(0.0264)	(0.0318)	(0.0322)	(0.0276)	(0.0467)
N	67195	109579	35535	34715	57490	19914	32480	52089	15621

Notes: Controlled direct effect: Effect of Exercise every week CDE-OLS denotes the estimated direct effect when the demediation function is estimated using Equation 2.12. CDE-FE denotes the estimated direct effect when the demediation function is estimated Equation 2.13. TE denotes the total effect of having lived in East Germany estimated on the same sample as CDE-OLS and CDE-FE. Controls in the second stage include dummies for age of birth in 5-year categories, a linear trend in age and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. Regressions are weighted using cross sectional weights provided by the SOEP. All standard errors are clustered at the household of origin level.

Table 2.10: *Controlled direct effect: Mediator: Locus of control*

Mental health summary score									
	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.912**	-0.4951	-0.7655	-0.8737*	-1.4449***	-1.6454**	-0.9063*	0.3694	0.0373
	(0.3947)	(0.3324)	(0.5149)	(0.4973)	(0.4456)	(0.7549)	(0.486)	(0.4616)	(0.6539)
TE	-1.1317***	-0.7603**	-0.7222	-0.9572*	-1.7296***	-1.5203*	-1.3333**	0.1388	-0.0517
	(0.4189)	(0.3708)	(0.5584)	(0.5238)	(0.485)	(0.7789)	(0.5306)	(0.5013)	(0.7318)
N	8718	13327	4607	4452	6895	2440	4266	6432	2167
Physical health summary score									
	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-1.7265***	-0.8812***	-0.1386	-1.5382***	-0.6965*	-0.1451	-1.9551***	-1.0739***	-0.1739
	(0.345)	(0.3041)	(0.3601)	(0.4545)	(0.4048)	(0.52)	(0.4929)	(0.4107)	(0.4897)
TE	-1.7683***	-0.9654***	-0.117	-1.5548***	-0.7454*	-0.1197	-2.0411***	-1.1981***	-0.1565
	(0.3437)	(0.318)	(0.3659)	(0.4544)	(0.4192)	(0.5182)	(0.4822)	(0.4289)	(0.5019)
N	8718	13327	4607	4452	6895	2440	4266	6432	2167
Number of diagnoses									
	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	0.1492***	0.0543	-0.0659**	0.1528**	0.0646	-0.0737*	0.1388**	0.0476	-0.0544
	(0.047)	(0.0346)	(0.0267)	(0.0615)	(0.0443)	(0.038)	(0.0641)	(0.0502)	(0.0364)
TE	0.1554***	0.0633*	-0.0653**	0.1568**	0.0724	-0.0726*	0.152**	0.0573	-0.0578
	(0.0473)	(0.0351)	(0.0267)	(0.0614)	(0.0446)	(0.0381)	(0.0655)	(0.0511)	(0.037)
N	7788	14615	6269	4003	7770	3556	3785	6845	2713
Self-assessed health									
	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.119***	-0.006	-0.0121	-0.0844**	-0.0188	-0.0194	-0.1587***	0.0044	-0.0045
	(0.0276)	(0.0241)	(0.0368)	(0.0348)	(0.0315)	(0.0392)	(0.0388)	(0.0326)	(0.0575)
TE	-0.1269***	-0.0208	-0.0136	-0.089**	-0.0332	-0.0187	-0.1731***	-0.0117	-0.0085
	(0.0281)	(0.0259)	(0.0375)	(0.0356)	(0.0334)	(0.0398)	(0.0391)	(0.0352)	(0.0588)
N	21887	35937	13612	11202	18816	7440	10685	17121	6172

Notes: Controlled direct effect: Effect of having lived in the GDR when keeping locus of control constant CDE-OLS denotes the estimated direct effect when the demediation function is estimated using Equation 2.12. TE denotes the effect of having lived in East Germany estimated on the same sample as CDE-OLS. Controls in the second stage include dummies for age of birth in 5-year categories, a linear trend in age and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. Regressions are weighted using cross sectional weights provided by the SOEP. All standard errors are clustered at the household of orign level.

2.6 Conclusion

In this work I exploit the German separation and reunification as a natural experiment to study the long-term effects of institutions on health. I document disparities in health between East and West Germans which persist more than two decades after reunification. East German individuals who were born prior to the separation exhibit a lower mental health and physical health summary score, report more diagnoses and rate their health lower than West Germans. Furthermore I seek to disentangle the impact of post-reunification factors from a direct effect of having lived in East Germany. Therefore I estimate direct controlled effects. Holding today's household income, unemployment or the locus of control constant does reduce the size of the estimated gap but a significant direct effect remain throughout. When attachment to a healthy diet is used as mediator, controlled direct effects some outcomes are no longer significant.

In recent years literature in Economics and other social sciences has been highlighting the importance of an early life environment (see for example Almond and Currie 2011,Kesternich et al. 2014,Kesternich et al. 2015). Accordingly one would expect the the impact of having lived in East Germany to be greatest for individuals who were exposed to the Socialist regime from birth onwards. Yet I detect particularly strong and robust disadvantages of East Germans among individuals born in or prior to 1949. Several interpretations are consistent with my findings. Firstly differences in health between East and West might only become apparent at older ages when health shocks occur more frequently. Therefore aging might be associated with a steeper deterioration of health for East Germans than West Germans. I do find some evidence consistent with this hypothesis. When measuring health by the physical health summary score, inequalities between East and West are increasing as people age. Findings are less clear for SAH and the number of diagnoses, while this pattern is not present for MCS. I conclude that age effects most likely contribute to the observed pattern of health inequalities between East and West Germans - especially when using an age sensitive measure of health - but do not explain the whole story. Secondly length of exposure might be a critical factor. Only individuals born in 1949 or earlier were exposed to the GDR for the full 41 years. Moreover the exposure to the GDR might in fact have provided a worse experience for individuals born prior to 1949 than for younger cohorts. Repression was most salient and most brutal during the first years of the GDR. Widespread opposition in the population met with a regime ready to apply Stalinist methods in order to back its claim to power. Also the shortcomings of the medical system worked against older individuals as they are more likely to develop chronic diseases. The lack of modern equipment particularly hit individuals with chronic diseases such as diabetes or patients in need of dialysis.

Additionally the oldest individuals in my sample experienced the extreme circumstances of WWII and the immediate post war period. While East Germans share this experience with their West German counterparts, exposure to these very adverse conditions in early life may have made them more fragile with respect to additional hardship in later life. Formulating this hypothesis in the language of a regression formula, there might be an interaction effect between exposure to WWII in early life and living in East Germany in later life. Other explanations include selection effects induced by large scale migration from East to West Germany prior to 1961 and the idea that older individuals were distressed by the reunification as they might found it difficult to adapt to the system of the FRG. I cannot fully rule out both concerns. However, I show that the treatment effect is robust to the classification of East-West migrants in a selected subsample. Furthermore East Germans born prior to 1949 are more likely to report an increase in life satisfaction since the fall of the Berlin Wall than West Germans from the same cohorts, a finding that is not consistent with the idea of reunification induced distress.

A limitation of the present study is the lack of information about individual live experiences in the GDR. I do not observe which East Germans were subject to repression or suffered from insufficient care. Thus, the exact mechanism, through which the "GDR-effect" operates, remains a question for further research.

Overall, I interpret my findings that the legacy of having lived under socialism in East Germany still provides a burden for individual health. As I document some evidence consistent with a widening in the gap as individuals age, health differences between East and West might not disappear in the immediate future.

Chapter 3

Does the laptop always help? Non-compliance and interviewer effects in cognitive tests

3.1 Introduction

Short cognitive tasks in longitudinal surveys provide social scientists with useful measurements of respondents' ability. At the same time, individual performance in a cognitive test does not only reflect inherent ability but also contextual circumstances. Consequently, researchers aim at administering cognitive tests in a standardized fashion to obtain comparability of test scores across respondents (Lang et al. 2007). This objective is significantly more difficult to achieve in a survey setting, where respondents are visited in their homes, than in a standardized test room environment (Herzog and Rodgers 1999). As longitudinal household studies are typically conducted in interviewer-administered modes, the ultimate responsibility for administrating cognitive tests in a standardized way, is passed on to the interviewer. For interviewers, the administration of a cognitive tests is a demanding task. As a result, observed test scores are likely to reflect heterogeneity in interviewers' ability and interviewing style (Wooden 2013). Moreover, previous literature has raised the concern that interviewer effects might be particularly large in settings, where the interviewer knows the correct answer to a question (T. F. Crossley et al. 2017).

One way to scale down the impact of interviewers is to shift administration of a cognitive test partly or fully to a technical device, such as the interviewer's laptop. While this approach eliminates measurement error caused by variation in interviewer ability, it can also introduce a new source of error. If interviewers do not always comply with the study protocol and administer the test with the help of the technical device for some but not all respondents, the cognitive test will effectively be administered in two distinct modes. The survey methodology literature has shown that the distribution of survey responses can depend on the mode of survey administration, so called "mode effects" (Al Baghal 2017; Cernat et al. 2016; Jäckle et al. 2010).

In this work, we study test scores of the word recall test, collected in the third wave of Understanding Society, the UK Household Longitudinal Study (McFall 2013). In this cognitive test, respondents hear a list of ten words and have to subsequently recall as many words as possible. Each respondent is asked to recall the words twice, once immediately after the words are read ("immediate word recall test") and a second time after conducting another cognitive test ("delayed word recall test"). Under the default procedure, the words are read to respondents by the laptop of the interviewer. However, about 20% of respondents in our sample hear the words from the interviewer instead. This group of respondents exhibits significantly lower test scores both in the immediate and in the delayed word recall test. Furthermore interviewer effects appear to be more severe when the standard procedure is not followed, as the share of the variance of test scores that can be attributed to interviewers (interviewer intra-class correlation, ICC) is elevated among respondents who hear the words from the interviewer.

We aim to answer three questions. We begin by asking which determinants drive imperfect compliance with the study protocol. Here, we are interested in respondents' characteristics predicting deviation from the default mode as well as heterogeneity across interviewers. Next, we turn to the question of what drives the difference in performance between the two modes. Respondents were not randomized to modes in our setting. Therefore differences in performance may either constitute mode effects, i.e. the test score a particular respondent achieves depends on the mode of administration, or they may stem from selection effects, that is cognitive ability is not evenly distributed across the two groups of respondents. We exploit the existence of test scores from additional cognitive tests that were administered to all respondents in the same mode, to understand whether the two groups of respondents differ in cognitive ability. Finally, we seek to answer whether administration via the computer successfully reduces interviewer effects in our setting. Different interviewers deviate from the default procedure for different respondents. This mechanism can contribute to a disparity in the interviewer intra-class correlation between the two groups, similar to non-response error variance (Brunton-Smith et al. 2012; West and Olson 2010).

Interviewers vary in the time they need to read the ten words and they also tend to repeat words if asked to (Herzog and Rodgers 1999). The laptop on the other hand, always plays the words in the same intonation and speed - regardless of the specific interviewer and respondent. However, faced with a strict protocol, leaving little discretion for catering to individual needs, interviewers will under certain circumstances choose not to - or not be able to - fully comply with the default procedure. We find that hearing the words from the interviewer is predicted by respondent characteristics which have been associated with the propensity to use technologies and cognitive ability. This is consistent with the interpretation that respondents with low previous exposure to computers may not feel comfortable with administration via the laptop and are therefore more likely to hear the words from the interviewer. The share of respondents who hear the words from the interviewer varies greatly across interviewers. This may reflect differences in preferences across interviewers, while differences in laptop quality are also likely to contribute to this pattern.

Selection effects appear to be the main driver of performance differences in our setting. Those respondents who are read the words to by the interviewer, perform significantly worse also in other cognitive tests. In the delayed word recall test, however, the difference between the two groups is larger than in any other cognitive test. Therefore we cannot fully rule out the presence of mode effects.

Finally, the differences in interviewer intra-class correlations between the two modes are greater in the word recall test than in all other cognitive tests. Therefore we conclude that the use of the laptop does indeed seem to reduce interviewer effects in the word recall test. Our findings suggest that the use of laptops in administration of the word recall test is preferable, despite the problem of non-compliance.

The study most closely related to ours is Al Baghal (2017) who find that respondents obtain higher test scores when a cognitive test is administered in the web mode rather than in the CAPI-mode. In their setting, some respondents were randomized to the CAPI-mode, while others were initially randomized to the web-mode but could self-select into the CAPI-mode. Although self-selection explains some of the differences in outcomes across modes, mode effects also appear to be present.

Our work is also related to literature studying the effect of fully controlled variation in the mode of administration on test scores in cognitive tests. Herzog and Rodgers (1999) compare test scores of old individuals randomly assigned to either a telephone interview or to a face to face interview and do not detect any significant differences across the two modes. Similarly, Gooch (2015) randomize participants to either self-administration or a face-to-face interview. While they detect differences in the marginal distributions of single items across the two modes, the results obtained by item-response-theory models do not suggest variation in cognitive ability across the two modes.

Assessments of cognitive functioning are increasingly administered within longitudinal studies such as the SOEP in Germany (Wagner et al. 2007), SHARE in Europe (Börsch-Supan et al. 2013) or HRS in the United States (Sonnega et al. 2014). These measurements help researchers to understand the impact of childhood schooling (Glymour et al. 2008) or retirement (Celidoni et al. 2017; Rohwedder and Willis 2010) on cognitive decline at old ages as well as the relationship between cognitive decline and aging in general (McArdle et al. 2007; Whitley et al. 2016). Cognitive ability is also an important determinant of economic decision making. For example Smith et al. (2010) link cognitive ability to financial numeracy. Furthermore, economists have studied the role of cognitive ability for labor market and behavioral outcomes (Heckman et al. 2006) and the returns of cognitive abilities in Germany (Heineck and Anger 2010).

Our results yield useful insights for survey administrators, as we suggest that the use of technical devices can reduce interviewer effects without introducing substantial mode effects. We also remind applied researchers to carefully examine the cognitive test data they are working with and approach potential biases with suitable methods. The remainder of this chapter proceeds as follows. Section 3.2 introduces the data and the sample. Section 3.3 analyzes selection and performance differences in the word recall tests. In Section 3.4, we discuss the implications of our findings and Section 3.5 concludes.

3.2 Data, variables and descriptive statistics

In this section we first describe our data and sample. We proceed with an overview of cognitive tests that were administered in the third wave of Understanding Society, the UK Household Longitudinal Study.

3.2.1 Data and sample

Our data come from Understanding Society - the UK household longitudinal study (UKHLS). Understanding Society follows around 40,000 households in England, Northern Ireland, Scotland and WalesUniversity of Essex (2016). These households were mostly sampled prior to the first wave. UKHLS employs a stratified clustered sample design, that is rather than selecting households from the whole population, primary sampling units (PSUs) - restricted geographic areas mostly overlapping with postcode sectors - are drawn in a first step and households are only selected within those PSUs (Boreham et al. 2012). Each household member above the age of 16 is interviewed on a yearly basis, with the questionnaires covering a wide range of topics such as health, work, education, income, family, and social life. We limit the analysis to the third wave, which includes a module to assess cognitive functioning. Data collection for the third wave took place between January 2011 and July 2013. The vast majority of interviews was conducted in the computer-assisted personal interview (CAPI) mode.[1] Interviewers attended a one day briefing prior to conducting the fieldwork (Knies 2016). We restrict our sample to individuals having completed a face-to-face interview who state that English is their first or childhood language[2] and have a valid test score in the word recall test.[3] Furthermore, we require non-missing information on basic individual as well as household characteristics.[4] In our analysis we are particularly interested in the role of the interviewer. As one interviewer typically conducts all interviews within a household, common unobserved characteristics of household members can increase correlation of responses within interviewers. Any analysis neglecting this source of clustering will potentially overstate the role of the interviewer. Therefore, we base our main analysis on a household sample that includes only one member from each household.

In total, this sample includes 24,323 observations in 6277 primary sample units. Each of the 686 interviewers conducted between 1 and 154 in 1 to 80 primary sampling units

[1] 481 interviews were conducted via telephone. We exclude these observations from the sample. Additionally respondents were asked to answer a self-completion questionnaire in the CASI mode.

[2] Previous literature has highlighted that some English-language cognitive tests might not provide valid measurements of bilingual individuals cognitive ability (Sanchez et al. 2010).

[3] For information on first or childhood language we draw on wave 1, 2 and 5.

[4] We exclude all observations, where one of the following variables is missing: age, sex, marital status, education, household income, born in UK, hearing problems, urban or rural area, longstanding illnes or disability.

Table 3.1: *Sample structure: Understanding Society (Wave 3)*

Observations	Number of PSUs	Number of Interviewers	PSU-Interviewer combinations
24323	686	6180	8505
Number of observations by interviewer:			
Mean	Median	Min	Max
35.46	29	1	154
Number of PSUs each interviewer is connected to:			
Mean	Median	Min	Max
12.4	10	1	80

(see Table 3.1). Restricting the analysis to one respondent from each household leads to overrepresentation of respondents from single and small households. On average, individuals in the household sample are about one year older than the average of the full sample, they are more likely to be female and only 47% are married compared to 54% in the full sample. In order to make the analysis representative for the full sample, we use the number of eligible observations in the household as sampling weights. Once we adjust for the probability of being selected for the household sample conditional on being in the full sample, differences become minor and mostly insignificant (see Table 3.2).

Table 3.2: *Sample characteristics: Understanding Society (Wave 3)*

	Household sample					Full sample			Diff
	Unweighted			Weighted		Unweighted			Weighted-Full
	Mean	SD	N	Mean	SD	Mean	SD	N	p-value
Male	0.42	0.49	24323	0.44	0.5	0.43	0.5	36591	0.25
Age	50.59	17.96	24323	49.43	17.88	49.62	17.83	36591	0.25
Single	0.28	0.45	24323	0.27	0.45	0.27	0.44	36591	0.76
Married	0.47	0.5	24323	0.54	0.5	0.54	0.5	36591	0.58
Educational attainment: Degree	0.34	0.47	24323	0.34	0.47	0.34	0.47	36591	0.98
Educational attainment: A-levels	0.19	0.4	24323	0.21	0.4	0.2	0.4	36591	0.73
Educational attainment: GCSE	0.2	0.4	24323	0.21	0.41	0.21	0.41	36591	0.75
Born in UK	0.93	0.25	24323	0.94	0.24	0.94	0.24	36591	0.51
Household size	2.5	1.36	24323	2.75	1.37	2.73	1.37	36591	0.06
Household income (log)	6.9	0.64	24323	6.89	0.62	6.89	0.62	36591	0.39
Urban area	0.75	0.43	24323	0.74	0.44	0.74	0.44	36591	0.46
Longstanding illness or disability	0.39	0.49	24323	0.37	0.48	0.37	0.48	36591	0.84
Hearing problems	0.03	0.18	24323	0.03	0.17	0.03	0.18	36591	0.61

Notes: Descriptive statistics. Household sample is the sample used for analysis, which includes one member per household. In the weighted household sample, each observation is weighted by the number of interviews eligible for the sample in its household. Full sample denotes the sample with all household members after applying the sample restrictions. P-values from Welch's two sample t-test to test equality of means between the weighted household sample and the full sample.

3.2.2 Cognitive tests in Understanding Society

Cognitive ability of respondents was assessed in the third wave of Understanding Society. McFall (2013) provide detailed background information on items and the testing procedure which we will summarize in this section. The cognitive functioning module relies on cognitive tests that are relatively brief and have been used in previous surveys. Interviewers received special instructions for administering the cognitive tests. Specifically they were advised to keep disturbances (including the presence of a third person) at a minimum level and they were discouraged from giving feedback on the performance. Additionally interviewers were reminded to administer the test exactly as specified in the instructions. The following cognitive tests are included in the module:

- **Word recall:** In the word recall test, respondents hear a list of ten words. Subsequently they are asked to recall the words both immediately after the list was read and again after conducting another cognitive test. The test score is the sum of correctly named words. In our data respondents were randomized into four different lists of words (see Figure C.1 in the Appendix). The word recall tests provides a measure of episodic memory, an ability that declines with age. This test is among the more difficult tests typically administrated within surveys (Herzog and Rodgers 1999).

- **Serial 7 subtraction:** This task requires the respondent to subtract 7 five times, starting at 100. The test score is the number of correct subtractions. This test assesses the ability to process, dispose and retrieve information on short term as well working memory.

- **Number series:** In the number series, respondents are presented with a series of numbers that includes a blank and are asked to fill in the blank. An adoptive testing procedure was applied, that is all respondent start with some baseline items and are assigned to subsequent items based on performance. The test score gives the imputed number of correct items had the respondent been asked to try all items.[5] The number series tests fluid reasoning.

[5] Respondents were randomized into two sets of task. Prior to testing respondents were given examples of the problems. In case the respondent did not seem to understand the task, interviewers were

Table 3.3: *Cognitive tests: Understanding Society (Wave 3)*

	Mean	SD	Min	Max	N	Missing in full sample
Immediate word recall	6.24	1.73	0	10	24323	0.01
Delayed word recall	5.2	2.1	0	10	24323	0.003
Immediate word recall (interviewer read)	5.83	1.9	0	10	4796	0.811
Immediate word recall (computer read)	6.34	1.67	0	10	19527	0.199
Delayed word recall (interviewer read)	4.54	2.24	0	10	4796	0.812
Delayed word recall (computer read)	5.36	2.04	0	10	19527	0.201
Serial 7 subtraction	4.43	1.07	0	5	23662	0.037
Number series: Set 1	530.18	31.9	409	584	22807	0.543
Number series: Set 2	529.8	33.65	413	584	11680	0.526
Verbal fluency: correct	21.9	6.87	0	58	24238	0.012
Numerical ability	3.59	1.11	0	5	24175	0.015

Descriptive statistics for cognitive tests. Columns (1)-(5) refer to the weighted household sample. The last columns gives the share of missing test scores for the full sample after restricting the sample to respondents with a face-to face interview and English as a first language, but prior to restricting the sample to respondents with a valid score in the word recall test.

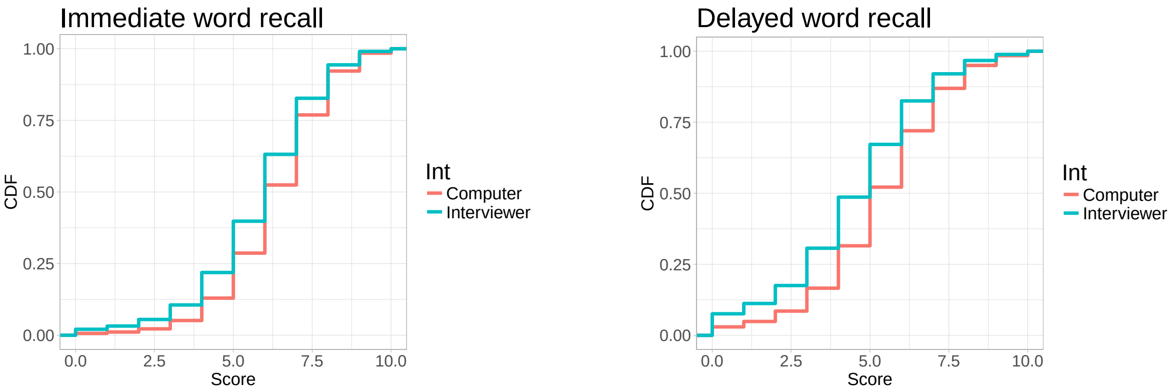
- **Verbal fluency:** Verbal fluency is assessed by having respondents name as many animals as possible within a minute. The test score is the number of distinct, correctly named animals. This test is related to executive functioning as well as mental flexibility.

- **Numeric ability:** Respondents are asked to solve 4-5 mathematical problems likely to come up in daily life. The test score is the number of correctly solved problems. This test assesses practical numerical knowledge.

In this study we restrict our analysis to respondents with a valid test score in the word recall test, neglecting selective non-response as a potential source of bias (Herzog and Rodgers 1999). In the sixth column of Table 3.3 we display the share of observations with a missing test score for the full sample after restricting the sample to respondents with a face-to face interview and English as a first language, but prior to restricting the sample to respondents with a valid score in the word recall test. In the world recall test, less than 1% of observations are missing. Also, except for the number series, where about 7% percent of respondents lack a test score in both sets, the share of missing test scores is quite low in the other cognitive tests. Consequently, item non-response is unlikely to be a decisive factor for this analysis. Nevertheless, we will explore the sensitivity of our results with respect to different ways of dealing with missing outcomes.

instructed to proceed to the next question. Due to a mistake in the CAPI code, some respondents randomized to the second set were given tasks they should have skipped. Therefore the test scores of both items are given in separate variables.

Figure 3.1: *Empirical distribution of test scores*



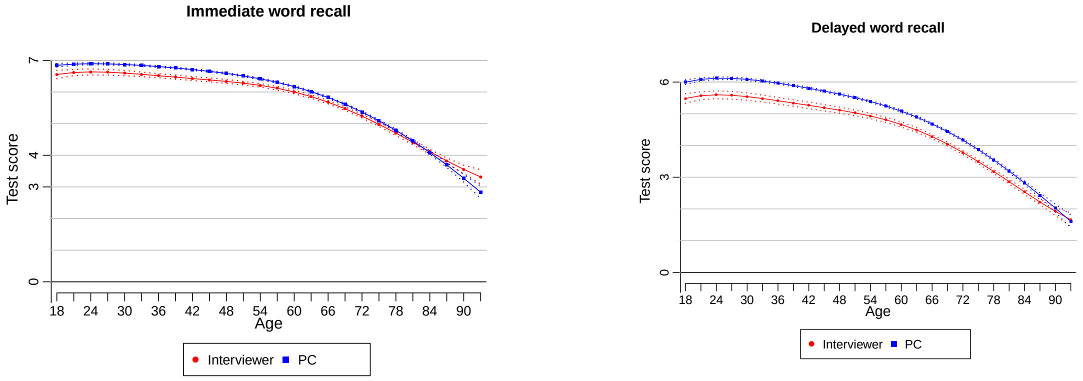
Notes: Empirical distribution of test scores for immediate and delayed word recall test by mode of test. Weighted household sample.

3.3 Word recall test

In the following we focus on the word recall test. By default, the list of words should be read to the respondents by the computer of the interviewer. If, however, the respondent is unable to hear the computer, interviewer instructions require interviewers to read the list of words to the respondents themselves. While about 80% of respondents are administered the word recall test under the default mode, for the remaining 20% of respondents interviewers read the words themselves (see Table 3.3). Overall, those individuals who hear the words from the interviewer perform worse both in the immediate as well as in the delayed recall task. In the delayed recall task this is partly driven by excess zeros in the group of respondents hearing the words from the interviewer. Furthermore, the difference between the modes is stronger in the delayed recall task than in the immediate recall task (see Figure 3.1). Additionally, in the delayed recall test respondents of all but the oldest age groups achieve lower scores when hearing the words from the interviewer. In contrast, in the immediate recall test only respondents below 70 perform worse under deviation from the study protocol (see Figure 3.2). While the interviewer intra-class correlations are not larger than for other cognitive outcomes under the default procedure, they are inflated when the interviewer reads the question herself (see Figure 3.3).

The rationale behind administering the word recall test with the help of a laptop is to shut down the channel of heterogeneity in the interviewers' reading styles as a source of variation in observed test scores. However, as not all respondents hear the word from the laptop, the word recall test is effectively administered in a mixed mode design. In

Figure 3.2: *Performance by age*



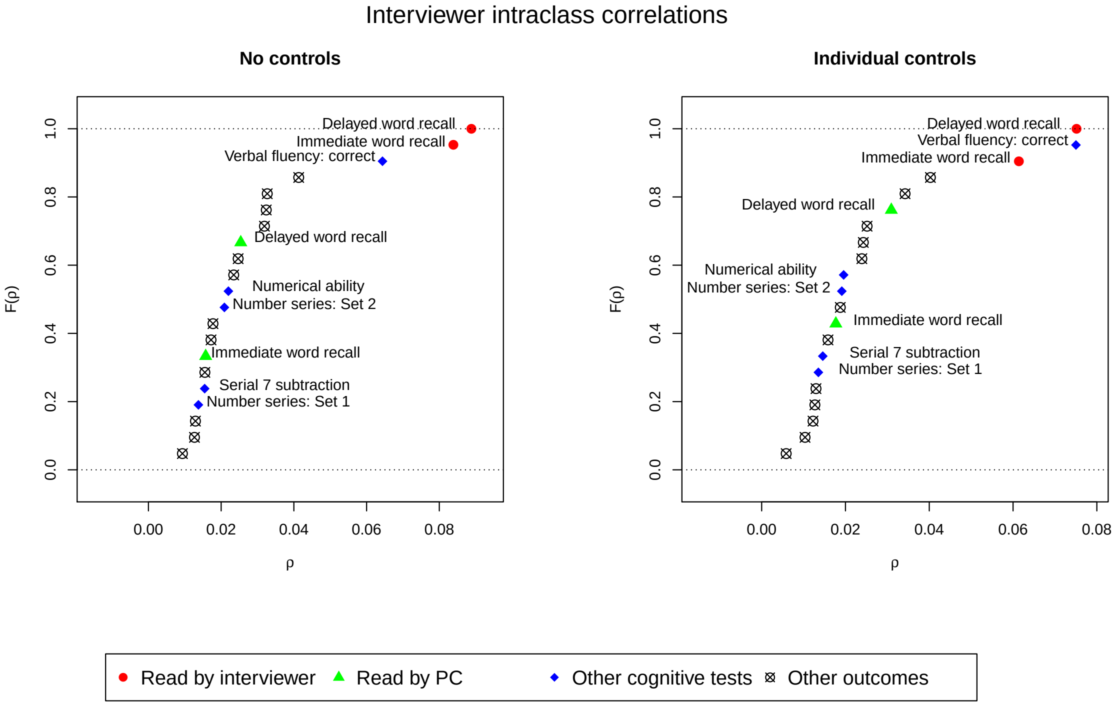
Notes: Performance in the word recall test by age and mode of administration.

this section, we explore determinants of imperfect compliance with the default procedure. We further investigate to which extend observed performance differences can be linked to mode effects and selection effects. Lastly, we turn to the question whether administration of the word recall test using the computer successfully reduces interviewer effects in our setting.

3.3.1 Determinants of imperfect compliance

We begin by asking why and under which circumstances interviewers read the words in the word recall test themselves. Broadly speaking, we can group sources of non-compliance with the default procedure in the administration of the word recall test into three categories: Firstly, some of the laptops distributed to interviewers may have problems with their audio function. In these cases, deviations from the standard procedure are a consequence of technical failure and their occurrence is partly random. The second category comprises factors related to the interviewer. Preferences for the mode of the recall test are likely to vary across interviewers. While some interviewers may have a high intrinsic motivation to follow the study protocol, others might not like hearing the computers' voice or they perceive reading the words themselves as a good way to help respondents in performing the task. Thirdly, respondents' characteristics constitute an additional potential determinant of deviations from the study protocol. Older individuals, less educated individuals and minorities have been found to report less use of technology (Czaja et al. 2006). And respondents with a low level of previous exposure to technology may feel intimidated by the prospect of interacting with a computer

***Figure 3.3:** CDF of interviewer effects*



Notes: Empirical CDF of interviewer intraclass correlations. Estimated using cross classified random effect models, with one random intercept for interviewers and one random intercept for PSUs. Outcomes:(1) Word recall (read by interviewer), (2) Word recall (read by computer), (3) Other cognitive tests (Serial 7 subtraction, Number series: Set 1, Number series: Set 2, Verbal fluency: correct, Numerical ability), (4) Other nonfactual items, that were assessed in the CAPI-mode (Prefers to move house, Frequence of internt use, TV hours, likes present neighbourhood, Standard of local services: Shopping, Worry about being affected by crime, Number of close friends, Feel safe walking alone at night, Supports a particular political party, Level of interest in politics, Perceived political influence, Proportion of friends with similar income,). Controls include: Hearing problems, Male, Marital status (3 levels), Dummies for age in 5 year categories, Linear age trend, Educational attainment (4 levels), Born in UK, Urban area, Household size, Log of monthly net income per household member, Longstanding illness or disability.

(Rosen and Weil 1995). Previous research has also shown that technophobia is a trait which creates discomfort when reading from a screen among senior adults (Hou et al. 2017), and a similar phenomenon might apply to audio tasks. Furthermore, hearing impairments can provide a physical barrier impeding administration via the computer. If a respondent claims to be unable to do the test under the standard procedure, administering the test in an alternative mode is in fact often the only way of obtaining a test score at all.[6]

It is also easy to think about ways how these three mechanisms interact with each other. For example, technophile interviewers will find it easier to deal with minor technical malfunctions and at the same time one might expect them to be unlikely to exhibit a preference for reading the word themselves. Different interviewers may also react to the same behavior of the respondent in heterogeneous ways. While some interviewers will be quick to read the words themselves whenever a respondent expresses discomfort with the computer, others will exert some effort to make administration under the default procedure possible.

Disentangling the roles of technical failure, interviewer-related factors and respondent-related factors in a causal way requires experimental variation. Specifically, we would need a setting where laptops and interviewers are randomly allocated to respondents. In our data this was not the case. Therefore, we must be careful before ascribing correlations in our data to causal channels. For example, a high interviewer intra-class correlation in the propensity to read the words may be caused by heterogeneity in interviewer preferences, but it could as well reflect clustering in observed respondent characteristics or variation in laptop quality.

What we can achieve in this section is to present indirect and descriptive evidence suggesting a relationship between respondents' characteristics and the mode of the test as well as heterogeneity across interviewers. While some of our findings are consistent with the interpretation of technical failure as additional source of non-compliance, we do not think this is the sole driver of imperfect compliance with the study protocol.

[6] This is vastly related to the question whether conversational or standardized hearing techniques produce smaller measurement errors. Schober and Conrad (1997) show that standardized procedures yield a greater measurement error when respondents are unsure about how a question maps onto their circumstances. While regular employees answer questions on their weekly working time easily, self-employed individuals often require clarification. In our setting the standardized procedure may leave little room to cater individual needs, inducing them to switch to reading the words themselves.

In the following, we assume the existence of a latent variable Int^*_{ij}, which can be interpreted as the latent propensity of interviewer j to read the words for respondent i. Int^*_{ij} is a linear function of observed characteristics and an error term.

$$\texttt{Int}^*_{ij}{=}Z_i\delta + w_{ij}, \tag{3.1}$$

The latent propensity to read the words maps into the observed mode of administration as follows:

$$\texttt{Int}_{ij} = \begin{cases} 0 & \text{if } Int^*_{ij} < 0 \\ 1 & \text{if } Int^*_{ij} \geq 0 \end{cases} \tag{3.2}$$

In our empirical specification Z_i includes age in five-year dummies, sex, marital status, educational attainment, hearing problems, longstanding illness or disability, a dummy for living in an urban region, number of household members, the log of monthly household income per household member, as well as a linear age trend.

We begin by estimating Equation 3.1 using pooled OLS and pooled logit models. We argue that interviewers play an important role in in process of assigning respondents to the two modes of administration. In a next step, we therefore model the error term to include an interviewer-specific component, i.e $w_{ij} = c_j + \omega_{ij}$. Since the assignment of interviewers to respondents was non-experimental, part of the estimated error variance attributed to the interviewer might in truth be caused by clustering of unobserved characteristics within interviewers. In fact, interviewers usually operate in spatially restricted areas, and at the same time inhabitants of one geographic area tend to share common characteristics (Schnell and Kreuter 2005). In order to mitigate this problem, we incorporate a primary sampling unit (PSU) specific component into the error term, that is we let $w_{ij} = c_j + \omega_{ij} + p_{i \in p}$.[7]

We estimate specifications including interviewer or PSU effects using linear probability fixed-effects and random effects models. In order for the random effect estimator to yield consistent estimates of the β coefficients, interviewer and PSU effects must be uncorrelated with the regressors. This would be violated if for example interviewers with a high propensity to read the question themselves were more likely to interview old re-

[7] Conceptually sampling point effects can be seen as averages of unobserved characteristics within a primary sampling unit.

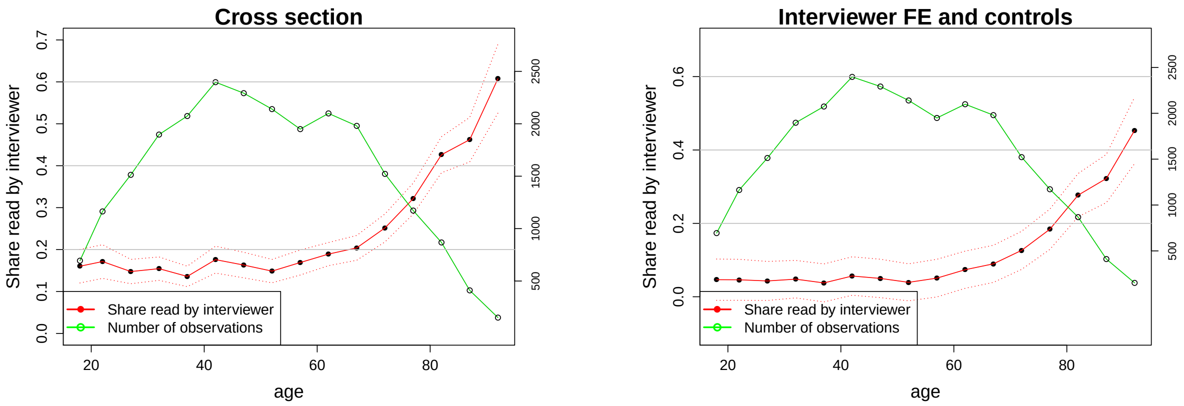
spondents. The fixed effects estimator on the other hand allows for correlation between unobserved effects and regressors but is less efficient. Furthermore interviewers and PSUs in our data are sparsely matched as each interviewer only visits a limited number of PSUs and challenges arise in the case of two-way effects estimation with sparsely matched data (Verdier 2017). Finally, we are also interested in the importance of interviewer effects compared to household effects. Therefore we also estimate a random effects model that includes a random intercept at the household level additional to the random intercept of the interviewer. The main results are presented in Table 3.4.

Respondents' characteristics

All specifications suggest a similar conclusion. Age, hearing problems, education and - to a lesser extent - marital status are significantly associated with the mode of administration (see Table 3.4). Previous literature has documented a relationship between some of these traits and the propensity to use technologies (Czaja et al. 2006). Our findings are therefore consistent with the idea of respondents with a low level of exposure to computers expressing discomfort with the default procedure.

Hearing problems are the strongest predictor for hearing the words from the interviewer. Reporting hearing problems in the health module increases the propensity to be administered the test in the non-default mode by 15 to 19 percentage points. Respondents who have obtained an university degree, A-levels or GCSE are all significantly less likely to hear the words from the interviewer than the base category of respondents without further education. This effect is monotonic in the level of education, with respondents holding a degree being the least likely group to deviate from the default mode. Moreover, single individuals are significantly more likely to have the words read by the interviewer than the base category of widowed or divorced individuals, while married individuals are not significantly different from the base category. Estimated coefficients from random effect models and fixed effect models are quantitatively very similar. This implies that correlation between interviewer effects and regressors does not strongly affect the estimates from the random effect models. Furthermore, average marginal effects reported for the logit model are quantitatively very close to the estimated coefficients from the OLS-model. Therefore ignoring the binary nature of the outcome variable should not shift conclusions drastically.

Figure 3.4: *Mode of administration by age*



Notes: Share of respondents for whom the words are read by the interviewer and number of observations in 5-year age categories by age. The left panel displays the raw shares, while the right panel adjusts for interviewer effects and the same controls as shown in Table 3.4. Weighted household sample.

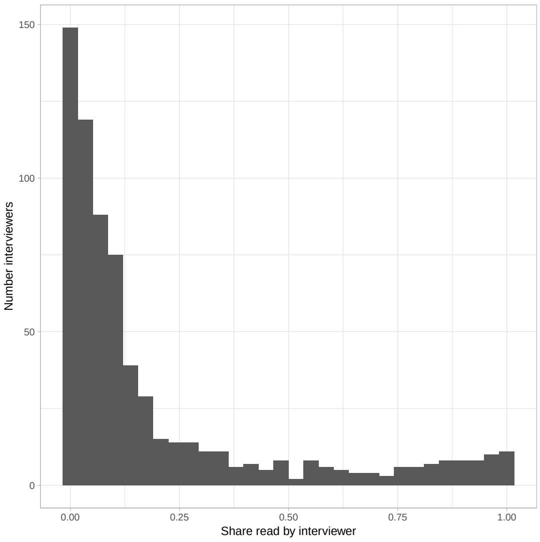
The share of individuals for whom the words are read by the interviewer for each age category is displayed in Figure 3.4. The left panel depicts the raw shares, while the right panel depicts shares after adjusting for interviewer effects and respondents' characteristics. Above the age of sixty, the likelihood of hearing the words from the interviewer increases sharply. However, the test is also administrated in the alternative mode a for about 15% of younger individuals. As the majority of respondents is between the ages 30 to 60, individuals below the age of 60 account for a significant share of respondents in the alternative mode and the mean age in this group is 56 years.

Table 3.4: *Word recall test: Selection into mode of administration*

	Pooled		RE		FE		RE (Full sample)
	OLS	Logit	Int	Int+PSU	Int	Int+PSU	Int+HH
Male	-0.0016	-0.0014	-0.0021	-0.0013	-0.0022	-0.0023	0.0035
	(0.0054)	(0.0055)	(0.0038)	(0.0036)	(0.004)	(0.0046)	(0.0024)
Single	0.0167*	0.0178**	0.0154**	0.0142**	0.0151**	0.0135*	0.0091*
	(0.0085)	(0.0089)	(0.0069)	(0.0065)	(0.0066)	(0.008)	(0.0055)
Married	-0.0026	-0.0031	0.0029	0.0027	0.0028	0.0037	-0.0065
	(0.0078)	(0.0075)	(0.0055)	(0.0051)	(0.0062)	(0.0068)	(0.0046)
Hearing problems	0.1629***	0.1456***	0.1818***	0.1853***	0.182***	0.1866***	0.1873***
	(0.0199)	(0.0188)	(0.0111)	(0.0104)	(0.0173)	(0.0213)	(0.008)
Educational attainment: Degree	-0.0353***	-0.0327***	-0.029***	-0.0267***	-0.0288***	-0.0243***	-0.0181***
	(0.0096)	(0.0088)	(0.0056)	(0.0053)	(0.0064)	(0.0071)	(0.0043)
Educational attainment: A-levels	-0.0341***	-0.0308***	-0.0244***	-0.0209***	-0.0238***	-0.0159**	-0.0192***
	(0.0091)	(0.0082)	(0.0061)	(0.0057)	(0.0069)	(0.0078)	(0.0045)
Educational attainment: GCSE	-0.0263***	-0.0231***	-0.0238***	-0.0184***	-0.0239***	-0.0133*	-0.0167***
	(0.0085)	(0.0077)	(0.006)	(0.0056)	(0.007)	(0.008)	(0.0044)
Born in UK	-0.0085	-0.0083	-0.0117	-0.0093	-0.0116	-0.0114	-0.0113*
	(0.0154)	(0.016)	(0.0081)	(0.0078)	(0.0086)	(0.0112)	(0.0062)
Urban area	-0.0274	-0.0275	0.0049	0.0063	0.0062	-0.0014	0.0031
	(0.0179)	(0.0179)	(0.005)	(0.0059)	(0.008)	(0.0102)	(0.0048)
Household size	0.0042	0.0046	0.0033*	0.0024	0.0034	7e-04	0.0021
	(0.0033)	(0.0035)	(0.0018)	(0.0017)	(0.0023)	(0.0021)	(0.0017)
Log of monthly net income per household member	-4e-04	-6e-04	0.0022	0.0028	0.0023	0.0026	-0.0019
	(0.0051)	(0.0053)	(0.0034)	(0.0032)	(0.0031)	(0.0036)	(0.0031)
N	24323	24323	24323	24323	24323	24323	36591
σ^2_{int}			0.0704	0.0658	0.0702	0.085	0.0689
σ^2_{psu}				0.022		0.0285	
σ^2_{hh}							0.0459
ρ_{int}			0.3568	0.354	0.3567	0.3539	0.4517

Notes: Dependent variable: Dummy for words read by interviewer. Column (1): LPM, standard errors clustered on interviewer level. Column (2): Logit model, average marginal effects, Standard errors clustered on interviewer levels. Column (3) and (4): Linear Mixed-Effects-Model with random intercepts for interviewers and (only 4) PSUs (Bates et al. 2015), Column (5) and (6): Linear Fixed-Effects-Model with interviewers and (only 6) PSU fixed effects, Column (7): Linear Mixed-Effects-Model with random intercepts for interviewers and households on the full sample. All models additionally include dummies for age in 5-year categories and a linear age trend.

Figure 3.5: *Share of deviations by interviewer*



Notes: Share of interviews with deviation from default mode by interviewer.

Interviewer effects

The share of the two modes of the word recall test varies across interviewers. While about 150 of 686 interviewers never read the words themselves and more than half read in less than 10% of their interviews, 10 % of interviewers, deviate in at least 75% of their interviews from the default procedure (see Figure 3.5). Therefore we are interested in assessing which share of the variance in assignment to test modes may be attributed to the interviewer.

The lower panel of Table 3.4 presents the estimated variances of interviewer, PSU and household effects as well as interviewer intra-class correlations defined as $\frac{\sigma^2_{int}}{\sigma^2_{int}+\sigma^2_{psu}+\sigma^2_{\omega}}$. In case of the fixed-effect models, we obtain estimators for the variances of the unobserved effects by first calculating the residuals as $\hat{w_{ij}} = \texttt{intread}_{ij} - Z_i\hat{\delta}^{FE}$ and then estimating the variance of the unobserved effects using random effect models.[8] Regardless of the specification, the estimated interviewer intra-class correlation in the household sample is always within the range of 0.35 to 0.36. Including PSU-effects in

[8] In principle it is also possible to estimate these variances directly as variances of the estimated fixed effects but then a bias correction to remove estimation error needs to be applied (see for example Andrews et al. 2008). In the specification that includes both interviewer fixed effects and PSU fixed effects, further complications arise. In our data, PSUs and interviewers form several connected networks and one interviewer effect in each network is not identified (Abowd et al. 1999). Therefore we cannot obtain an estimate for variance between these connected networks. Also within the largest connectivity group, which comprises the majority of the sample, interviewers and sampling points in our data form a network with low connectivity. The resulting variance estimators will have a high bias and a high variance (Jochmans and Weidner 2017).

addition to interviewer effects reduces the estimated interviewer intra-class correlation only marginally. When using the full sample and including a random intercept for the household, the share of the residual variation attributed to the interviewer is even larger ($\rho_{int} = 0.45$). These findings suggest that whether a respondent hears the words from the interviewer or the computer is to a large extend determined by which interviewer she is interviewed by.

Technical failure

Unfortunately our data do not include the information whether an interviewer reported a problem with her laptop. Above we have seen that the mode of administration is associated with certain characteristics of respondents. We have no reason to believe that interviewers with malfunctioning laptops are sent to older or less educated respondents more often. Therefore we think it is unlikely that technical failure accounts for the majority of deviations from the study protocol. Nevertheless, in this section we present some evidence consistent with the idea of technical failure being an additional determinant of non-compliance with the study protocol. We can, however, not formally identify this channel.

While it is impossible for us to know whether a laptop's audio function failed during a specific interview, we argue that technical failure should create a certain pattern in the sequence of modes. If an issue with a laptop's audio function occurs, this issue will likely be present for several consecutive interviews until the problem is fixed. In order to check whether deviations from the study protocol occur sequentially, we regress $\texttt{int}_{ij}$ on a dummy variable which is equal to one, if the last word recall test in a previous household interviewed by the same interviewer was administered in the alternative mode.[9]

The results are shown in Table 3.5. In the cross-section respondents are about 55 percentage points more likely to hear the words from the interviewer when the words were also read for the last respondent in the previous household. This effect includes heterogeneity in the propensity to read the words across interviewers. Once we include interviewer fixed effects, the estimated effect reduces to 25 percentage points and is

[9] The last respondent in the previous household is not necessarily included in the household sample. We omit respondents living in the first household visited by each interviewer.

Table 3.5: *Words read in previous household*

	OLS		FE	
	(1)	(2)	(3)	(2)
Read for last respondent in previous household	0.5534***	0.5506***	0.2465***	0.2432***
	(0.0235)	(0.0239)	(0.0211)	(0.0213)
N	21283	21283	21283	21283
Controls	No	Yes	No	Yes
Interviewer FE	No	No	Yes	Yes

Notes: Dependent variable: Words read by interviewer. This table displays the estimated coefficients of a dummy variable, which is equal to one if in the most recent word recall test in a previous household the words were read by the interviewer. Weighted household sample. We omit the first household visited by each interviewer.

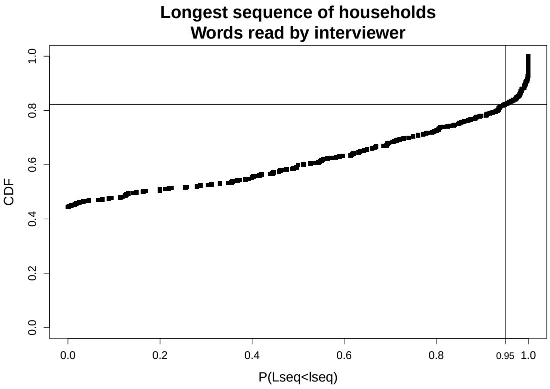
still highly significant. Interestingly, controlling for individual and household characteristics hardly affects the estimated coefficients. This suggests that, the observed pattern in the sequence of mode cannot be explained by households that are similar in terms of observed characteristics being visited consecutively.[10] Another way to check for patterns consistent with technical failure is to look at the longest sequence of households where all households members hear the words from the interviewers. For each interviewer, we calculate the probability of observing a shorter longest sequence than the one observed, if the sequence was created randomly given the number of interviews in both modes.[11] The results are displayed in Figure 3.6. The x-axis shows the probability of observing a longest sequence shorter than the one we observe, the y-axis gives the CDF of these probabilities. About 44% of interviewers never read the words for all household members in two consecutive households (x-axis=0). On the other hand, for slightly below 20% of interviewers, we would observe a shorter longest sequence with at least 95% probability if the sequences were randomly allocated. This suggests that some interviewers do not randomize their sequence of modes and one likely explanation for this pattern is technical failure.

We also investigate heterogeneity in modes by timing of the interview. We normalize

[10] In theory it is still possible that households that are similar in terms of unobserved characteristics are visited sequentially. However it is not clear why these unobserved characteristics should be uncorrelated to observed characteristics.

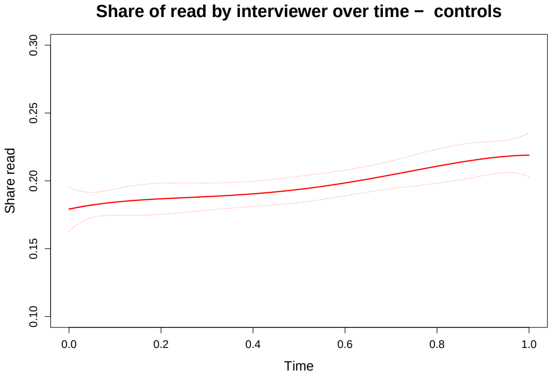
[11] We exclude interviewers without variation in the mode and all interviewers visiting less than three households. Consider an interviewer who interviews 10 households, reads the words in 4 interviews, of which 3 occur subsequently. There are 210 ways of allocating 4 times reading the words to 10 interviews. In 161 of these possible allocations, the length of the longest sequence of interviews in the non-default mode is less than 3. Therefore the probability of observing a shorter-longest sequence than the one observed is 0.77. For an explanation on how to calculate these probabilities see Schilling (1990).

***Figure 3.6:** Probability of shorter sequence*



Notes: The x-axis gives the probability of observing a shorter longest sequence in households with all household members being interviewed in the alternative mode for each interviewer. The y-axis shows the empirical CDF of this probability.

***Figure 3.7:** Share of deviations by time*



Notes: Share of interviews with deviation from default mode by time of interview. The first interview date for each interviewer is normalized to 0 and the last date is normalized to 1. The graph shows the predicted share of interviews with words read by the interviewer when all individual an household characteristics as shown in Table 3.4 are set to their respective means.

the date of the interview across interviewers by setting the date of the first interview to zero for each interviewer and the date of the last interview to one. The share of words read over the observation period is depicted in Figure 3.7, where we control for observable characteristics. There is a small upwards sloping trend throughout the observation period. While this trend is consistent with the interpretation of computers becoming worse over time, it could also be generated by interviewers developing a preference for a certain mode during the period of administration.

3.3.2 Performance differences

Respondents who hear the words from the computer remember on average 0.82 words less in the delayed recall test than respondents for whom the laptop reads the words. This raises the question whether promoting administration of the word using the computer significantly reduces measurement error after all. While the computer removes measurement error caused by heterogeneity in interviewers' ability to read the words, it may introduce mode effects as new source of undesired variation.

Assignment of respondents to modes of administration is clearly non-random. Therefore we cannot readily interpret differences in average test scores as causal mode effects. In this section, we seek to determine whether selection effects may account for the differences in average performance between the two groups. Assuming that all cognitive tests measure a related form of cognitive ability, we will contrast performance differences between the two groups in the word recall test and in other cognitive tests.

Framework

Let ξ_i denote the latent, true cognitive functioning of respondent i. As ξ_i is a latent construct, it does not possess any natural scale and we can normalize the expectation of ξ_i to zero. ξ_i may be decomposed into a part explained by observable characteristics as well as unobserved individual heterogeneity, i.e:

$$\xi_i = X_i\beta + \nu_i \tag{3.3}$$

We formulate the relationship between latent cognitive functioning and observed test scores in the language of a common-factor model (Alwin 2007; Cernat et al. 2016).

$$Y_{ijg} = \tau_g + \kappa_g\xi_i + e_{ijg} \tag{3.4}$$

Here j indexes interviewers and g indexes measurements. We refer to both, different modes of one cognitive test as well as different cognitive tests, as distinct measurements. Two parameters govern the structural relationship between latent cognitive ability and observed test scores. The slope parameter κ_g determines the strength of the association and - following from the normalization of ξ_i - the intercept τ_g coincides with expected test score in the population. As in the previous section, we model the error term e_{ijg} to be the sum of idiosyncratic error as well as an interviewer effect, i.e $e_{ijg} = u_{jg} + \epsilon_{ijg}$. [12] In the following, the variable $\texttt{Int}_{ij}$ is defined as in Section 3.3.1. $\texttt{Int}_{ij}$ categorizes respondents into two groups, depending on the mode of the word recall test. Furthermore, let $Y_{ijg'}$ denote a test score in the word recall test when the computer reads the words, let $Y_{ijg''}$ denote a test score in the word recall test when the interviewer reads

[12] We assume the allocation of interviewers to respondents to be unrelated to true cognitive functioning and other sources of error, such that $cov(u_{jg}, \xi_i) = 0$ and $cov(u_{jg}, \epsilon_{ig}) = 0$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$ and $g \in \mathcal{G}$. The error term e_{ijg} and latent cognitive functioning ξ_i are uncorrelated by definition as the slope λ_g captures the relationship between cognitive functioning and observed test scores. In reality, interviewers can effect observed responses in additional ways. T. F. Crossley et al. (2017) point out that interviewers might further affect measurement error by moderating respondent error. Also Brunton-Smith et al. (2017) allow for location and scale effects. Furthermore interviewers might also affect the structural relationship between true ability and observed test scores.

Table 3.6: *Predictions for observed test scores under pure mode effects and pure selection effects.*

	Prediction for word recall test	Prediction for other cognitive tests
1. Pure mode effect	$(Y_{ijg'} \mid \texttt{Int}{=}0) > (Y_{ijg''} \mid \texttt{Int}{=}1)$ $(Y_{ijg'} \mid \texttt{Int}{=}0, X) > (Y_{ijg''} \mid \texttt{Int} = 1, X)$	$(Y_{ijg'''} \mid \texttt{Int}{=}0) = (Y_{ijg'''} \mid \texttt{Int}{=}1)$ $(Y_{ijg'''} \mid \texttt{Int}{=}0, X) = (Y_{ijg'''} \mid \texttt{Int}{=}1, X)$
2. Selection on observables	$(Y_{ijg'} \mid \texttt{Int}{=}0) > (Y_{ijg''} \mid \texttt{Int}{=}1)$ $(Y_{ijg'} \mid \texttt{Int}{=}0, X) = (Y_{ijg''} \mid \texttt{Int} = 1, X)$	$(Y_{ijg'''} \mid \texttt{Int}{=}0) > (Y_{ijg'''} \mid \texttt{Int}{=}1)$ $(Y_{ijg'''} \mid \texttt{Int}{=}0, X) = (Y_{ijg'''} \mid \texttt{Int}{=}1, X)$
3a. Selection on unobservables	$(Y_{ijg'} \mid \texttt{Int}{=}0) > (Y_{ijg''} \mid \texttt{Int}{=}1)$ $(Y_{ijg'} \mid \texttt{Int}{=}0, X) > (Y_{ijg''} \mid \texttt{Int} = 1, X)$	$(Y_{ijg'''} \mid \texttt{Int}{=}0) > (Y_{ijg'''} \mid \texttt{Int}{=}1)$ $(Y_{ijg'''} \mid \texttt{Int}{=}0, X) > (Y_{ijg'''} \mid \texttt{Int}{=}1, X)$
3b. Selection and interviewer heterogeneity	$[(Y_{ijg'} \mid \texttt{Int}{=}0) - (Y_{ijg''} \mid \texttt{Int}{=}1) \mid c_j^{high}]$ $< [(Y_{ijg'} \mid \texttt{Int}{=}0) - (Y_{ijg''} \mid \texttt{Int}{=}1) \mid c_j^{low}]$	$[(Y_{ijg'''} \mid \texttt{Int}{=}0) - (Y_{ijg'''} \mid \texttt{Int}{=}1) \mid c_j^{high}]$ $< [(Y_{ijg'''} \mid \texttt{Int}{=}0) - (Y_{ijg'''} \mid \texttt{Int}{=}1) \mid c_j^{low}]$

Notes: Predictions for observed test scores under (1) pure mode effects, (2) selection on observables, (3a) selection on unobservables (3b) interviewer heterogeneity in selection.

the words and let $Y_{ijg'''}$ denote a test score in any other cognitve test. Test scores in the word recall test and test scores of other cognitive tests are assumed to be congeneric measures, that is they are linearly related in their true scores (Alwin 2007, Chapter 3). We only observe $Y_{ijg'}|\texttt{Int} = 0$ and $Y_{ijg''}|\texttt{Int} = 1$, but we observe $Y_{ijg'''}$ irrespective of the realization of $\texttt{Int}_i$. In order to distinguish possible sources behind the performances differences in the word recall test, we consider several scenarios: Pure mode effects, selection on observables, selection on unobservables, and interviewer heterogeneity in selection. Table 3.6 summarizes the predictions regarding observed test scores for each scenario.

Firstly, the fact that we observe $\bar{Y}_{ijg'} > \bar{Y}_{ijg''}$ might be entirely driven by a mode effect. Under this scenario, the expected value of latent cognitive functioning is zero in both groups. Therefore we would not expect to see any difference in average performance between respondents who heard the words from the interviewer and respondents who heard the words from the computer, when looking at the test scores of an additional cognitive test, i.e $\bar{Y}_{ijg'''|Int=0}{=}\bar{Y}_{ijg'''|Int=1}$

Secondly, average performance differences between the group of respondents with $\texttt{Int}{=}1$ and those with $\texttt{Int}{=}0$ might be entirely driven by selection on observable characteristics. For example, in the previous section high age and low education turned out to be strong predictors for hearing the words from the interviewer. Moreover, both variables are associated with cognitive ability (Van Hooren et al. 2007; Verhaeghen and Salthouse 1997). Adjusting for age and education should therefore remove a major part of the differences in observed test scores. Under selection on observables we expect to see a difference in mean test scores of both the word recall test and any other cognitive test. However, once we condition on observed characteristics, these differences should disappear, i.e:

$$E\left(Y_{ij}|X_i, \texttt{Int}\right) = \tau + \kappa X_i\beta + \kappa E\left(\nu_i|\texttt{Int}\right) + E\left(e_{ijg}|\texttt{Int}\right) = \tau + \kappa X_i\beta = E\left(Y_{ij}|X\right) \quad (3.5)$$

Thirdly, performance differences might be caused by selection on both observable and unobservable characteristics. This is the case when unobserved heterogeneity in latent cognitive functioning ν_i is correlated with the error term in Equation 3.1, w_{ij}. For example, we expect technophobe individuals to be more likely to hear the words from the interviewer, while technophobia has been shown to be negatively correlated with cognitive ability (Czaja et al. 2006). This type of selection corresponds to the framework of a Heckman selection model (Heckman 1979). Imposing a standard normal distribution on w_{ij} and assuming that the conditional expectation $E\left(\nu_i|w_{ij}\right)$ is linear (Semykina and Wooldridge 2010), the expected observed test score conditional on $\texttt{Int}_{ij}$ and X_i is given by:

$$E\left(Y_{ij}|\texttt{Int}, X, Z\right) = \tau + \kappa X_i\beta + \kappa\rho_{w\nu}\sigma_\nu\lambda(Z_i\delta), \quad (3.6)$$

where σ_ν denotes the standard deviation of ν_i and $\rho_{w\nu}$ is the correlation between w_{ij} and ν_i. As we expect respondents with higher levels of cognitive functioning to be less likely to hear the words from the interviewer, $\rho_{w\nu}$ will be negative in our setting. $\lambda(.)$ denotes the inverse mills ratio. When $\texttt{Int}_i$ is equal to one, the inverse mills ratio is positive, $\lambda(Z_i\delta)=\frac{\phi(Z_i\delta)}{\Phi(Z_i\delta)}$, and expected observed test scores will be smaller than expected observed test scores in the population. When $\texttt{Int}_i$ is equal to zero, the inverse mills ratio is negative, $\lambda(Z_i\delta)=-\frac{\phi(Z_i\delta)}{1-\Phi(Z_i\delta)}$, and expected observed test scores will be greater than expected observed test scores in the population. Under the scenario of selection on unobservables, we expect to see a difference in mean test scores, both for the word recall test and other cognitive tests. Moreover, these differences should not disappear once we condition on observable characteristics.

In Section 3.3.1 we decomposed the error term of the selection equation, w_{ij}, into an interviewer-specific component, c_j and an idiosyncratic error term, ω_{ij}. Consequently, c_j selection also depends on interviewer characteristics and and we can incorporate c_j into the inverse mills ratio. As we assume the allocation of interviewers to respondents to be unrelated to cognitive ability, c_j and ν_i are uncorrelated. However one needs to consider a potential correlation between the systematic interviewer effect in the error term of observed test scores, u_j, and c_j. Therefore, the expected interviewer effect conditional on $\texttt{Int}_{ij}$ is not necessarily zero. Allowing for heterogeneity on the interviewer level changes the expression for expected test scores:

$$E\left(Y_{ij}|\texttt{Int}, X, Z\right) = \tau + \kappa X_i\beta + \kappa\rho_{\omega\nu}\sigma_\nu E\left(\lambda(Z_i\delta + c_j)|Z\right) + E\left(u_j|\texttt{Int}\right) \tag{3.7}$$

For $\texttt{Int}_{ij}$=1, the inverse mills-ratio is a convex (but approximately linear), monotonically decreasing function of c_j. This implies that the expectation of the inverse mills ratio over c_j is not less than the expected inverse mills ratio when c_j is at its population mean, $E\left(c_j\right)$=0. However, respondents interviewed by interviewers with a high c_j will be overrepresented in the group of respondents for whom the interviewer reads the words, implying $E\left(c_j|\texttt{Int} = 1\right) \geq E\left(c_j\right)$ and therefore interviewer-induced heterogeneity can alleviate the degree of selection. Furthermore, correlation between the propensity to read the words and interviewer effects u_j, can also impact the mean of observed test scores. If u_j and c_j are negatively correlated, implying those interviewers who deviate from the default mode often, cause respondents to have lower test scores, $E\left(u_j|\texttt{Int}_{ij}\text{=}1\right)$ will be less than zero, intensifying selection effects. A positive correlation between u_j and c_j, on the other hand, will partially offset the selection effect or might even reverse it.

Interviewer heterogeneity in selection on cognitive ability will cause the pool of respondents within a specific mode to vary across interviewers. Some interviewers will only read the words for respondents with very low ability. For these interviewers the gap between the two modes will be large. For interviewers who also read to respondents with medium cognitive ability, performance differences between the two groups will be less severe. Assuming that the assignment of interviewers to respondents is unrelated to interviewers' reading ability and preferences, this generates the prediction that performance differences are a decreasing function of an interviewers' propensity to read the words.[13]

[13] A potential threat to this assumption is interviewers' learning. It might be the case that interviewers who are randomly allocated to predominantly old individuals develop better reading skills.

Results

In the following, we show that those respondents who heard the words from the interviewer fall behind in all cognitive tests. This holds although all cognitive tests other than the word recall tests were administered using the same procedure. We further investigate to which extent controlling for individual characteristics and cognitive ability eliminates the gap in the word recall test, and we link performance differences to interviewers' propensity to read the words.

To adjust for differences in observable characteristics between the two groups of respondents, we rely on inverse-probability weighting (IPW). Specifically, we predict the probability of hearing the words from the interviewer using a logit-model (see Column 2 of Table 3.4) and estimate a quantity that would correspond to the population average treatment effect (see for example Hirano et al. 2003; Imbens 2004), if the treatment was randomized conditional on observable characteristics.[14]

$$\text{ATE}_{Int} = \sum_{i=1}^{N} \frac{\texttt{Int}_{ij} Y_{ij}}{\hat{p}_1 \sum_{i=1}^{N} \frac{\texttt{Int}_{ij}}{\hat{p}_1}} - \sum_{i=1}^{N} \frac{(1 - \texttt{Int}_{ij})\, Y_{ij}}{\hat{p}_0 \sum_{i=1}^{N} \frac{(1-\texttt{Int}_{ij})}{\hat{p}_0}} \tag{3.8}$$

Here $\hat{p}_1$ denotes the predicted probability of hearing the words from the interviewer and $\hat{p}_0$ is defined as $1 - \hat{p}_1$.[15]

Results are displayed in Table 3.7. All outcome variables are standardized to have a mean of zero and a standard deviation of one. The first column shows raw differences, in column (2) we adjust for individual characteristics, in column (3) we further control for age and experience of the interviewer, and in column (4) we control for irregularities during administration of the test reported by the interviewer. These irregularities include the presence of others for all tests. For the serial 7 subtraction task we also control for whether an aid was used and in the word recall test we account for interruptions during the test and whether problems hearing the words were recorded.[16]

For all cognitive tests, respondents who heard the words in the word recall test from

[14] Possible alternative estimands include the average treatment effect for the treated, where the control group is reweighted to resemble the treatment group and the average treatment effect for the untreated.

[15] We use weights, that are normalized to unity in the population as suggested in Imbens (2004).

[16] In the household sample, problems hearing the words were recorded for 857 respondents, 288 of whom heard the words from the interviewer.

__Table 3.7:__ Performance differences between individuals for who the words were read by the interviewer and other respondents: All cognitive tests

	Immediate word recall			
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.289***	-0.0868***	-0.0878***	-0.0766***
	(0.0288)	(0.0229)	(0.0234)	(0.0216)
N	24323	24323	24313	24313
	Delayed word recall			
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.3813***	-0.1919***	-0.1906***	-0.1842***
	(0.0317)	(0.0291)	(0.0293)	(0.0289)
N	24323	24323	24313	24313
	Serial 7 subtraction			
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.1131***	-0.0461*	-0.0443	-0.0444
	(0.0285)	(0.0273)	(0.0275)	(0.0274)
N	23662	23662	23652	23658
	Number series: Set 1			
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.1376***	0.0075	0.0045	0.0092
	(0.0303)	(0.0253)	(0.0251)	(0.0251)
N	11127	11127	11121	11126
	Number series: Set 2			
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.1242***	-0.0129	-0.0109	-0.0123
	(0.0331)	(0.0294)	(0.0299)	(0.0294)
N	11680	11680	11677	11680
	Verbal fluency: correct			
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.2568***	-0.0947***	-0.0973***	-0.0943***
	(0.03)	(0.0277)	(0.0278)	(0.0278)
N	24238	24238	24229	24237
	Numerical ability			
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.1877***	-0.067***	-0.067***	-0.0649***
	(0.0259)	(0.0224)	(0.0225)	(0.0225)
N	24175	24175	24166	24173
Individual controls	No	Yes	Yes	Yes
Interviewer characteristics	No	No	Yes	No
Unusual events during test	No	No	No	Yes

Notes: ATEs from inverse probability weighting models. All outcomes are standardized. Individual controls include: Hearing problems, Male, Marital status (3 levels), Dummies for age in 5 year categories, Linear age trend, Educational attainment (4 levels), Born in UKUrban area, Household size, Log of monthly net income per household member, Longstanding illness or disability. Interviewer characteristics include, Interviewer: Year of birth, Year started working as interviewerUnusal events vary by test and include at least the presence of others. For the recall tests they also include, whether the respondent had hearing problems during the test, used aides or was interrupted. Standard errors clustered at the interviewer level.

the interviewer display on average significantly lower test scores. This suggests that respondents with lower cognitive ability are significantly more likely to be administered the word recall test in the alternative mode. Performance differences range from 0.12 standard deviations in the number series to 0.38 standard deviations in the delayed word recall test. The fact that performance differences in the delayed word recall test clearly exceed differences in all other outcomes, suggests that mode effects might play an additional role here.

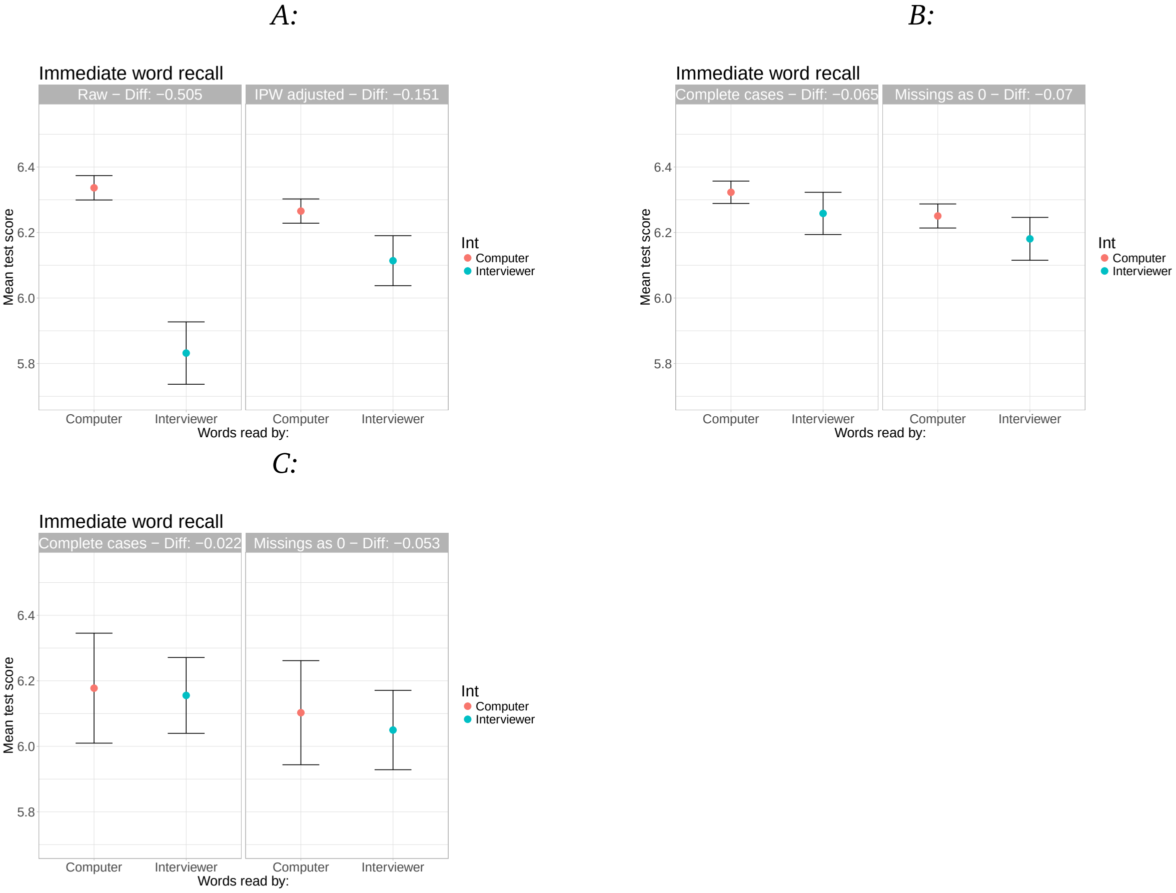
Adjusting for observed characteristics eliminates a large part of the differences in test scores. Nevertheless, for the delayed and immediate word recall test as well as numeric ability, verbal fluency and - to some extent - number 7 subtraction, a significant gap remains. Further controlling for interviewer characteristics or irregularities during the test, does not change the results.

In the additional cognitive tests, the number of missing test scores is higher than in the word recall tests. While our results in the main analysis are based on all available test scores, we explore two alternative approaches in the Appendix. In Table C.1, we repeat the analysis using only respondents for whom all test scores are available. To obtain the results in Table C.2, we make the assumption that the main reason for a missing test score is a respondents' inability to conduct the test. Therefore we set all test scores to the minimum test score. Both changes to the sample do not substantially alter the results qualitatively.

If performance differences are entirely driven by individual characteristics and cognitive ability, adjusting for both factors should eliminate the differences in average test scores. In Figure 3.8 and Figure 3.9, we display the sequence of raw and adjusted mean in test scores. Panel A repeats the exercise of Table 3.7, showing that adjusting for observed individual characteristics removes a major part of the differences in performance. In Panel B we estimate the propensity score using both observable characteristics and test scores in other cognitive tests.[17] The idea here is to use performance in other cognitive tests a proxy for unobserved cognitive ability. For the immediate word recall, the gap decreases to 0.15 words once we control for individual characteristics. If we further adjust for performance in other cognitive tests, the gap reduces to 0.06-0.07 words

[17] We omit the number series, as a mistake in the testing program appeared only in the second set of questions. Therefore test scores from both sets should not be pooled.

Figure 3.8: *Weighted means: Immediate word recall*



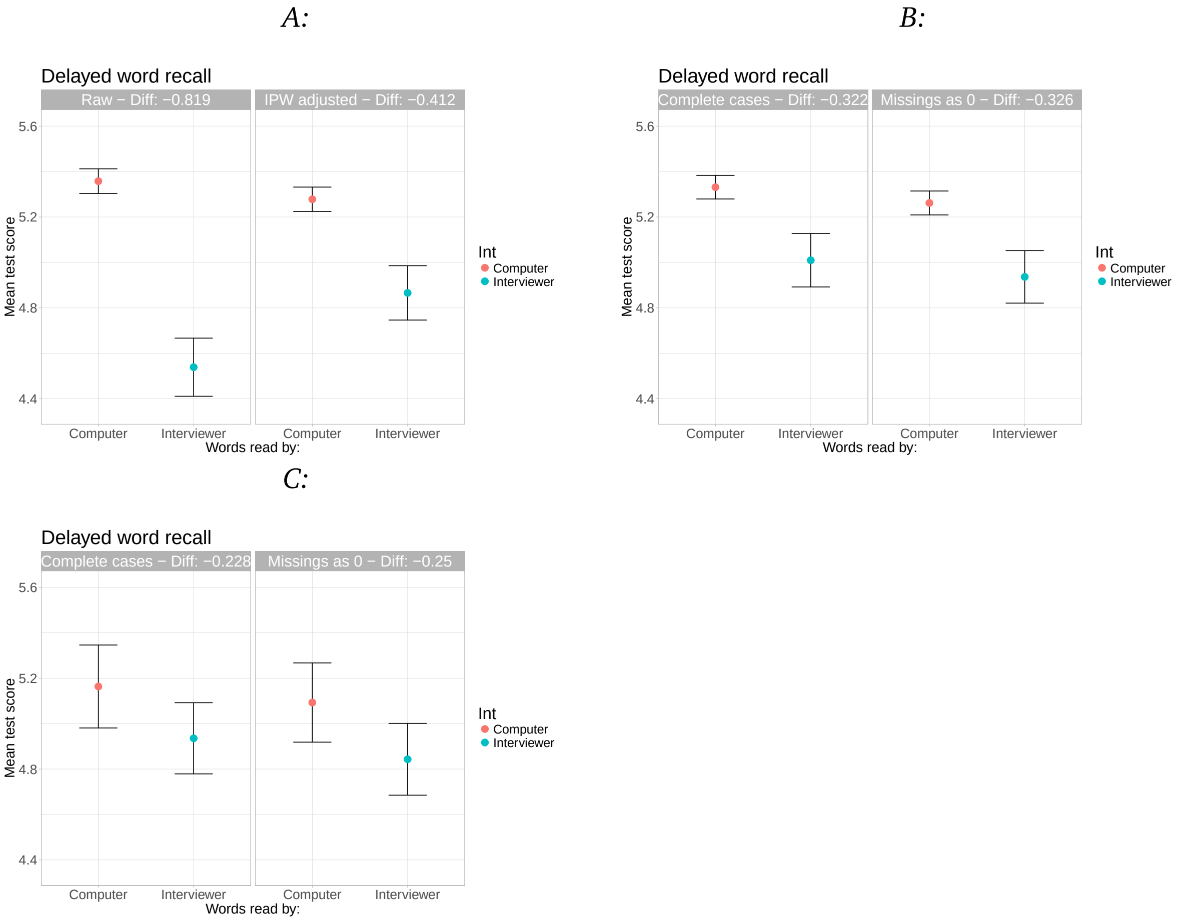
A:

B:

C:

Notes: Raw and adjusted differences for immediate word recall test. This table shows propensity score weighted means of test scores in the word recall test. Panel A: Raw test scores and test scores adjusted for individual and household controls. Panel B: Test scores adjusted for individual and household controls as well as test scores in the subtract 7, numeric ability and verbal fluency tests. Panel C: Similar to B, but estimated propensity scores include the mode of predicted interviewer effects. Weighted household sample.

Figure 3.9: *Weighted means: Delayed word recall*


A:

B:

C:

Notes: Raw and adjusted differences for delayed word recall test. This table shows propensity score weighted means of test scores in the word recall test. Panel A: Raw test scores and test scores adjusted for individual and household controls. Panel B: Test scores adjusted for individual and household controls as well as test scores in the subtract 7, numeric ability and verbal fluency tests. Panel C: Similar to B, but estimated propensity scores include the mode of predicted interviewer effects. Weighted household sample.

and is no longer statistically significant. In case of the delayed word recall test, we also see a reduction in performance differences when controlling for test scores in other cognitive tests. However, respondents who heard the words from the interviewer still obtain significantly lower test scores. Finally, we believe that the mode a respondent is assigned to, heavily depends on the interviewer she is visited by. To incorporate interviewer heterogeneity, we estimate a logit model that includes a random intercept for the interviewer. We predict the propensity score conditional of the mode of the random interviewer effect (Bates et al. 2015), i.e $\hat{p}_1 = \Lambda\Big(X\hat{\beta} + \hat{c_j}\Big)$, where $\Lambda(.)$ denotes the CDF of the logistic distribution.

This procedure brings the point estimates closer to zero.[18] In the delayed recall test, respondents who heard the words from the word recall test, still remember 0.23 words less than those who heard the words from the computer, even when adjusting for individual characteristics, performance in other cognitive tests and incorporating interviewer heterogeneity. Therefore we cannot rule out the existence of a mode effect. Nevertheless, the fact that the estimated gap has reduced from 0.82 words to 0.23 words suggests that selection effects play a greater role in our setting than mode effects.

We also explore, whether performance differences can be linked to interviewers' propensity to read the words. Figure 3.10 and Figure 3.11 plot performance differences within interviewers against the the share of interviews conducted in the alternative mode by interviewer. In Figure 3.10 we do not control for individual characteristics. The pattern displayed in this graph is highly consistent with the idea of interviewer heterogeneity in selection. For all outcomes, respondents who heard the words from the interviewer perform much worse if the interviewer only reads the words in a small portion of her interviews. These performance differences decrease as the proportion of of interviewers in the alternative mode increases. The gradient is most pronounced for the word recall test. Therefore we cannot rule out that a second mechanism also contributes to this pattern in the word recall test. Interviewers who read the questions more often might simply be better at reading the words. Nevertheless, the fact that a gradient is also present for the other cognitive tests, backs the interpretation that heterogeneity in selection is present in our data. Once we also control for individual characteristics (Figure 3.11), the patterns become less clear. There still exists a gradient in the word

[18] Our confidence intervals treat c_j as fixed and may therefore be treated with caution.

recall test as well the verb fluency test, while for other cognitive tests this does not hold. Taken together the evidence presented suggests that interviewers indeed vary in their assessment of respondents' ability to hear the words from the computer. For those interviewers who only deem a very small proportion unfit, performance differences are very large. However, much of this variation seems to be related to basic demographic characteristics - for example some interviewers read the words to respondents with small hearing problems, others do not - rather than the unobserved part of cognitive ability.[19]

3.3.3 Interviewer effects

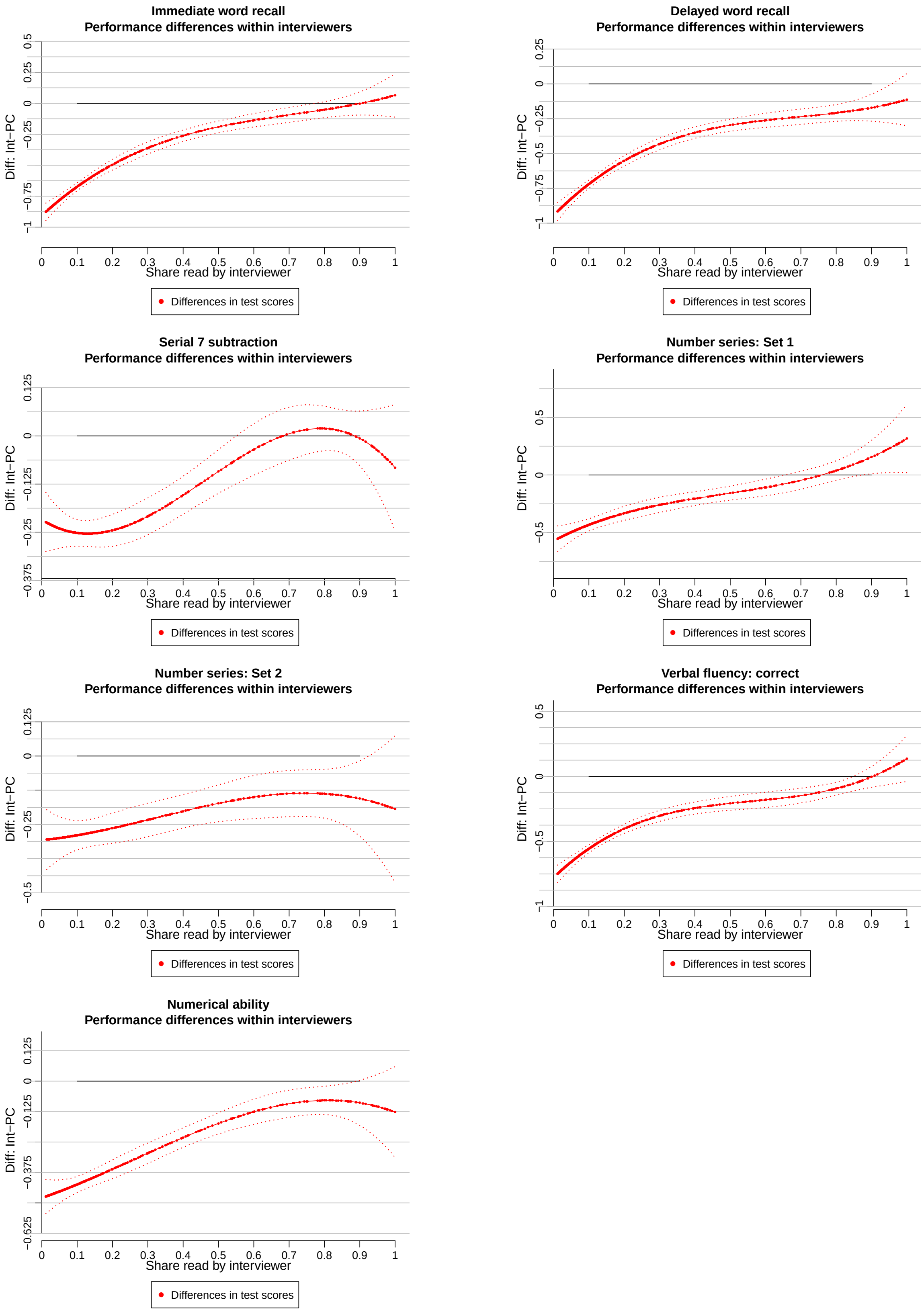
Now we turn to the question, whether administration of the word recall test via the computer successfully reduces interviewer intra-class correlations after all. A look at estimated interviewer intra-class correlations displayed in Figure 3.3, reveals two notable findings. Firstly, interviewer intra-class correlations in the word recall test are greater when the interviewer reads the questions herself than when the words are read by the computer. Secondly, in the delayed word recall test the estimated interviewer intra-class correlation is greater than for most other cognitive tests even when the computer reads the words.[20]

Similarly to performance differences, differences in the magnitude of interviewer intra-class correlation between the to groups of respondents, do not necessarily stem from mode effects alone. A recent literature has identified non-response error variance as additional source of observed interviewer intra-class correlations in surveys (Brunton-Smith et al. 2012; West and Olson 2010). This literature highlights the fact that different interviewers obtain participation from different sources of respondents. Differences in the pool respondents across to interviewers, contribute to observed interviewer intra-

[19] Figure C.2 in the Appendix displays raw average test scores for the two groups in the cross-section. This graph looks similar to Figure 3.11. Moreover it shows that average performance of the larger group - that is respondents who heard the words from the computer - is not affected by selection, while test scores of respondents who heard the words from the computer increase as interviewers read more frequently.

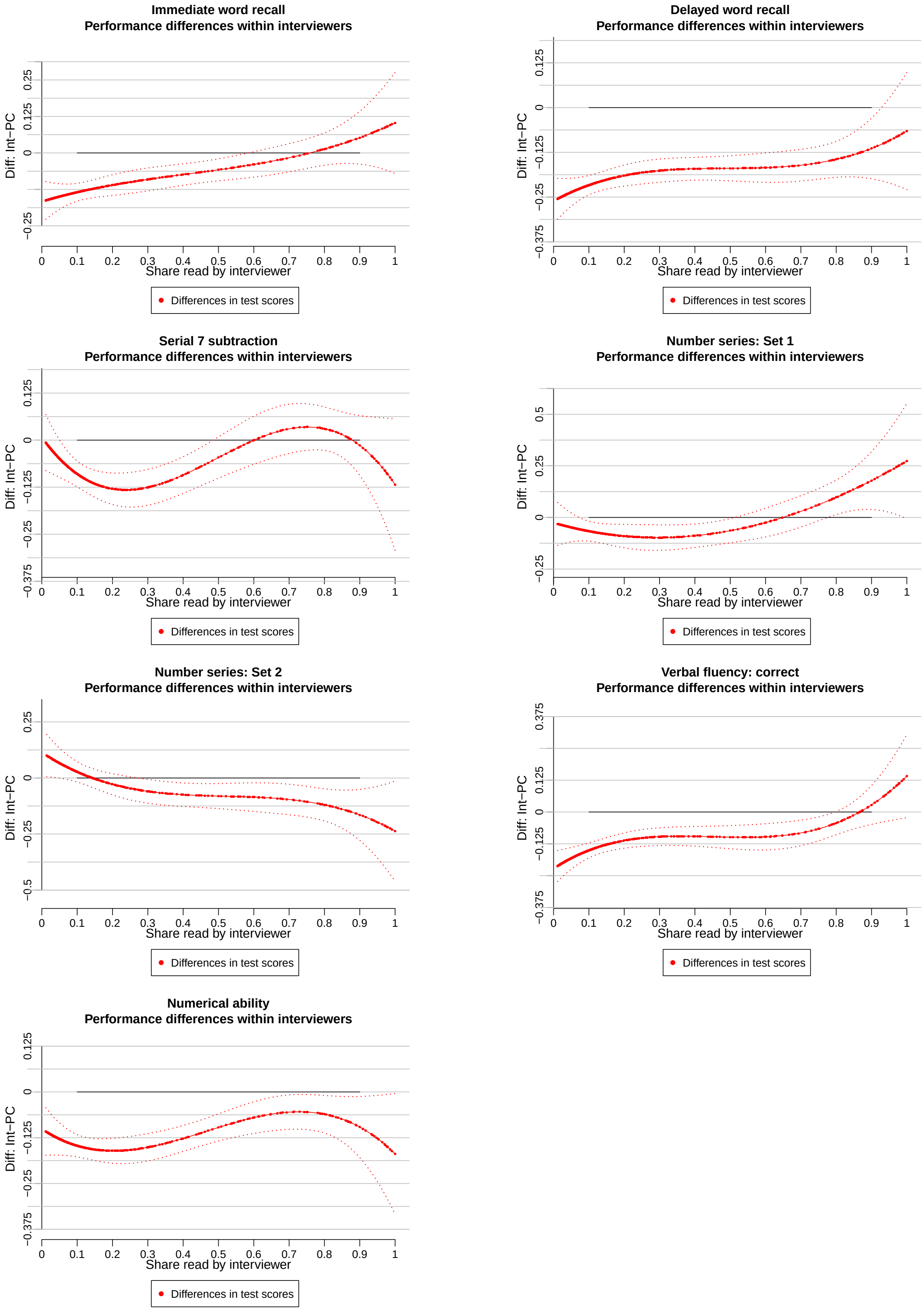
[20] The only cognitive test displaying a greater interviewer intra-class correlation than the word recall test, is the verbal fluency test. Administering this test is highly challenging for interviewers, as they need to write down all the animals a respondent names within a given time frame. Very high interviewer intra-class correlations in this task are not surprising, given heterogeneity in interviewers' ability to concentrate and writing speed.

Figure 3.10: *Performance by share read: Fixed effects (raw)*



Notes: Differences in test scores between alternative and default mode. We estimate the model $y = f_1(sread) + \psi_j + \epsilon$, where y is a standardized outcome, ψ_j is a set of interviewer dummies and $sread$ is the share of interviews, where the interviewer reads the words. $f_1(sread)$ is a semiparametric estimate of the difference between the two groups within interviewers and is approximated using cubic B-splines. The graphs plots $f_1(sread)$ over $sread$.

Figure 3.11: *Performance by share read: Fixed effects (controls)*



Notes: Differences in test scores between alternative and default mode. We estimate the model $y = f_1(sread) + X\beta + \psi_j + \epsilon$, and $sread$ is the share of interviews, where the interviewer reads the words. $f_1(sread)$ is a semiparametric estimate of the difference between the two groups within interviewers and is approximated using cubic B-splines. The graphs plots $f_1(sread)$ over $sread$.

class correlations. Our setting is related in the sense that interviewers differ in the pool of respondents they administer the word recall test in a specific mode.

Framework

We are interested in interviewer intra-class correlations (ICC), i.e. the share of residual variance that can be attributed to the variance of interviewer effects. In the absence of interviewer induced heterogeneity in selection, this quantity is - in our framework - defined as follows:

$$\text{ICC}_{full} = \frac{\sigma_u^2}{\kappa^2\sigma_\nu^2 + \sigma_u^2 + \sigma_\epsilon^2}, \tag{3.9}$$

Here σ_ν^2, σ_u^2, σ_ϵ^2 are the variances of unobserved cognitive ability, interviewer effects and the idiosyncratic error of observed test scores and κ is the slope parameter linking true cognitive ability to expected test scores.

To understand, how interviewer heterogeneity in selection might contribute to observed interviewer intra-class correlations, we again treat observed test scores within a mode of administration, as incidentally truncated variable. Assuming a bivariate normal distribution for both ω_i and ν_i, there exists a closed-form expression for the variance of observed test scores for a respondent interviewed by interviewer j conditional on the mode and observed characteristics (W. Greene 2012, p. 913):

$$\text{Var}\left(Y_{ij}|j, \texttt{Int}, X, Z\right) = \kappa^2\sigma_v^2\left(1 - \rho_{w\nu}^2\lambda\left(Z_i\gamma + c_j\right)\left(\lambda\left(Z_i\gamma + c_j\right) + \left(Z_i\gamma + c_j\right)\right)\right) + \sigma_\epsilon^2, \tag{3.10}$$

where the inverse mills ratio $\lambda(.)$ is defined as above. Truncation reduces the variance of observed test scores (W. Greene 2012, p. 913). The total variance of incidentally truncated test scores is the sum of the expectation of the variance within interviewers and the variance of expected test scores across interviewers.

$$\begin{aligned} \text{Var}\left(Y_{ij}|\texttt{Int}, X, Z\right) &= \text{E}\left(\text{Var}\left(Y_{ij}|j, \texttt{Int}, X, Z\right)|\text{Int}, X, Z\right) + \text{Var}\left(E\left(Y_{ij}|j, \texttt{Int}, X, Z\right)|\texttt{Int}, X, Z\right) \\ &= \text{E}\left(\kappa^2\sigma_v^2\left(1 - \delta\left(Z_i\gamma + c_j\right)\right)|X, Z\right) + \text{Var}\left(\kappa\rho_{w\nu}\sigma_\nu\lambda(Z_i\delta + c_j) + u_j|X, Z\right) + \sigma_\epsilon^2, \end{aligned} \tag{3.11}$$

where $\delta\left(Z_i\gamma + c_j\right) = \rho_{w\nu}^2\lambda\left(Z_i\gamma + c_j\right)\left(\lambda\left(Z_i\gamma + c_j\right) + \left(Z_i\gamma + c_j\right)\right)$. As $\delta\left(Z_i\gamma + c_j\right)$ is between zero and one, the contribution of the variance within interviewers to the total variance of incidentally truncated test scores is less than $\kappa^2\sigma_\nu^2$. If c_j and u_j are uncorrelated it follows that:

$$\text{Var}\left(E\left(Y_{ij}|j,\texttt{Int},X,Z\right)|\texttt{Int},X,Z\right)=\text{Var}\left(\kappa\rho_{w\nu}\sigma_\nu\lambda(Z_i\delta+c_j)\right)+\sigma_u^2\geq\sigma_u^2 \tag{3.12}$$

Therefore the share of the variance attributed to interviewers is elevated under incidentally truncation. Taken together this implies:

$$\text{ICC}_{Int}=\frac{\text{Var}\left(E\left(Y_{ij}|j,\texttt{Int},X,Z\right)|\texttt{Int},X,Z\right)}{\text{E}\left(\text{Var}\left(Y_{ij}|j,\texttt{Int},X,Z\right)|\text{Int},X,Z\right)+\text{Var}\left(E\left(Y_{ij}|j,\texttt{Int},X,Z\right)|\texttt{Int},X,Z\right)}>\text{ICC}_{full} \tag{3.13}$$

In the absence of a correlation between c_j and u_j, incidental truncation always increases the interviewer intra-class correlations. The amount of the increase depends on the degree of selection and will usually not be the same in the two groups.

If u_j and c_j are correlated, the change in the interviewer intra-class correlation depends on the exact variance-covariance structure. A negative correlation between c_j and u_j implies a positive correlation between u_j and the inverse mills ratio, contributing to an increase in the interviewer intra-class correlation. The reverse holds for a positive correlation between u_j and c_j. On the other hand, any correlation between u_j and c_j will make interviewers within one mode more similar and therefore decrease the variance of interviewer effects within a given mode.

Results

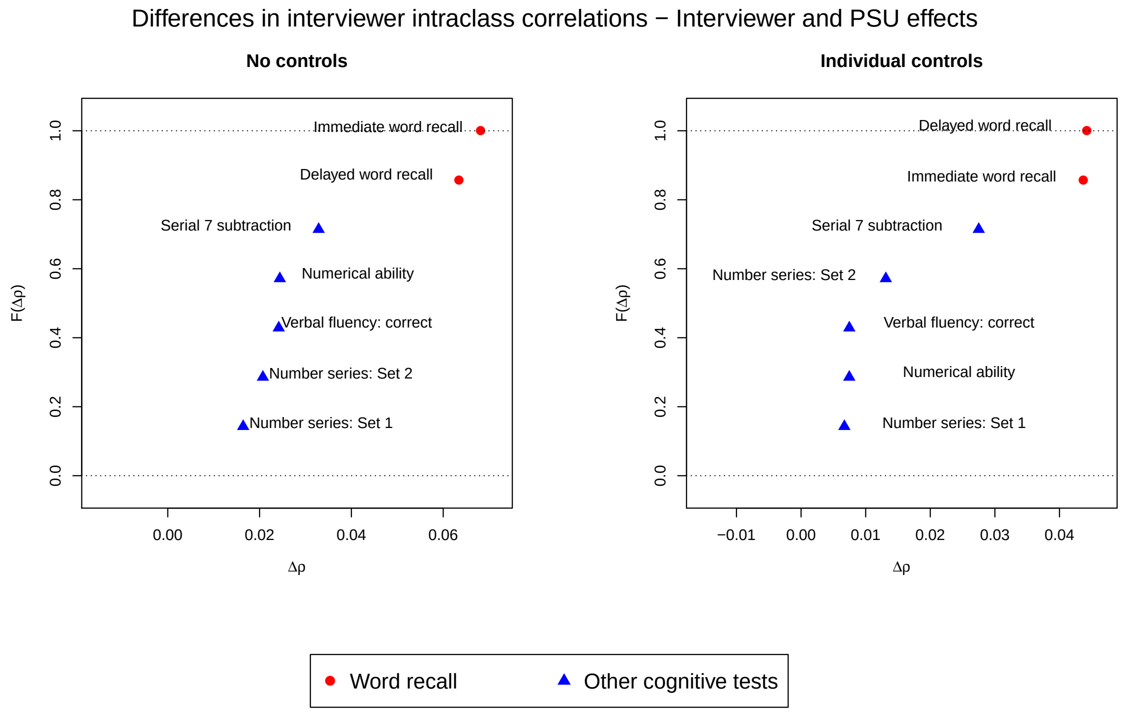
Differences between the interviewer intra-class correlations in the word recall test may mainly stem from heterogeneity in the pool of respondents across interviewers. In this case we expect to see elevated interviewer intra-class correlation for the group of respondents, who heard the words from the interviewer, also in other cognitive tests. If, on the other hand, shifting the task of reading the words from the interviewer to the PC, does in fact reduce interviewer effects, this should not impact interviewer intra-class correlations in other cognitive tests.

To obtain interviewer intra-class correlations, we estimate linear models that include a random intercept for both interviewers and PSUs (Bates et al. 2015) separately for the two groups of respondents. We calculate interviewer intra-class correlations as:

$\hat{\rho}=\frac{\hat{\sigma}^2_{int}}{\hat{\sigma}^2_{int}+\hat{\sigma}^2_{PSU}+\hat{\sigma}^2_{\epsilon}}$, where $\hat{\sigma}^2_{int}$ is the estimated variance of the random interviewer effect, $\hat{\sigma}^2_{psu}$ is the estimated variance of the random PSU effect and $\hat{\sigma}^2_{\epsilon}$ is the residual variance. For each outcome we calculate the difference in interviewer intra-class correlation between the two groups as: $\rho_{intread}-\rho_{PCread}$.

For all cognitive tests, the interviewer intra-class correlation is higher in the group of respondents hearing the words from the interviewer (see Figure 3.12). However, for most cognitive tests other than the word recall test, the gap is very close to zero. Taken together, these findings imply the existence of a small selection effect, while administration via the computer does seem to reduce interviewer effects.

Figure 3.12: *Differences in intra-class correlation (PSU)*



Notes: Empirical CDF of differences in interviewer intra-class correlations: $\Delta\rho = \rho_{intread} - \rho_{PCread}$. Estimated using cross classified random effect models, with one random intercept for interviewers and one random intercept for PSUs. Outcomes:(1) Word recall, (2) Other cognitive tests (Serial 7 subtraction, Number series: Set 1, Number series: Set 2, Verbal fluency: correct, Numerical ability). Controls include: Hearing problems, Male, Marital status (3 levels), Dummies for age in 5 year categories, Linear age trend, Educational attainment (4 levels), Born in UK, Urban area, Household size, Log of monthly net income per household member, Longstanding illness or disability.

3.4 Discussion

3.4.1 Implications for applied researchers

When using cognitive test data collected under imperfect compliance with the study protocol, applied researchers need to decide which observations they include in their sample. Part of the variation in observed test scores may stem from mode effects rather than variation in cognitive ability. Restricting the analysis to respondents for whom the test was administered under the default procedure solves the problem of mode effects. However, a reduction in sample size comes along with a loss in power and moreover, observed test scores in the restricted sample may be affected by incidental truncation. We argue any decision should be preceded by a careful analysis to which extend mode effects and incidental truncation appear to be present in the data. In the following, we briefly discuss how these two problems can affect results in simple OLS-regressions and derive some suggestions on how to address these.

To understand the impact of mode effects, we consider the case when assignment to the mode is random. We denote observed test scores by Y and we are interested in the impact of cognitive ability on an outcome Z. As Y does not perfectly measure cognitive ability, the estimated coefficient of a regression of Z on Y always suffers from attenuation bias. Mode effects can aggravate this attenuation bias. When Y is administered using either mode g' or mode g'', we observe $Y_{g'}$ with probability $p_{g'}$ and $Y_{g''}$ with probability $(1 - p_{g'})$. We assume that $Y_{g'}$ and $Y_{g''}$ differ in their means, but not in the slope parameter κ. Regressing Z on pooled test scores from both modes yields in the case of no further controls:

$$\text{plim}\hat{\beta}_{\xi pooled} = \beta_{\xi}\frac{\sigma_{\xi}^2}{\kappa\sigma_{\xi}^2 + \sigma_e^2 + p_{g'}(1-p_{g'})\delta_{g'g''}^2}, \tag{3.14}$$

where $\hat{\beta}_{\xi}$ denotes the effect of true cognitive ability on z, σ_e^2 is the variance of the measurement error of Y and $\delta_{g'g''}$ denotes the difference in expectations between $Y_{g'}$ and $Y_{g''}$. Therefore, we advise to either run such regressions on the restricted sample or to include a mode dummy.[21] If the researcher has reason to believe that the mode also

[21] From the Frisch–Waugh–Lovell theorem, the estimated coefficient $\hat{\beta}_y$ resulting from estimating regression model $Z = \beta_0 + \beta_y Y + \delta\texttt{Int} + \epsilon$ is equivalent to the $\tilde{\beta}_y$ resulting from the regression model

impacts the slope parameter κ, results obtained on a restricted sample will be more informative.

On the other hand, if the mode of administration is random conditional on observables, using Y instead of $Y_{g'}$ or $Y_{g''}$ as left hand side variable does not bias the estimation. Potential increases in residual error variance and the resulting loss in precision may even be offset by gains in sample size.

Next we consider selection effects. It is well-known that incidental truncation leads to inconsistency of the OLS-estimator, as the inverse mills ratio becomes part of the error term and causes omitted variable bias (see for example W. Greene 2012, Chapter 19.5). Therefore, one should use the full sample in case the data exhibit selection that is not accounted for by observed characteristics. In contrast, if we are interested in the effect of test scores on an additional outcome - and selection is not related to this additional outcome - both using the full sample and using a restricted sample will yield consistent estimates.[22]

Mode effects and incidental truncation may also occur simultaneously. Here, restricting the sample to a certain mode or using the full sample and including a mode dummy will both lead to consistent estimates for the effect of cognitive tests scores on Z. When using test scores as dependent variable, we suggest to do the analysis on a restricted sample and apply an appropriate econometric method to adjust for selection effects.[23]

Finally, heterogeneity across interviewers can further complicate the problem. Estimators that deal with incidental truncation in the case of panel data are proposed in Kyriazidou (1997), Semykina and Wooldridge (2010), Vella and Verbeek (1999), and Wooldridge (1995).

3.4.2 Implications for survey designers

In our setting, administration of the word recall test with the help of a laptop appears to reduce interviewer effects without introducing substantial mode effects. Therefore we think that - overall - administration of the word recall test using the computer leads

$M_{\texttt{Int}}Z = M_{\texttt{Int}}\beta_0 + \beta_y M_{\texttt{Int}}Y + M_{\texttt{Int}\epsilon}$, where $M_{\texttt{Int}} = I_n - \texttt{Int}\,(\texttt{Int}'\texttt{Int})^{-1}\,\texttt{Int}'$. As Z is orthogonal to the mode of the test, $M_{\texttt{Int}}Z = Z$. Multiplying Y by $M_{\texttt{Int}}$, nets out the mode effect.

[22] Assuming there is common support in the test scores of the two modes.

[23] Using the full sample and not controlling for the mode, introduces omitted variable bias. Additionally controlling for the mode via the use of a mode-dummy, introduces selection bias.

to an improvement in data quality. Nevertheless, for the delayed word recall test, the gap between the two modes does not fully disappear when adjusting for observed characteristics and performance in other tests. This suggests that a small trade-off remains. Survey designers should therefore aim at keeping non-compliance with the default procedure at a minimum level. Interviewer training can help to increase compliance with the study protocol (Billiet and Loosveldt 1988). Moreover, we emphasize the following two points that appear to contribute to non-compliance in our setting. Firstly, for some interviewers the pattern in the sequence of modes is consistent with technical malfunction of the laptops' audio function. Survey administrators should therefore encourage interviewers to report and fix problems with their laptops. Secondly, hearing problems and an old age are the most important predictors for hearing the words from the interviewer. Therefore it should be made sure that the words can be played on a high volume facilitating administration for these groups. Finally, researchers can use statistical methods to address mode- or selection effects in their data. Therefore it is important that para- data on deviations from the study protocol is provided together with the actual data.

3.5 Conclusion

In order to eliminate the impact of interviewers, surveys may shift administration of a cognitive test partially or fully to technical devices. In this work, we study an unintended side effect of this approach. Interviewers may not always fully comply with the default procedure and forgo the help of the technical device for some respondents. As a result, the cognitive test is effectively administered in a mixed-mode design.

In our setting, the words in the word recall test are supposed to be read to respondents by the laptop of the interviewer. However, for about 20% of our observations, interviewer read the words themselves. Within the group of respondents who hear the words from the interviewer, we observe worse performance and higher interviewer intra-class correlations. We find that differences between the two groups of respondents mainly stem from selection. Furthermore, shifting the administration of the word recall test to the computer, does seem to successfully reduce interviewer effects. Therefore we conclude, that involvement of technical devices is beneficial for data quality.

Appendix A

Birth in Times of War - An investigation of health, mortality and social class using historical clinical records

A.1 Birth records

Almost all birth records since the foundation of the hospital in 1884 have been preserved. Birth records span around four to eight pages and generally contain background information on the mother, information on the pregnancy, medical examinations, a labour protocol including detailed notes, characteristics of the newborn child and observations during childbed. A compressed version of the birth records is provided by two series of journals, called birth journals and main journals. Birth journals have been filled in by midwives shortly after childbirth. Main journals make a more official appearance, suggesting that they were kept by a hospital clerk. Both journals contain the birth number, name, age and parity of the mother, the date of birth, sex, length and weight of the child, and short notes on medical issues. Main journals additionally give the date of discharge and the fetal position. Birth journals include information on the socio-economic status, mostly in form of the occupation of either the father or the mother. Main journals are only available for the common section. We digitized the information contained in the main journal and birth journals for a period starting in November 1937 and ending in October 1941. Since main journals do not exist for the private section, the date of discharge and the fetal position were added from the birth records. Apart from birth records, parts of the correspondence of the management of the hospital have been preserved in archives and the hospital itself. We use this material to corroborate our findings with qualitative evidence.

A.2 HISCO-HISCLASS

As mentioned in the main text the birth journals - both from the general and private ward - contain parental status and/or occupation. This allows us to derive a measure of social classes which will be explained in this section. These specific occupations were originally recorded by hospital personnel with additional (grand-)parental socio-economic information in the medical files to ensure Aryan ancestry and the patient's health insurance among other things.[1] If fatherhood is known and stated in the medical files, usually the (civil) profession of the child's father (e.g. grocer, in German "Krämer") or her relation to him (e.g. grocer's wife, in German "Krämersfrau") was registered in the birth journals.[2]

In a first step we standardized the spelling of the occupations (to the male form) and separated non-occupational information. In the following step we assigned a numerical 5-digit code according to the "Historical International Classification of Occupations" (HISCO) which was developed by Van Leeuwen et al. (2002) and is provided as an online database called "History Of Work Information System".[3] HISCO combines information on occupational tasks and duties and forms a system of 1675 historical and international comparable occupations. It was developed upon ILO's modern-day "International Classification of Occupations"' from 1968 (ISCO68) and adjusted with 18th-20th century occupations from several countries in Europe and America. HISCO's hierarchical structure - similar to ISCO68 - into 9 major groups, 76 minor groups, 296 unit groups and finally 1675 occupations has descriptions for each level and therefore allows comparisons to modern-day occupational groups and professions as well. HISCO has three additional variables (Status, Relation, Product) from which the variable "Status" is the

[1] Due to data privacy regulations we were not able to use this valuable information.

[2] In very rare cases the relation of the pregnant woman to her father's occupation (e.g. grocer's daughter in German "Krämerstochter") was entered if she was too young to have a own job and most likely unwed. In other cases the female notation of an occupation was recorded (e.g. in German "Krämerin"). This is a sign that the pregnant woman is unwed, but in some cases it might just indicate her job. For some observations more and/or other non-occupational information is recorded, e.g. "unwed", "student" or "housewife". If just "housewife" was recorded, a cross-check with the medical files most often revealed the relevant occupation.

[3] `http://historyofwork.iisg.nl`

most important one. It contains information about supervisory tasks and skill levels within an occupation (e.g. master backer, journeyman baker, apprentice baker, baker's helper) which would otherwise be lost because HISCO codes only incorporate the raw definition of an occupation (e.g. baker).

In a last step we translated the HISCO into HISCLASS codes, the measure of social status which we will later use in our empirical analysis.[4] The Historical International Social Class Scheme (HISCLASS) invented and explained by Van Leeuwen and Maas (2011) builds upon HISCO, assigns each occupational code one out of 12 social classes, and defines a social class as "a set of individuals with the same life chances" (Van Leeuwen and Maas 2011, p.18). These social classes are derived in the following step-wise procedure: First Van Leeuwen and Maas (2011) identified (1)"type of work" (manual vs. non-manual), (2) "skill level" (4 levels), (3) "supervisory tasks" (yes vs. no) and (4) economic sector (primary vs. other sectors) as the four relevant dimensions of social class through an intensive literature review of existing historical class/status schemes. Second they used the American Dictionary of Occupational Titles (DOT) to grade the 1675 HISCO occupations along these dimensions. Third if there is additional information over and above the simple occupation name (e.g. baker) in the "Status" variable mentioned before (e.g. master, apprentice, helper, etc.) this is taken into account by promoting or demoting individuals into a higher/lower social class respectively. Since the DOT was constructed for modern-day occupations these grades were adjusted with help of expert historians which was only necessary in a few cases and finally led to 12 distinct social classes. In our empirical analysis we rely on the previous literature and use a compressed 7-class version of HISCLASS (see Abramitzky et al. (2011) and Schumacher and Lorenzetti (2005) and references therein). This simplifies the interpretation of regression coefficients, attenuates possible coding errors and increases sample size within classes. Table A.1 in the appendix shows the original and compressed HISCLASS versions along with the underlying dimensions of social class and the number of observations in each class.

To adjust HISCO/HISCLASS to the specific Bavarian background and our data set some-

[4] For the actual translation we relied on a SPSS program provided in the "History Of Work Information System" database (`http://historyofwork.iisg.nl/docs/hisco_hisclass12_book@_numerical.inc`) which we corrected and translated into a Stata. A commented do-file is available on request from the authors.

times we had to refine or deviate from the suggested coding procedure by Van Leeuwen and Maas (2011) and Van Leeuwen et al. (2002). First and foremost we had to rely on the occupational information about the child's father/ pregnant woman's husband in most cases, because either the pregnant women were not working at all or their own job was not recorded in the birth journals. This is in contrast to Van Leeuwen and Maas (2011) who don't assign a HISCLASS code to them at all. Nevertheless we are confident that this measure captures the relevant social class of a family since - as mentioned in section 1.2.1 - Nazi-propaganda promoted housewife-dom and the husband was the head of the most households. Secondly due to the fact that Munich is a state capital, there are a lot of public sector occupations which were strictly hierarchically ranked according to the "Führerprinzip" and comparable to military ranks.[5] This allowed us to use equivalent HISCLASS codes of military ranks as a benchmark for police, postal, railway, educational and other governmental HISCO codes and adjust the previously assigned HISCLASS codes if there were large discrepancies.

[5] In English: "leader principle" (see Frei 2013)

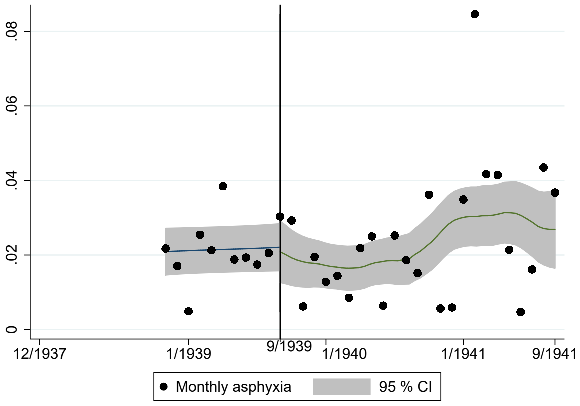
Table A.1: Original and compressed HISCLASS

HISCLASS				Dimensions of social class				# of obs.
Original classification		Compressed classification		Type	Skill	Supervisory	Economic	in
Nr	Name	Nr	Name	of work	level	tasks	sector	data set
1	Higher managers	1	Higher managers & professionals	non-manual	high	yes	mainly other	292
2	Higher professionals					no		397
3	Lower managers	2	Lower managers clerical and professionals, clerical and sales		medium	yes		554
4	Lower professionals and clerical and sales personnel					no		507
5	Lower clerical and sales personnel				low			968
6	Foremen	3	Skilled workers	manual	medium	yes		222
7	Medium skilled workers					no		2,014
8	Farmers and fishermen	4	Farmers and fishermen				primary	664
9	Lower skilled workers	5	Lower-skilled workers		low		other	1,343
10	Lower skilled farm workers	6	Farm workers				primary	118
11	Unskilled workers	7	Unskilled workers		unskilled		other	2,719
12	Unskilled farm workers	6	Farm workers				primary	115

Source: In style of Schumacher and Lorenzetti (2005) and Van Leeuwen and Maas (2011) **Notes:** Other economic sector refers to the industrial or service sector. Classes 1 and 3 of the original HISCLASS system contain only 3 occupations which are in the primary sector.

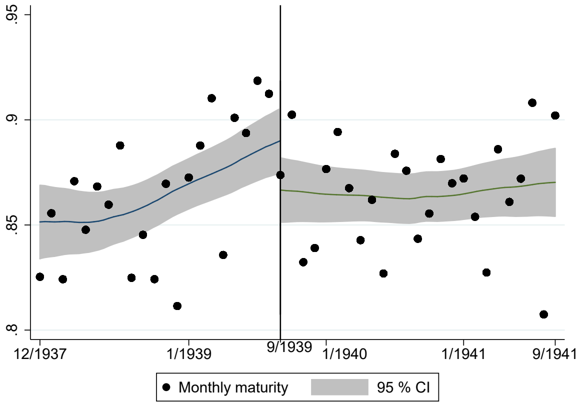
A.3 Additional Figures and Tables

Figure A.1: *Raw asphyxia rates by month of birth*



Notes: Asphyxia rates (monthly averaged) and local linear regressions with a ROT bandwidth and an Epanechnikov kernel separately for the pre-war and the war period.

Figure A.2: *Raw average maturity by month of birth*



Notes: Maturity (monthly averaged) and local linear regressions with a ROT bandwidth and an Epanechnikov kernel separately for the pre-war and the war period.

Table A.2: *Livebirths 1938-1940*

	Hospital	Munich	Bavaria
1938	2171	12164	168391
1939	2297	13028	179129
1940	2269	13741	174311

Source: Bayerisches Statistisches Landesamt (1937-1942)

Table A.3: *Descriptive statistics - Miscarriages*

General characteristics	N	Mean	SD	Min	Max
Birth after 9/1939	1194	0.560	0.497	0	1
General ward	1194	0.956	0.204	0	1
Mother	N	Mean	SD	Min	Max
Age of mother	1194	29.775	6.364	14	48
Parity	1184	2.994	2.308	1	18
Status is wife	1194	0.680	0.467	0	1
Status is own job	1194	0.270	0.444	0	1
Status is single, divorced or widowed	1194	0.033	0.178	0	1
Social status	N	Mean	SD	Min	Max
Higher managers & professionals	1125	0.063	0.243	0	1
Lower managers & professionals, cleric	1125	0.276	0.447	0	1
Foremen & skilled workers	1125	0.238	0.426	0	1
Farmers	1125	0.028	0.166	0	1
Lower skilled workers	1125	0.156	0.363	0	1
Unskilled workers	1125	0.228	0.419	0	1
Farm workers	1125	0.012	0.107	0	1
Infant	N	Mean	SD	Min	Max
Male	174	0.667	0.473	0	1
Birth weight	146	389.322	272.975	20	1870
Length of infant	178	24.458	8.197	9	90
No. of infants	1194	1.020	0.140	1	2

Notes: Descriptive statistics of miscarriages.

***Table A.4:** Mean comparison - Miscarriages*

General characteristics	Mean before war	Mean after war	Diff	SD	p	N before war	N after war
General ward	0.950	0.961	0.0107	0.012	0.371	525	669
Mother	Mean before war	Mean after war	Diff	SD	p	N before war	N after war
Age of mother	29.667	29.859	0.1928	0.371	0.603	525	669
Parity	3.033	2.964	-0.0689	0.135	0.611	518	666
Status is wife	0.632	0.717	0.0851**	0.027	0.002	525	669
Status is own job	0.310	0.238	-0.0728**	0.026	0.005	525	669
Status is single, divorced or widowed	0.036	0.030	-0.0063	0.010	0.544	525	669
Social status	Mean before war	Mean after war	Diff	SD	p	N before war	N after war
Higher managers & professionals	0.077	0.052	-0.0249	0.015	0.089	493	632
Lower managers & professionals, cleric	0.252	0.294	0.0428	0.027	0.111	493	632
Foremen & skilled workers	0.209	0.261	0.0522*	0.026	0.042	493	632
Farmers	0.037	0.022	-0.0144	0.010	0.151	493	632
Lower skilled workers	0.152	0.158	0.0061	0.022	0.780	493	632
Unskilled workers	0.260	0.203	-0.0571*	0.025	0.023	493	632
Farm workers	0.014	0.009	-0.0047	0.006	0.464	493	632
Infant	Mean before war	Mean after war	Diff	SD	p	N before war	N after war
Male	0.609	0.724	0.1149	0.071	0.109	87	87
Birth weight	392.877	385.767	-7.1096	45.336	0.876	73	73
Length of infant	24.333	24.585	0.2519	1.232	0.838	90	88
No. of infants	1.019	1.021	0.0019	0.008	0.819	525	669

Notes: T-tests on the equality of means by war. Only miscarriages. Significance levels: ***$p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

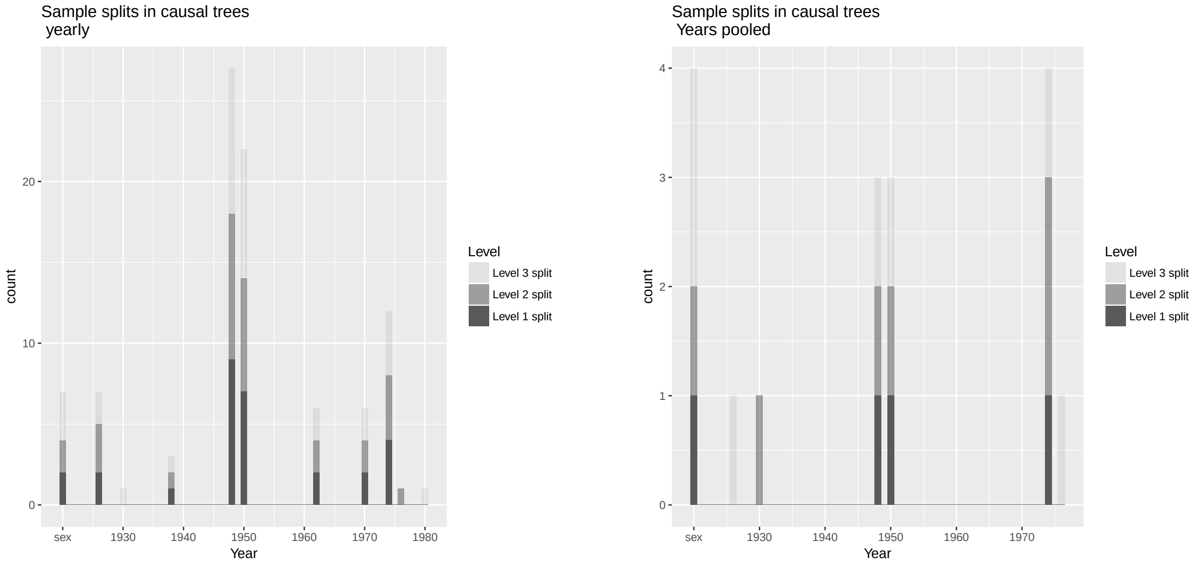
Appendix B

Does the wall still exist? Health differences between East and West Germans

B.1 Data driven partitioning of the sample

When splitting my sample into cohort groups I draw on historical events to define the cut-offs. In this appendix I additionally employ a data driven approach to identify cut-offs predicting cohort level heterogeneity in the effect of having lived in East Germany prior to 1989. Specifically I estimate a modified version of regression trees - so called "causal trees" - introduced in Athey and Imbens 2016. While standard regression trees split the sampling according to covariates in a way that minimizes prediction error in the outcome variable, causal trees are designed to detect treatment effect heterogeneity. Since the true treatment effect is generally unobserved, causal trees use an unbiased estimator of the mean-squared error as goodness-of-fit criterion. A tree is constructed by sequentially evaluating all potential sample splits and then choosing the partitioning which maximizes the goodness-of-fit criterion. In order to avoid overfitting, cross validation is used to select the optimal complexity parameter preventing the tree from growing to large. Rather than using the same sample for both constructing the tree and estimating the treatment effect at the leaves Athey and Imbens (2016) propose what they call "honest splitting". The treatment effect at each leave is estimated using a separate estimation sample, therefore the partition can be treated as if it was exogenously given. Athey and Imbens (2016) provide adjusted goodness-of-fit criteria for constructing the tree under anticipation of honest-splitting, which neglect systematic bias in estimation. In this exercise I am only interested in learning at which years of birth to split the sample (constructing the tree) and not in estimating treatment effects at each leave. Therefore I use the whole sample for constructing the tree. As I use the honest goodness-of-fit criteria, I construct trees in a way that pretends a second sample to estimate treatment effects was available. I construct both one tree for each outcome and survey year separately as well as one tree for each outcome pooling all survey years. Depending on the year and outcome, the algorithm does not necessarily suggest to split the sample at all. In Figure B.1 I show the distribution of splits for the first three levels of the tree. When considering the specifications by year and outcome, suggested splits occur most frequently during the early 1970s. This corresponds to the second cut-off used for the analysis, which exploits a cut-off rule in school entry. However, for some outcomes and years, the algorithm proposes a slightly earlier cut-off in the late 1960s. An additional cluster of splitting points is observed for the years 1948-1951, coinciding closely with the first cut-off. Furthermore the graph suggests the existence of further treatment effect heterogeneity within the old cohort, with the very early cohorts 1925-1930 being different from later born cohorts. When looking at the specifications where I pool all survey years, the picture looks very similar. All in all this exercise suggests that the cut-offs employed for the main analysis do indeed capture heterogeneity in the effect of having lived in East Germany prior to 1989. There seems to exist further heterogeneity within the oldest cohort group that is not explored in this study.

Figure B.1: *Sample splits from causal trees*



Notes: Sample splits from causal trees. Causal trees are constructed using the R package "causalTree" based on Athey and Imbens 2016 using the honest splitting and cross-validation criteria. I allow for partitions according to year of birth and require a minimum leaf size of 200 observations for both treatment and control group. In the figure on the left-hand side one tree is constructed for each outcome and survey year, while all years are pooled in the figure on the right-hand side. Regression are weighted using cross sectional weights provided by the SOEP.

B.2 Demediation function

$$
\begin{aligned}
&E\left(Y_i-\gamma(t,M_i,x)\mid T_i=t,X_i=x\right) \qquad \text{(B.1)}\\
=&E\left(Y_i-E\left(Y_i(t,M_i)-Y_i(t,0)\right)\mid X_i=x\right)\mid T_i=t,X_i=x)\\
=&E\left(Y_i-E\left(Y_i(t,M_i)\right)+E\left(Y_i(t,0)\right)\mid T_i=t,X_i=x\right)\\
=&E\left(Y_i-E\left(Y_i(t,M_i)\right)\mid T_i=t,X_i=x\right)+E\left(Y_i(t,0)\mid T_i=t,X_i=x\right)\\
=&E\left(Y_i(t,M_i(a))-E\left(Y_i(t,M_i(t))\right)\mid T_i=t,X_i=x\right)+E\left(Y_i(t,0)\mid T_i=t,X_i=x\right)\\
=&E\left(Y_i(t,0)\mid X_i=x\right)
\end{aligned}
$$

B.3 Estimation of standard errors of the direct effect

Ignoring the fact that the demediation function itself is estimated in a first step will lead to inconsistent estimation of the standard error of the direct effect in the second step. As suggested by Acharya et al. (2016) I obtain consistent standard errors for two-step estimators derived in Newey and McFadden (1994). Estimation of the first and the second stage can be carried out jointly within a GMM framework, where the sample moment conditions are given by

$$\sum_{i=1}^{n} \tilde{g}_i(D_i, M_i, X_i, L_i, Y_i\delta, \beta) = \sum_{i=1}^{n} \left(m_i(D_i, M_i, X_i, L_i, Y_i, \delta)', g_i(D_i, X_i, Y_i, M_i, \delta_m, \beta)'\right)' = 0$$

and the identity matrix is used as weighting matrix. Here n denotes the number of observations. The first component of the moment function:

$$m_i(D_i, M_i, X_i, L_i, Y_i, \delta) = (D_i, M_i, X_i, L_i)' \cdot (Y_i - \delta_0 - D_i\delta_d - M_i\delta_m - X_i\delta_x - L_i\delta_l)\,,$$

provides the moment conditions for estimating the vector of coefficients $\hat{\delta}$ in the first stage, while setting the expectation of

$$g_i(D_i, X_i, Y_i, M_i, \delta_m, \beta) = (D, X)' \cdot (Y_i - M_i\delta_m - \beta_0 - D_i\beta_{CDE} - X_i\beta)$$

to zero will return the second step estimator $\hat{\beta}$ as a function of δ_m. Defining $\tilde{G}$ to be expectation of the matrix of the first derivatives of $\tilde{g}$ with respect to δ and β, the asymptotic variance of $(\hat{\delta}', \hat{\beta}')'$ is given by

$$var(\hat{\delta}', \hat{\beta}')' = \frac{1}{n}\tilde{G}^{-1} E\left(\tilde{g}(D, M, X, L, Y, \delta, \beta)\tilde{g}(D, M, X, L, Y, \delta, \beta)'\right)\tilde{G}^{-1}.$$

When replacing the population moments by sample analogs, a consistent variance estimator is obtained. Since most variation of residency in 1989 stems from variation across households rather than variation within households, I cluster all standard errors at the household of origin level. This is archived by replacing the population moment $E\left(\tilde{g}(D, M, X, L, Y, \delta, \beta)\tilde{g}(D, M, X, L, Y, \delta, \beta)'\right)$ by the following sample moment:

$$\frac{nc}{(nc-1)}\frac{1}{n}\sum_{c=1}^{nc}\left(\tilde{g}_c(D, M, X, L, Y, \hat{\delta}, \hat{\beta})\tilde{g}_c(D, M, X, L, Y, \hat{\delta}, \hat{\beta})'\right),$$

where $c \in \{1, ...nc\}$ denotes the clusters and

$$\tilde{g}_c(D, M, X, L, Y, \hat{\delta}, \hat{\beta}) = \sum_{i \in c} \tilde{g}_i(D_i, M, X, L, Y, \hat{\delta}, \hat{\beta}).$$

B.4 Estimating the CDE using nonlinear models

Sequential g-estimation can also be modified to allow for the estimation of nonlinear models. For rare binary outcomes Vansteelandt (2009) propose a multiplicative direct effects model. They model the relative change in the probability that the potential outcome is equal to one when the treatment changes from 0 to d and the Mediator is held constant at m as:

$$\frac{Pr(Y(d,m)=1)}{Pr(Y(0,m)=1)} = \exp{(\beta_{treat} * d)}, \tag{B.2}$$

and the relative risk when M changes from 0 to m as:

$$\frac{Pr(Y(d,m)=1|D=d,L)}{Pr(Y(d,0)=1|D=d,L)} = \exp{(\gamma_m m)}, \tag{B.3}$$

In this framework $\exp{(\gamma_m m)}$ measures how the relative risk ratio changes when M changes from 0 to m and the treatment is held constant. Invoking Assumptions 2.1 and 2.3 and applying the law of iterated expectations the potential outcome when M is fixed at zero -$Y(d,0)$- can be approximated as $Y\exp{(\gamma_m M)}$. When the outcome is rare, the relative risk ratio can be approximated by the odds ratio, that is γ_m can be estimated as the regression coefficient in a logistic regression. Here I collapse SAH into a binary variable that takes the value one if an individual reports a good or very good health status. As around 30% of observations in my sample have a good health status, approximating the relative risk ratio by the odds ratio does not work very well. Instead I directly use odds ratios. In a logistic regression framework the following equalities hold:

$$\frac{Pr(Y(d,0)=1|D=d,X)}{Pr(Y(d,0)=0|D=d,X)} = \exp{(\beta_{cde} d + \beta_x X)} \tag{B.4}$$

and

$$\frac{\Pr(Y(d,m)=1|D=d,X,L)}{\Pr(Y(d,m)=0|D=d,X)} = \frac{\Pr(Y=1|M=m,D=d,X,L)}{\Pr(Y=0|M=m,D=d,X,L)} = \exp{(\gamma_m m + \gamma_L L + \gamma_X X)}, \tag{B.5}$$

where I use Assumptions 2.1 and 2.2 for the first equality. Furthermore one can show:

$$\frac{\frac{\Pr(Y(d,m){=}1|D{=}d,L,X)}{\Pr(Y(d,m){=}0|D{=}d,L,X)}}{\frac{\Pr(Y(d,0){=}1|D{=}d,L,X)}{\Pr(Y(d,0){=}0|D{=}d,L,X)}} = \frac{\frac{\Pr(Y(d,m){=}1|D{=}d,X)}{\Pr(Y(d,m){=}0|D{=}d,X)}}{\frac{\Pr(Y(d,0){=}1|D{=}d,X)}{\Pr(Y(d,0){=}0|D{=}d,X)}} = \exp{(\gamma_m * M)}, \tag{B.6}$$

where I use Assumption 2.4 of no intermediate interactions in the first equality. Equation B.4 provides the CDE on a OR-scale, but $Y(d,0)$ is not generally observed.

Combining Equation B.4 and Equation B.6 yields:

$$\frac{\Pr(Y(d,m){=}1|D{=}d,X)}{\Pr(Y(d,m){=}0|D{=}d,X)} = \frac{\Pr(Y(d,0){=}1|D{=}d,X)}{\Pr(Y(d,0){=}0|D{=}d,X)} * \exp{(\gamma_m * M)}$$
$$= \frac{\Pr(Y{=}1|D{=}d,M{=}m,X)}{\Pr(Y{=}0|D{=}d,M{=}m,X,)} = \exp{(\beta_{cde}d{+}\beta_x X)} * \exp{(\gamma_m M)} \tag{B.7}$$

That implies that the odds of observing Y equal one conditional on D, M and X is the product of the odds of $Pr(Y = 1)$ when M is held constant at zero times the change in the odds ratio when M changes from zero to M. I replace γ_m by the estimated coefficient $\hat{\gamma_m}$ from a logistic regression model on Y on D, M, L and X. To estimate β_{cde} and β_x in the second stage I fit a logistic regression model of Y on D, X and $M * \hat{\gamma_m}$, where I restrict the coefficient of $\gamma_m * M$ at to be equal to one. Standard errors for the estimated coefficients are obtained as explained in Appendix B.3.

The CDE for count data such as the number of diagnoses can be obtained in a similar fashion as shown by Vansteelandt (2009) for a binary outcome modeled on a risk ratio scale (see also Valeri and VanderWeele 2013). Assuming a log link (i.e. $E(Y|X) = \exp{(X\beta)}$) the change in the expected potential outcome when M changes from 0 to m and D is held constant at d is given by:

$$\frac{E\,(Y(d,m)|D{=}d,X,L)}{E\,(Y(d,0)|D{=}d,X,L)} =$$
$$\frac{E\,(Y|D{=}d,M=m,X,L)}{E\,(Y(d,0)|D{=}d,M=0,X,L)} = \exp{(\gamma_m m)}, \tag{B.8}$$

where I again use Assumptions 2.1 and 2.2 for the first equality. Rearranging Equation B.8 gives:

$$\begin{aligned}
E\left(Y(d,0)|D{=}d,X,L\right) &= E\left(Y\exp\left(-\gamma_m m\right)|D{=}d,X,L\right)\\
E\left[E\left(Y(d,0)|D{=}d,X,L\right)|D=d,X\right] &= E\left[E\left(Y\exp\left(-\gamma_m m\right)|D{=}d,X,L\right)|D=d,X\right]\\
E\left(Y(d,0)|D{=}d,X\right) &= E\left(Y\exp\left(-\gamma_m m\right)|D{=}d,X\right)
\end{aligned} \tag{B.9}$$

Therefore $Y\exp\left(-\gamma_m m\right)$ forms the demediated outcome and I estimate the second stage by running a Poisson regression of the demediated outcome on D and X.

I apply the framework outlined above to the number of diagnoses and a binary version of SAH. Table B.1 and Table B.2 display the differences in the predicted outcomes between East and West Germans. For the number of diagnoses the estimated effects are quantitatively very similar to the estimated effects in linear models. For SAH the effects are not directly comparable since I now use a binary outcome. Only in the oldest cohort group East Germans exhibit a significantly lower probability to report a *good* or *very good* health status - an effect that is entirely driven by the male sample.

Table B.1: *Controlled direct effect estimated using poisson models: Outcome: Number of diagnoses*

Mediator: Log of net household income per household member									
	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE	0.1583***	0.0196	-0.0511**	0.1658***	0.0049	-0.0619	0.1448**	0.0304	-0.0366
	(0.0418)	(0.0294)	(0.0257)	(0.0518)	(0.0364)	(0.0387)	(0.0612)	(0.0442)	(0.0313)
TE	0.1561***	0.0559*	-0.0367	0.1594***	0.0543	-0.0517	0.1501***	0.059	-0.0237
	(0.039)	(0.0294)	(0.0232)	(0.0498)	(0.0381)	(0.0352)	(0.0551)	(0.0433)	(0.0287)
N	18868	31352	12529	9746	16636	6970	9122	14716	5559
Mediator: Unemployed									
	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE	0.2432**	0.0489*	-0.0401*	0.1553	0.0349	-0.0574*	0.3584**	0.0626	-0.0274
	(0.1056)	(0.029)	(0.0227)	(0.1284)	(0.0373)	(0.0339)	(0.1627)	(0.0428)	(0.0281)
TE	0.1998**	0.0573*	-0.0395*	0.1557	0.055	-0.052	0.2641*	0.0603	-0.0289
	(0.1002)	(0.0296)	(0.0235)	(0.126)	(0.0384)	(0.0352)	(0.148)	(0.0435)	(0.0298)
N	1726	30880	12533	888	16402	6971	838	14478	5562

Predicted average differences between East and West Germans when income and unemployment are used as mediator. The CDE is estimated using poisson models as described in Appendix B.4. TE denotes the effect of having lived in East Germany estimated on the same sample as CDE-OLS. Controls in the second stage include dummies for year of birth in 5-year categories, a linear trend in birth year and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. All standard errors are clustered at the household of orign level and obtained using the delta method.

Table B.2: *Controlled direct effect estimated using logit models: Outcome: Self-assessed health*

Mediator: Log of net household income per household member

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE	-0.0549***	0.0072	0.0102	-0.0396***	0.01	0.0122	-0.0711***	0.0043	0.0065
	(0.0093)	(0.0079)	(0.0081)	(0.0115)	(0.0096)	(0.0106)	(0.0128)	(0.0105)	(0.0118)
TE	-0.0723***	-0.0162**	-5e-04	-0.0518***	-0.0126	0.0027	-0.095***	-0.0202**	-0.004
	(0.0091)	(0.0078)	(0.0081)	(0.0112)	(0.0096)	(0.0106)	(0.0122)	(0.0103)	(0.0114)
N	93134	156215	55260	47940	82266	31224	45194	73949	24036

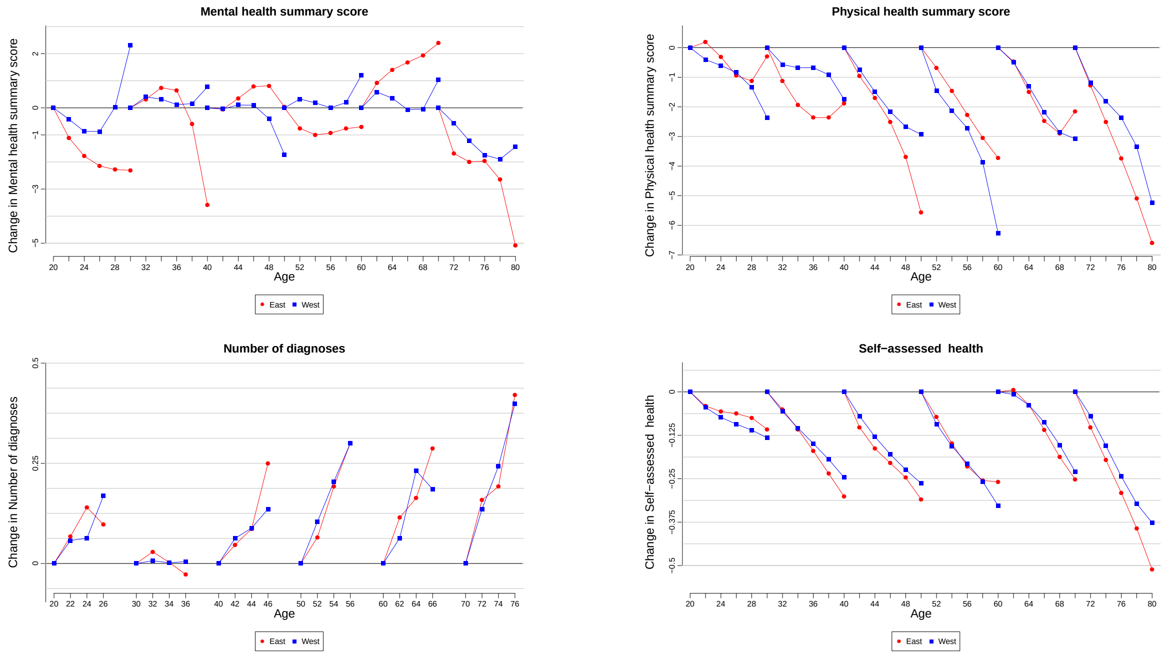
Mediator: Unemployed

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE	-0.0835***	-0.011	0.0028	-0.0567***	-0.0057	0.0062	-0.1127***	-0.0164	-7e-04
	(0.014)	(0.0079)	(0.0081)	(0.018)	(0.0098)	(0.0107)	(0.0182)	(0.0104)	(0.0114)
TE	-0.0828***	-0.0158**	-6e-04	-0.0541***	-0.0124	0.0025	-0.1136***	-0.0195*	-0.0041
	(0.0136)	(0.0078)	(0.0081)	(0.0175)	(0.0096)	(0.0106)	(0.0177)	(0.0103)	(0.0115)
N	34333	155544	55281	17374	81932	31234	16959	73612	24047

Predicted average differences between East and West Germans when income and unemployment are used as mediator. The CDE is estimated using logit models as described in Appendix B.4. TE denotes the effect of having lived in East Germany estimated on the same sample as CDE-OLS. Controls in the second stage include dummies for year of birth in 5-year categories, a linear trend in birth year and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. All standard errors are clustered at the household of orign level and obtained using the delta method.

B.5 Additional Figures and Tables

Figure B.2: *East-West differences in ageing*



Notes: Health outcomes: East-West differences in ageing. I estimate the following linear regression: $y_{normj} = \alpha + f_{east}(age) + f_{west}(age) + u$, where y_{normj} denotes a health outcome normalized to the respective base age $j = 20, 30...70$. I use B-splines of degree 3 without internal knots to approximate $f_{east}(age)$ and$f_{west}(age)$. Regression are weighted using cross sectional weights provided by the SOEP.

Table B.3: *East-West differences by cohort group: Nonlinear Models*

Number of diagnoses						
	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
East*Born 1925-1945	0.1405***	0.1304***	0.1404***	0.1152**	0.1399***	0.1509***
	(0.0385)	(0.042)	(0.0497)	(0.0539)	(0.054)	(0.0584)
East*Born 1945-1973	0.0548*	0.0453	0.0543	0.0479	0.0563	0.0443
	(0.0285)	(0.0311)	(0.037)	(0.0408)	(0.0416)	(0.0444)
East*Born 1973-1989	-0.036*	-0.0286	-0.033	-0.0143	-0.0351	-0.037
	(0.0202)	(0.0218)	(0.0312)	(0.0351)	(0.0257)	(0.025)
Observations	66839	57344	35515	30491	31324	26853
Individuals	26684	22648	14258	12113	12426	10535
p-value	0.0109	0.0637	0.1265	0.506	0.0947	0.0765

Self-assessed health						
	All		Female		Male	
	(1)	(2)	(3)	(4)	(5)	(6)
East*Born 1925-1945	-0.0529***	-0.0391***	-0.0689***	-0.0495***	-0.0354***	-0.0656***
	(0.0095)	(0.0116)	(0.0131)	(0.0108)	(0.0133)	(0.0151)
East*Born 1945-1973	0.004	-0.0032	0.0105	-1e-04	-0.0111	0.0104
	(0.009)	(0.0115)	(0.0119)	(0.0103)	(0.0134)	(0.0141)
East*Born 1973-1989	-0.0032	0.0082	-0.0139	-0.0272*	-0.0082	-0.0466**
	(0.0121)	(0.0143)	(0.0168)	(0.0143)	(0.017)	(0.0201)
Observations	367517	193448	174069	296376	157001	139375
Individuals	43147	22530	20617	31321	16581	14740
p-value	0	0.0141	0	0.0013	0.2459	3e-04
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Additional controls	No	Yes	No	Yes	No	Yes

Notes: Results from estimation of Model 2.3 using poisson regression (Number of diagnoses) and and logistic regression (SAH= *Good* or *very good*) models. Reported are the average differences in predicted outcomes between East and West separately by cohort group. Standard errors clustered on the household of origin level. Significance levels: : * p<0.10, ** p<0.05, *** p<0.01.

Table B.4: *Controlled direct effect: Mediator: Someone unemployed in household*

Mental health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-1.1126***	-0.6621**	-0.5009	-0.9243**	-1.1923***	-0.7915	-1.3479***	-0.1628	-0.2138
	(0.3036)	(0.2703)	(0.3698)	(0.3783)	(0.3617)	(0.4901)	(0.38)	(0.3465)	(0.5121)
CDE-FE	-1.0669***	-0.6826**	-0.5579	-0.8642**	-1.2799***	-0.8334*	-1.3205***	-0.1363	-0.2905
	(0.2992)	(0.2727)	(0.3731)	(0.3749)	(0.3574)	(0.4922)	(0.3755)	(0.3513)	(0.5143)
TE	-1.1366***	-0.6925***	-0.3548	-0.9579**	-1.4152***	-0.7729*	-1.3551***	-0.0161	0.0338
	(0.2972)	(0.2647)	(0.3494)	(0.3725)	(0.35)	(0.466)	(0.3714)	(0.3408)	(0.4728)
N	34639	54978	19839	17696	28789	10822	16943	26189	9017

Physical health summary score

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-1.6747***	-0.6754***	-0.2807	-1.464***	-0.5992**	-0.3104	-1.9548***	-0.7464**	-0.2472
	(0.2681)	(0.2265)	(0.2583)	(0.3554)	(0.3032)	(0.3268)	(0.3715)	(0.2983)	(0.3803)
CDE-FE	-1.6865***	-0.7795***	-0.3831	-1.4947***	-0.8582***	-0.4323	-1.9224***	-0.7128**	-0.3497
	(0.2685)	(0.2348)	(0.285)	(0.3525)	(0.3138)	(0.3308)	(0.3775)	(0.3137)	(0.4386)
TE	-1.7084***	-0.8488***	-0.4362*	-1.4863***	-0.7694**	-0.4473	-1.9827***	-0.9404***	-0.4364
	(0.263)	(0.2294)	(0.2526)	(0.3514)	(0.3039)	(0.3159)	(0.3589)	(0.3031)	(0.3711)
N	34639	54978	19839	17696	28789	10822	16943	26189	9017

Number of diagnoses

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	0.1576***	0.0414	-0.0426**	0.1597***	0.0256	-0.0616*	0.1538***	0.0572	-0.0271
	(0.0398)	(0.0293)	(0.0216)	(0.0513)	(0.0366)	(0.0334)	(0.0564)	(0.0441)	(0.0286)
CDE-FE	0.1579***	0.0615**	-0.0501**	0.1578***	0.056	-0.0515	0.1573***	0.0693	-0.0477
	(0.0395)	(0.0299)	(0.0232)	(0.0511)	(0.0381)	(0.0331)	(0.0558)	(0.0446)	(0.0323)
TE	0.1568***	0.0578*	-0.0399*	0.1602***	0.0559	-0.0512	0.1508***	0.0612	-0.0293
	(0.0394)	(0.0298)	(0.0215)	(0.051)	(0.0389)	(0.0332)	(0.0555)	(0.0435)	(0.0285)
N	18869	31355	12533	9747	16638	6971	9122	14717	5562

Self-assessed health

	All			Female			Male		
Cohort	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990	1925-1949	1949-1973	1973-1990
CDE-OLS	-0.1093***	-0.0068	-0.0244	-0.0953***	-0.021	-0.0184	-0.1275***	0.0075	-0.027
	(0.0233)	(0.0193)	(0.0267)	(0.0306)	(0.0252)	(0.0297)	(0.0311)	(0.0258)	(0.0423)
CDE-FE	-0.1066***	-0.0208	-0.0365	-0.0895***	-0.0444*	-0.0371	-0.1276***	0.001	-0.0327
	(0.0232)	(0.0197)	(0.0294)	(0.0304)	(0.0254)	(0.0303)	(0.0311)	(0.0264)	(0.0474)
TE	-0.1102***	-0.0289	-0.0372	-0.0948***	-0.0471*	-0.0308	-0.1289***	-0.0121	-0.0418
	(0.023)	(0.0195)	(0.0284)	(0.0301)	(0.0253)	(0.0292)	(0.0306)	(0.0261)	(0.0454)
N	93141	156231	55282	47946	82272	31235	45195	73959	24047

Notes: Controlled direct effect: Effect of having lived in the GDR in 1989 when no current household member is unemployed CDE-OLS denotes the estimated direct effect when the demediation function is estimated using Equation 2.12. CDE-FE denotes the estimated direct effect when the demediation function is estimated Equation 2.13. TE denotes the total effect of having lived in East Germany estimated on the same sample as CDE-OLS and CDE-FE. Controls in the second stage include dummies for age of birth in 5-year categories, a linear trend in age and sex. In the first stage I additionally control for education, stated risk preferences, marital status, household size and whether an individual has reached the age of 65. Regressions are weighted using cross sectional weights provided by the SOEP. All standard errors are clustered at the household of orign level.

Appendix C

Does the laptop always help? Non-compliance and interviewer effects in cognitive tests

C.1 Additional Figures and Tables

Figure C.1: *List of words for word recall test*

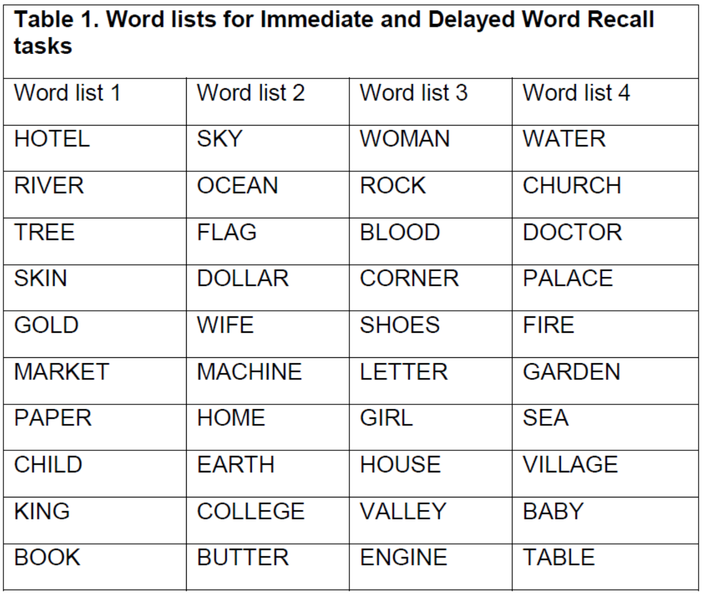
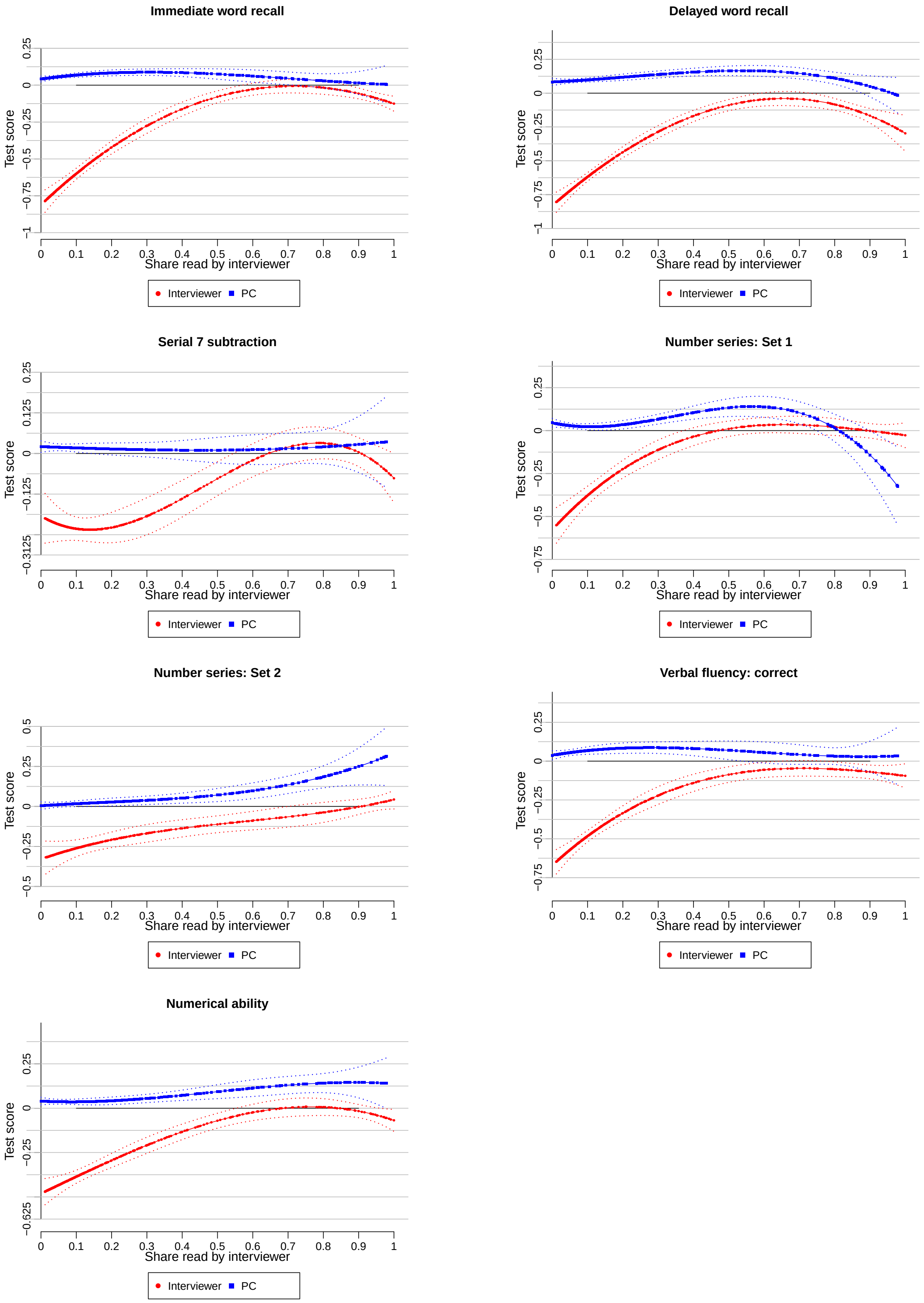


Table 1. Word lists for Immediate and Delayed Word Recall tasks

Word list 1	Word list 2	Word list 3	Word list 4
HOTEL	SKY	WOMAN	WATER
RIVER	OCEAN	ROCK	CHURCH
TREE	FLAG	BLOOD	DOCTOR
SKIN	DOLLAR	CORNER	PALACE
GOLD	WIFE	SHOES	FIRE
MARKET	MACHINE	LETTER	GARDEN
PAPER	HOME	GIRL	SEA
CHILD	EARTH	HOUSE	VILLAGE
KING	COLLEGE	VALLEY	BABY
BOOK	BUTTER	ENGINE	TABLE

Notes: List of words for word recall test. *Source:* McFall (2013, Table 1)

***Figure C.2:** Performance by share read: cross section*


Immediate word recall

Delayed word recall

Serial 7 subtraction

Number series: Set 1

Number series: Set 2

Verbal fluency: correct

Numerical ability

Notes: Average test scores in the two modes by share of interviews conducted in the alternative mode (cross section). We estimate the model $y = f_0(sread) + f_1(sread) + \epsilon$,where y is a standardized outcome, $\tilde{X}$ is a set of control variables normalized to have a mean of zero in the sample, $sread$ is the share of interviews, where the interviewer reads the words by interviewer and $f_0(sread)$, $f_1(sread)$ are approximated using cubic B-splines. The graphs plots $f_0(sread)$ and $f_1(sread)$ over $sread$.

Table C.1: *Performance differences between individuals for who the words were read by the interviewer and other respondents: Complete cases.*

	(1)	(2)	(3)	(4)
Immediate word recall				
Recall read by interviewer	-0.2484***	-0.0703***	-0.0712***	-0.0672***
	(0.0265)	(0.0216)	(0.0222)	(0.021)
N	23550	23550	23541	23548
Delayed word recall				
Recall read by interviewer	-0.3533***	-0.1813***	-0.1798***	-0.1783***
	(0.0316)	(0.0296)	(0.0298)	(0.0294)
N	23550	23550	23541	23548
Serial 7 subtraction				
Recall read by interviewer	-0.1122***	-0.0469*	-0.0449	-0.0453*
	(0.0285)	(0.0274)	(0.0276)	(0.0274)
N	23550	23550	23541	23546
Number series: Set 1				
Recall read by interviewer	-0.1305***	0.007	0.0035	0.008
	(0.0295)	(0.025)	(0.0248)	(0.0248)
N	11010	11010	11005	11009
Number series: Set 2				
Recall read by interviewer	-0.1191***	-0.0118	-0.0089	-0.011
	(0.0323)	(0.0286)	(0.0292)	(0.0284)
N	11529	11529	11526	11529
Verbal fluency: correct				
Recall read by interviewer	-0.2228***	-0.0793***	-0.082***	-0.0793***
	(0.0291)	(0.0274)	(0.0275)	(0.0275)
N	23550	23550	23541	23550
Numerical ability				
Recall read by interviewer	-0.15***	-0.0516**	-0.0507**	-0.0503**
	(0.0243)	(0.0211)	(0.0212)	(0.0212)
N	23550	23550	23541	23550
Individual controls	No	Yes	Yes	Yes
Interviewer characteristics	No	No	Yes	No
Unusual events during test	No	No	No	Yes

Notes: ATEs from inverse probability weighting models. All outcomes are standardized. Only observations without missing test scores. Individual controls include: Hearing problems, Male, Marital status (3 levels), Dummies for age in 5 year categories, Linear age trend, Educational attainment (4 levels), Born in UKUrban area, Household size, Log of monthly net income per household member, Longstanding illness or disability. Interviewer characteristics include, Interviewer: Year of birth, Year started working as interviewer. Unusal events vary by test and include at least the presence of others. For the recall tests they also include, whether the respondent had hearing problems during the test, used aides or was interrupted. Standard errors clustered at the interviewer level.

Table C.2: *Performance differences between individuals for who the words were read by the interviewer and other respondents: Missings as zero.*

		Immediate word recall		
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.289***	-0.0868***	-0.0878***	-0.0766***
	(0.0288)	(0.0229)	(0.0234)	(0.0216)
N	24323	24323	24313	24313
		Delayed word recall		
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.3813***	-0.1919***	-0.1906***	-0.1842***
	(0.0317)	(0.0291)	(0.0293)	(0.0289)
N	24323	24323	24313	24313
		Serial 7 subtraction		
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.1953***	-0.0862***	-0.0841***	-0.0444
	(0.0341)	(0.0319)	(0.0321)	(0.0274)
N	24323	24323	24313	23658
		Number series: Set 1		
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.3923***	-0.1133***	-0.1186***	-0.109***
	(0.0481)	(0.0395)	(0.0397)	(0.0396)
N	12099	12099	12092	12088
		Number series: Set 2		
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.2467***	-0.0626*	-0.0589*	-0.0573
	(0.0412)	(0.035)	(0.0357)	(0.035)
N	12224	12224	12221	12216
		Verbal fluency: correct		
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.2505***	-0.0869***	-0.0888***	-0.0943***
	(0.0305)	(0.0279)	(0.028)	(0.0278)
N	24323	24323	24313	24237
		Numerical ability		
	(1)	(2)	(3)	(4)
Recall read by interviewer	-0.1954***	-0.0698***	-0.0684***	-0.0652***
	(0.0264)	(0.0224)	(0.0226)	(0.0225)
N	24323	24323	24313	24313
Individual controls	No	Yes	Yes	Yes
Interviewer characteristics	No	No	Yes	No
Unusual events during test	No	No	No	Yes

Notes: ATEs from inverse probability weighting models. All outcomes are standardized. Missing test scores are set to zero. Individual controls include: Hearing problems, Male, Marital status (3 levels), Dummies for age in 5 year categories, Linear age trend, Educational attainment (4 levels), Born in UKUrban area, Household size, Log of monthly net income per household member, Longstanding illness or disability. Interviewer characteristics include, Interviewer: Year of birth, Year started working as interviewerUnusal events vary by test and include at least the presence of others. For the recall tests they also include, whether the respondent had hearing problems during the test, used aides or was interrupted. Standard errors clustered at the interviewer level.

Bibliography

Abowd, John M, Francis Kramarz, and David N Margolis (1999) ‘High wage workers and high wage firms’. *Econometrica* 67 (2), pp. 251–333.

Abramitzky, Ran, Adeline Delavande, and Luis Vasconcelos (2011) ‘Marrying up: The Role of Sex Ratio in Assortative Matching’. *American Economic Journal: Applied Economics* 3 (3), pp. 124–157.

Absolon, Rudolf (1960) *Wehrgesetz und Wehrdienst 1935-1945. Das Personalwesen in der Wehrmacht*. Schriften des Bundesarchivs 5. Boldt.

Acharya, Avidit, Matthew Blackwell, and Maya Sen (2016) ‘Explaining causal findings without bias: Detecting and assessing direct effects’. *The American Political Science Review* 110 (3), p. 512.

Al Baghal, Tarek (2017) ‘The Effect of Online and Mixed-Mode Measurement of Cognitive Ability’. *Social Science Computer Review*, p. 0894439317746328.

Alesina, Alberto and Nicola Fuchs-Schündeln (2007) ‘Goodbye Lenin (or Not?): The Effect of Communism on People’. *American Economic Review* 97 (4), pp. 1507–1528.

Alesina, Alberto and Paola Giuliano (2015) ‘Culture and institutions’. *Journal of Economic Literature* 53 (4), pp. 898–944.

Almond, Douglas and Janet Currie (2011) ‘Killing Me Softly: The Fetal Origins Hypothesis’. *Journal of Economic Perspectives* 25 (3), pp. 153–72.

Alwin, Duane F (2007) *Margins of error: A study of reliability in survey measurement*. Vol. 547. John Wiley & Sons.

Andersen, Hanfried H, Axel Mühlbacher, Matthias Nübling, Jürgen Schupp, and Gert G Wagner (2007) ‘Computation of standard values for physical and mental health scale scores using the SOEP version of SF-12v2’. *Schmollers Jahrbuch* 127 (1), pp. 171–182.

Andrews, Martyn J, Len Gill, Thorsten Schank, and Richard Upward (2008) ‘High wage workers and low wage firms: negative assortative matching or limited mobility bias?’ *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (3), pp. 673–697.

Angrist, Joshua D and Jörn-Steffen Pischke (2009) *Mostly harmless econometrics: An empiricist’s companion*. Princetion, NJ: Princeton University Press.

Ärztekammer und KVD, Landesstelle Bayern (1939) ‘Beschleunigte Bestallung von Medizinalpraktikanten als Ärzte’. *Ärzteblatt für Bayern* 6 (18), p. 387.

Atella, Vincenzo, Edoardo Di Porto, and Joanna Kopinska (2016) *Stress, Famine and The Fetal Programming: The Long Term Effect of WWII in Italy*. Tor Vergata University, CEIS Research Paper Series Vol. 14, Issue 9, No. 385.

Athey, Susan and Guido Imbens (2016) 'Recursive partitioning for heterogeneous causal effects'. *Proceedings of the National Academy of Sciences* 113 (27), pp. 7353–7360.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015) 'Fitting Linear Mixed-Effects Models Using lme4'. *Journal of Statistical Software* 67 (1), pp. 1–48.

Bauer, Michael, Stefan Priebe, Bettina Blaring, and Kerstin Adamczak (1993) 'Long-term mental sequelae of political imprisonment in East Germany.' *The Journal of Nervous and Mental disease* 181 (4), pp. 257–262.

Bayerisches Statistisches Landesamt, ed. (1937-1942) *Zeitschrift des Bayerischen Statistischen Landesamts*.

Behrman, Jere R and Mark R Rosenzweig (2004) 'Returns to Birthweight'. *Review of Economics and Statistics* 86 (2), pp. 586–601.

Billiet, Jacques and Geert Loosveldt (1988) 'Improvement of the quality of responses to factual survey questions by interviewer training'. *Public Opinion Quarterly* 52 (2), pp. 190–211.

Boreham, Richard, Diana Boldysevaite, and Caroline Killpack (2012) 'UKHLS: Wave 1 technical report'. *London: NatCen*.

Börsch-Supan, Axel, Martina Brandt, Christian Hunkler, Thorsten Kneip, Julie Korbmacher, Frederic Malter, Barbara Schaan, Stephanie Stuck, and Sabrina Zuber (2013) 'Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE)'. *International journal of epidemiology* 42 (4), pp. 992–1001.

Bouvier, Beatrix (1999) 'Verfolgung und Repression in der SBZ/DDR von den vierziger bis zu den sechziger Jahren und ihre Wahrnehmung in Ost und West'. In: *Politische Repression in der SBZ, DDR und ihre Wahrnehmung in der Bundesrepublik Vorträge einer Sektion auf dem Berliner "Geschichtsforum 1949 - 1989 - 1999. Getrennte Vergangenheit - Gemeinsame Geschichte?", am 29. Mai 1999*. Gesprächskreis Geschichte. Bonn: Historisches Forschungszentrum.

Bozzoli, Carlos and Climent Quintana-Domeque (2014) 'The Weight of the Crisis: Evidence From Newborns in Argentina'. *The Review of Economics and Statistics* 96 (3), pp. 550–562.

Brunton-Smith, Ian, Patrick Sturgis, and George Leckie (2017) 'Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180 (2), pp. 551–568.

Brunton-Smith, Ian, Patrick Sturgis, and Joel Williams (2012) 'Is success in obtaining contact and cooperation correlated with the magnitude of interviewer variance?' *Public Opinion Quarterly*, nfr067.

Busse, Reinhard and Annette Riesberg (2004) 'Health care systems in transition: Germany'.

Carlson, Kyle (2015) ‘Fear itself: The effects of distressing economic news on birth outcomes’. *Journal of Health Economics* 41, pp. 117–132.

Case, Anne and Angus Deaton (2017) ‘Mortality and morbidity in the 21st century’. *Brookings Papers on Economic Activity* 2017, p. 397.

Cawley, John and Christopher J. Ruhm (2011) ‘Chapter Three - The Economics of Risky Health Behaviors1’. In: *Handbook of Health Economics*. Ed. by Thomas G. Mcguire Mark V. Pauly and Pedro P. Barros. Vol. 2. Handbook of Health Economics. Elsevier, pp. 95–199.

Celidoni, Martina, Chiara Dal Bianco, and Guglielmo Weber (2017) ‘Retirement and cognitive decline. A longitudinal analysis using SHARE data’. *Journal of Health Economics* 56, pp. 113–125.

Cernat, Alexandru, Mick P. Couper, and Mary Beth Ofstedal (2016) ‘Estimation of Mode Effects in the Health and Retirement Study Using Measurement Models’. *Journal of Survey Statistics and Methodology* 4 (4), pp. 501–524.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler (2016) ‘The association between income and life expectancy in the United States, 2001-2014’. *Jama* 315 (16), pp. 1750–1766.

Christians, Annemone (2013) *Amtsgewalt und Volksgesundheit: das öffentliche Gesundheitswesen im nationalsozialistischen München*. München im Nationalsozialismus. Wallstein.

Cobb-Clark, Deborah A and Stefanie Schurer (2013) ‘Two economists’ musings on the stability of locus of control’. *The Economic Journal* 123 (570), F358–F400.

Crossley, Thomas F., Tobias Schmidt, Panagiota Tzamourani, and Joachim K. Winter (2017) *Interviewer effects and the measurement of financial literacy*. eng. ISER Working Paper Series 2017-06. Colchester.

Crossley, Thomas and Steven Kennedy (2002) ‘The reliability of self-assessed health status’. *Journal of Health Economics* 21 (4), pp. 643–658.

Currie, Janet (2009) ‘Healthy, wealthy, and wise: Is there a causal relationship between child health and human capital development?’ *Journal of Economic Literature* 47 (1), pp. 87–122.

Currie, Janet and Maya Rossin-Slater (2013) ‘Weathering the storm: Hurricanes and birth outcomes’. *Journal of Health Economics* 32 (3), pp. 487–503.

Cutler, David, Angus Deaton, and Adriana Lleras-Muney (2006) ‘The determinants of mortality’. *The Journal of Economic Perspectives* 20 (3), pp. 97–120.

Czaja, Sara J, Neil Charness, Arthur D Fisk, Christopher Hertzog, Sankaran N Nair, Wendy A Rogers, and Joseph Sharit (2006) ‘Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (CREATE).’ *Psychology and aging* 21 (2), p. 333.

Deaton, Angus (2003) ‘Health, inequality, and economic development’. *Journal of Economic Literature* 41 (1), pp. 113–158.

Diederichs, Claudia and Hannelore Neuhauser (2014) 'Regional variations in hypertension prevalence and management in Germany: results from the German Health Interview and Examination Survey (DEGS1)'. *Journal of Hypertension* 32 (7), pp. 1405–1414.

Dragone, Davide and Nicolas Ziebarth (2017) 'Non-separable time preferences, novelty consumption and body weight: Theory and evidence from the East German transition to capitalism'. *Journal of Health Economics* 51, pp. 41–65.

Eckart, Wolfgang U, Volker Sellin, and Eike Wolgast (2006) *Die Universität Heidelberg im Nationalsozialismus*. Springer.

Eibich, Peter and Nicolas Ziebarth (2014) 'Examining the structure of spatial health effects in Germany using Hierarchical Bayes Models'. *Regional Science and Urban Economics* 49, pp. 305–320.

Evans, Richard J. (2004) *Das dritte Reich*. Dt. Verl.-Anst.

Fallwell, Lynne (2013) *Modern german midwifery: 1885 - 1960*. Studies for the Society for the Social History of Medicine 13. Pickering & Chatto.

Farbmacher, Helmut, Raphael Guber, and Johan Vikström (2016) 'Increasing the credibility of the twin birth instrument'. *Journal of Applied Econometrics*.

Farré, Lıdia, Francesco Fasani, and Hannes Felix Mueller (2015) *Feeling useless: the effect of unemployment on mental health in the Great Recession*. Tech. rep. IZA Discussion Paper.

Floris, Joel, Kaspar Staub, and Ulrich Woitek (2016) *The Benefits of Intervention: Birth Weights in Basle 1912-1920*. University of Zurich, Department of Economics, Working Paper Series No. 236.

Forouzanfar, Mohammad H, Ashkan Afshin, Lily T Alexander, H Ross Anderson, Zulfiqar A Bhutta, Stan Biryukov, Michael Brauer, Richard Burnett, Kelly Cercy, Fiona J Charlson, et al. (2016) 'Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015'. *The Lancet* 388 (10053), pp. 1659–1724.

Frei, Norbert (2013) *Der Führerstaat. Nationalsozialistische Herrschaft 1933 bis 1945*. Beck.

Friehe, Tim, Markus Pannenberg, and Michael Wedow (2015) *Let Bygones be Bygones? Political Regimes and Personalities in Germany*. Tech. rep.

Frijters, Paul, John Haisken-DeNew, and Michael Shields (2004) 'Money Does Matter! Evidence from Increasing Real Income and Life Satisfaction in East Germany Following Reunification'. *American Economic Review* 94 (3), pp. 730–740.

Fuchs-Schündeln, Nicola and Paolo Masella (2016) 'Long-lasting effects of socialist education'. *Review of Economics and Statistics* 98 (3), pp. 428–441.

Fuchs-Schündeln, Nicola and Matthias Schündeln (2005) 'Precautionary savings and Self-Selection: Evidence from the german reunification "Experiment"'. *The Quarterly Journal of Economics* 120 (3), pp. 1085–1120.

Gandek, Barbara, John E Ware, Neil K Aaronson, Giovanni Apolone, Jakob B Bjorner, John E Brazier, Monika Bullinger, Stein Kaasa, Alain Leplege, Luis Prieto, et al. (1998) 'Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project'. *Journal of Clinical Epidemiology* 51 (11), pp. 1171–1178.

Gerstorf, Sandra and Jürgen Schupp (2016) *SOEP wave report 2015*. Tech. rep. SOEP Wave Report.

Gjonça, Arjan, Hilke Brockmann, and Heiner Maier (2000) 'Old-age mortality in Germany prior to and after reunification'. *Demographic Research* 3.

Glymour, M Maria, Ichiro Kawachi, Christopher S Jencks, and Lisa F Berkman (2008) 'Does childhood schooling affect old age memory or mental status? Using state schooling laws as natural experiments'. *Journal of Epidemiology & Community Health* 62 (6), pp. 532–537.

Goetgeluk, Sylvie, Stijn Vansteelandt, and Els Goetghebeur (2008) 'Estimation of controlled direct effects'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (5), pp. 1049–1066.

Gooch, Andrew (2015) 'Measurements of cognitive skill by survey mode: Marginal differences and scaling similarities'. *Research & Politics* 2 (3), p. 2053168015590681.

Görges, Luise and Miriam Beblo (2015) 'Breaking down the wall between nature and nurture: An exploration of gendered work preferences in East and West Germany'.

Greene, William (2012) *Econometric analysis*. 7th edition. Pearson Education Limited.

Grossman, Michael (1972) 'On the concept of health capital and the demand for health'. *Journal of Political Economy* 80 (2), pp. 223–255.

Groves, Robert M. (2004) *Survey errors and survey costs*. Wiley series in survey methodology. Hoboken, NJ: Wiley-Interscience.

Gustavsson, A, J Bjorkman, C Ljungcrantz, Annika Rhodin, M Rivano-Fischer, K-F Sjolund, and C Mannheimer (2012) 'Socio-economic burden of patients with a diagnosis related to chronic pain–Register data of 840,000 Swedish patients'. *European journal of pain* 16 (2), pp. 289–299.

Hansen, Bruce E (2001) 'The new econometrics of structural change: Dating breaks in US labor productivity'. *The Journal of Economic Perspectives* 15 (4), pp. 117–128.

Heckman, James J (1979) 'Sample Selection Bias as a Specification Error'. *Econometrica* 47 (1), pp. 153–161.

Heckman, James J, Jora Stixrud, and Sergio Urzua (2006) 'The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior'. *Journal of Labor economics* 24 (3), pp. 411–482.

Heidemann, Christin, Ronny Kuhnert, Sabine Born, and Christa Scheidt-Nave (2017) '12-Month prevalence of known diabetes mellitus in Germany'. *Journal of Health Monitoring* 2 (1), pp. 43–50.

Heineck, Guido and Silke Anger (2010) 'The returns to cognitive abilities and personality traits in Germany'. *Labour Economics* 17 (3), pp. 535–546.

Herzog, Regula and Rodgers (1999) 'Cognition, aging, and self-reports'. In: ed. by Willard L, N Schwarz, DC Park, B Knauper, and S Sudman. Psychology Press, Philadelphia. Chap. Cognitive performance measures in survey research on older adults, pp. 327–340.

Hirano, Keisuke, Guido Imbens, and Geert Ridder (2003) 'Efficient estimation of average treatment effects using the estimated propensity score'. *Econometrica* 71 (4), pp. 1161–1189.

Hou, Jinghui, Yijie Wu, and Erin Harrell (2017) 'Reading on Paper and Screen among Senior Adults: Cognitive Map and Technophobia'. *Frontiers in Psychology* 8, p. 2225.

Huber, Martin, Michael Lechner, and Giovanni Mellace (2016) 'The finite sample performance of estimators for mediation analysis under sequential conditional independence'. *Journal of Business & Economic Statistics* 34 (1), pp. 139–160.

Huber, Martin, Michael Lechner, and Anthony Strittmatter (2017) 'Direct and indirect effects of training vouchers for the unemployed'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Humphreys, Brad R, Logan McLeod, and Jane E Ruseski (2014) 'Physical activity and health outcomes: evidence from Canada'. *Health Economics* 23 (1), pp. 33–54.

Imai, Kosuke and In Song Kim (2016) *When should we use linear fixed effects regression models for causal inference with panel data*. Tech. rep. Princeton University. Mimeo.

Imbens, Guido (2004) 'Nonparametric estimation of average treatment effects under exogeneity: A review'. *Review of Economics and Statistics* 86 (1), pp. 4–29.

Jäckle, Annette, Caroline Roberts, and Peter Lynn (2010) 'Assessing the effect of data collection mode on measurement'. *International Statistical Review* 78 (1), pp. 3–20.

Jochmans, Koen and Martin Weidner (2017) 'Fixed-effect regressions on network data'. *arXiv preprint arXiv:1608.01532v2*.

Joffe, Marshall and Tom Greene (2009) 'Related causal frameworks for surrogate outcomes'. *Biometrics* 65 (2), pp. 530–538.

Jürges, Hendrik (2013) 'Collateral damage: The German food crisis, educational attainment and labor market outcomes of German post-war cohorts'. *Journal of Health Economics* 32 (1), pp. 286–303.

Jylhä, Marja (2009) 'What is self-rated health and why does it predict mortality? Towards a unified conceptual model'. *Social Science & Medicine* 69 (3), pp. 307–316.

Kawachi, Ichiro, Nancy E Adler, and William H Dow (2010) 'Money, schooling, and health: Mechanisms and causal evidence'. *Annals of the New York Academy of Sciences* 1186 (1), pp. 56–68.

Kesternich, Iris, Bettina Siflinger, James Smith, and Joachim Winter (2014) 'The Effects of World War II on Economic and Health Outcomes across Europe'. *The Review of Economics and Statistics* 96 (1), pp. 103–118.

Kesternich, Iris, Bettina Siflinger, James Smith, and Joachim Winter (2015) 'Individual Behavior as a Pathway between Early-Life Shocks and Adult Health: Evidence from Hunger Episodes in Post-War Germany'. *The Economic Journal* 125 (588), F372–F393.

Klärner, Andreas (2015) 'The low importance of marriage in eastern Germany-social norms and the role of peoples' perceptions of the past'. *Demographic Research* 33, p. 239.

Knies, Gundi (2016) *Understanding Society: Waves 1-7, 2009-2016 and harmonised British Household Panel Survey: Waves 1-18, 1991-2009, UserGuide*. Tech. rep.

König (1939) 'An die Bayerische Ärzteschaft'. *Ärzteblatt für Bayern* 6 (18), pp. 385–393.

Krämer, U., R. Schmitz, J. Ring, and H. Behrendt (2015) 'What can reunification of East and West Germany tell us about the cause of the allergy epidemic?' *Clinical Experimental Allergy* 45 (1), pp. 94–107.

Kroh, Martin (2009) *Short-Documentation of the Update of the SOEP-Weights, 1984–2008*. Tech. rep.

Kroll, Lars E and Thomas Lampert (2010) 'Regionale Unterschiede in der Gesundheit am Beispiel von Adipositas und Diabetes mellitus'. *Robert Koch-Institut, editor. Daten und Fakten: Ergebnisse der Studie» Gesundheit in Deutschland aktuell*, pp. 51–59.

Kyriazidou, Ekaterini (1997) 'Estimation of a Panel Data Sample Selection Model'. *Econometrica* 65 (6), pp. 1335–1364.

Lang, Frieder R, David Weiss, Andreas Stocker, Bernhard von Rosenbladt, et al. (2007) 'Assessing cognitive capacities in computer-assisted survey research: Two ultra-short tests of intellectual ability in the German Socio-Economic Panel (SOEP)'. *Schmollers Jahrbuch* 127 (1), pp. 183–192.

Lee, I-Min, Eric J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair, Peter T Katzmarzyk, Lancet Physical Activity Series Working Group, et al. (2012) 'Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy'. *The lancet* 380 (9838), pp. 219–229.

Liebert, Helge and Beatrice Mäder (2016) *Marginal effects of physician coverage on infant and disease mortality*. University of St.Gallen, School of Economics and Political Science, Department of Economics, Discussion Paper No. 2016-20.

Lindeboom, Maarten and Reyn Van Ewijk (2015) 'Babies of the War: The effect of war exposure early in life on mortality throughout life'. *Biodemography and Social Biology* 61 (2), pp. 167–186.

Lippe, Elena von der, Angela Fehr, and Cornelia Lange (2017) 'Limitations to usual activities due to health problems in German'. *Journal of Health Monitoring* 2 (3).

Lleras-Muney, Adriana (forthcoming) 'Mind the Gap: a review of The Health Gap by Sir Michael Marmot'. *Journal of Economic Perspectives*.

Luechinger, Simon (2009) 'Valuing Air Quality Using the Life Satisfaction Approach'. *The Economic Journal* 119 (536), pp. 482–515.

Lumey, Lambert H, Aryeh D Stein, and Ezra Susser (2011) 'Prenatal famine and adult health'. *Annual Review of Public Health* 32, pp. 237–262.

Luy, Marc (2004) 'Mortality differences between Western and Eastern Germany before and after Reunification: A macro and micro level analysis of developments and responsible factors'. *Genus* 60 (3/4), pp. 99–141.

MacKenzie (2006) 'High-Risk Pregnancy: Management Options'. In: ed. by D James, P Steer, C Weiner, and B Gonik. Philadelphia: Saunders Elsevier. Chap. Unstable lie, malpresentations, and malpositions, pp. 359–1375.

Marcus, Jan (2013) 'The effect of unemployment on the mental health of spouses–Evidence from plant closures in Germany'. *Journal of Health Economics* 32 (3), pp. 546–558.

Marmot, Michael (2015) *The Health Gap: The Challenge of an Unequal World*. Bloomsbury Publishing.

Marmot, Michael, Stephen Stansfeld, Chandra Patel, Fiona North, Jenny Head, Ian White, Eric Brunner, Amanda Feeney, and G Davey Smith (1991) 'Health inequalities among British civil servants: the Whitehall II study'. *The Lancet* 337 (8754), pp. 1387–1393.

Marmot, Michael and R. Wilkinson (2005) *Social determinants of health*. OUP Oxford.

Mayer, Wolfgang (2002) *Flucht und Ausreise: Botschaftsbesetzungen als wirksame Form des Widerstands und Mittel gegen die politische Verfolgung in der DDR*. Anita Tykve Verlag.

McArdle, John J, Gwenith G Fisher, and Kelly M Kadlec (2007) 'Latent variable analyses of age trends of cognition in the Health and Retirement Study, 1992-2004.' *Psychology and aging* 22 (3), p. 525.

McFall, Stephanie (2013) *Understanding Society: UK Household Longitudinal Study: Cognitive Ability Measures*. Tech. rep. Version 1.1. Institute for Social and Economic Research University of Essex.

McIntire, Donald D., Steven L. Bloom, Brian M. Casey, and Kenneth J. Leveno (1999) 'Birth Weight in Relation to Morbidity and Mortality among Newborn Infants'. *New England Journal of Medicine* 340 (16), pp. 1234–1238.

Miller, Ute (1964) *Zur Geschichte der Münchner Kinderkrankenhäuser*. Dissertation. University of Munich.

Monjok, Emmanuel, Ita B Okokon, Margaret M Opiah, Justin A Ingwu, John E Ekabua, and Ekere J Essien (2012) 'Obstructed labour in resource-poor settings: the need for revival of symphysiotomy in Nigeria'. *African Journal of Reproductive Health* 16 (3), pp. 93–100.

Müller-Nordhorn, J, K Rossnagel, W Mey, and S N Willich (2004) 'Regional variation and time trends in mortality from ischaemic heart disease: East and West Germany 10 years after reunification'. *Journal of Epidemiology & Community Health* 58 (6), pp. 481–485.

Müller-Nordhorn, Jacqueline, Sylvia Binting, Stephanie Roll, and Stefan N. Willich (2008) 'An update on regional variation in cardiovascular mortality within Europe'. *European Heart Journal* 29 (10), pp. 1316–1326.

Mylonas, Ioannis and Klaus Friese (2015) 'Indications for and Risks of Elective Cesarean Section'. *Deutsches Ärzteblatt International* 112 (29-30), p. 489.

Neuhauser, Hannelore, Ronny Kuhnert, and Sabine Born (2017) '12-Month prevalence of hypertension in Germany'. *Journal of Health Monitoring* 2 (1), pp. 51–57.

Newey, Whitney K and Daniel McFadden (1994) ‘Large sample estimation and hypothesis testing’. *Handbook of Econometrics* 4, pp. 2111–2245.

Nilsen, Thomas S, Ingunn Brandt, Per Magnus, and Jennifer R Harris (2012) ‘The Norwegian twin registry’. *Twin Research and Human Genetics* 15 (6), pp. 775–780.

Nolte, Ellen, Rembrandt Scholz, Vladimir Shkolnikov, and Martin McKee (2002) ‘The contribution of medical care to changing life expectancy in Germany and Poland’. *Social Science & Medicine* 55 (11), pp. 1905–1921.

Nolte, Ellen, Vladimir Shkolnikov, and Martin McKee (2000) ‘Changing mortality patterns in East and West Germany and Poland. II: Short-term trends during transition and in the 1990s’. *Journal of Epidemiology & Community Health* 54 (12), pp. 899–906.

OECD (2017) *Dataset: Level of GDP per capita and productivity*. URL: `http://stats.oecd.org//Index.aspx?QueryId=54369` (visited on 12/16/2017).

Overmanns, Rüdiger (2009) *Deutsche militärische Verluste im Zweiten Weltkrieg*. 3rd ed. Beiträge zur Militärgeschichte 46. Oldebourg.

Pearl, Judea (2001) ‘Direct and indirect effects’. In: *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 411–420.

Permooser, Irmtraud (1997) *Der Luftkrieg über München : 1942 - 1945 ; Bomben auf die Hauptstadt der Bewegung*. Zugl.: München, Univ., Diss., 1991. Oberhaching: Aviatic-Verl.

Quintana-Domeque, Climent and Pedro Ródenas-Serrano (2017) ‘The hidden costs of terrorism: The effects on health at birth’. *Journal of Health Economics* 56, pp. 47–60.

Reichsminister des Inneren (1939) ‘Runderlass, betr. Hausentbindungen und Anstaltsentbindungen vom 6. September 1939’. *Reichsgesundheitsblatt* 14 (63), p. 873.

Richter, Hedwig (2009) *Die DDR*. UTB S. Stuttgart: UTB GmbH.

Robins, James M (1994) ‘Correcting for non-compliance in randomized trials using structural nested mean models’. *Communications in Statistics-Theory and methods* 23 (8), pp. 2379–2412.

Robins, James M and Sander Greenland (1992) ‘Identifiability and exchangeability for direct and indirect effects’. *Epidemiology*, pp. 143–155.

Rohwedder, Susann and Robert J Willis (2010) ‘Mental retirement’. *Journal of Economic Perspectives* 24 (1), pp. 119–38.

Rosen, Larry D. and Michelle M. Weil (1995) ‘Computer availability, computer experience and technophobia among public school teachers’. *Computers in Human Behavior* 11 (1), pp. 9–31.

Rubin, Donald B (1974) ‘Estimating causal effects of treatments in randomized and nonrandomized studies.’ *Journal of Educational Psychology* 66 (5), p. 688.

Sanchez, Christopher A, Jennifer Wiley, Timothy K Miura, Gregory JH Colflesh, Travis R Ricks, Melinda S Jensen, and Andrew RA Conway (2010) ‘Assessing working memory capacity in a non-native language’. *Learning and Individual Differences* 20 (5), pp. 488–493.

Schilling, Mark F (1990) ‘The longest run of heads’. *College Math. J* 21 (3), pp. 196–207.

Schipf, S., A. Werner, T. Tamayo, R. Holle, M. Schunk, W. Maier, C. Meisinger, B. Thorand, K. Berger, G. Mueller, S. Moebus, B. Bokhof, A. Kluttig, K. H. Greiser, H. Neuhauser, U. Ellert, A. Icks, W. Rathmann, and H. Völzke (2012) ‘Regional differences in the prevalence of known Type 2 diabetes mellitus in 45–74 years old individuals: Results from six population-based studies in Germany (DIAB-CORE Consortium)’. *Diabetic Medicine* 29 (7), e88–e95.

Schnell, Rainer and Frauke Kreuter (2005) ‘Separating Interviewer and Sampling-Point Effects’. *Journal of Official Statistics* 21 (3), p. 389.

Schober, Michael F. and Frederick G. Conrad (1997) ‘Does Conversational Interviewing Reduce Survey Measurement Error?’ *The Public Opinion Quarterly* 61 (4), pp. 576–602.

Schumacher, Reto and Luigi Lorenzetti (2005) ‘“We Have No Proletariat”: Social Stratification and Occupational Homogamy in Industrial Switzerland, Winterthur 1909/10–1928’. *International Review of Social History* null (Supplement S13), pp. 65–91.

Schützwohl, Matthias and Andreas Maercker (2000) ‘Anger in former East German political prisoners: Relationship to posttraumatic stress reactions and social support’. *The Journal of Nervous and Mental Disease* 188 (8), pp. 483–489.

Semykina, Anastasia and Jeffrey M Wooldridge (2010) ‘Estimating panel data models in the presence of endogeneity and selection’. *Journal of Econometrics* 157 (2), pp. 375–380.

Sensch, Jürgen (2006) *histat-Datenkompilation online: Grunddaten zur historischen Entwicklung des Gesundheitswesens in Deutschland von 1876 bis 1999.* Data from: GESIS Datenarchiv, Köln. histat. Study number 8209.

Smith, James P., John J. McArdle, and Robert Willis (2010) ‘Financial Decision Making and Cognition in a Family Context*’. *The Economic Journal* 120 (548), F363–F380.

SOEP 2012 - Erhebungsinstrumente 2012 (Welle 29) des Sozio-oekonomischen Panels: Personenfragebogen, Altstichproben (2013). ger. SOEP Survey Papers 157. Berlin.

Sonnega, Amanda, Jessica D Faul, Mary Beth Ofstedal, Kenneth M Langa, John WR Phillips, and David R Weir (2014) ‘Cohort profile: The health and retirement study (HRS)’. *International journal of epidemiology* 43 (2), pp. 576–585.

Spitzer, Carsten, Ines Ulrich, Kathryn Plock, Jörn Mothes, Anne Drescher, Lena Gürtler, Harald J Freyberger, and Sven Barnow (2007) ‘Beobachtet, verfolgt, zersetzt-psychische Erkrankungen bei Betroffenen nichtstrafrechtlicher Repressionen in der ehemaligen DDR’. *Psychiatrische Praxis* 34 (02), pp. 81–86.

Stadtarchiv München, ed. (1939-1940) *Kriegswirtschaftsberichte*.

Statistisches Reichsamt, ed. (1933-1940) *Statistisches Jahrbuch für das Deutsche Reich*.

Stauber, M. (2012) ‘Herausforderungen – 100 Jahre Bayerische Gesellschaft für Geburtshilfe und Frauenheilkunde’. In: ed. by C. Anthuber, M.W. Beckmann, J. Dietl, F. Dross, and W. Frobenius. Thieme. Chap. Vergangenheitsbewältigung in der bayerischen

Gynäkologie – Erfahrungen an der I. Universitätsfrauenklinik München, pp. 237–256.

Stein, Aryeh D, Patricia A Zybert, Margot Van de Bor, and LH Lumey (2004) 'Intrauterine famine exposure and body proportions at birth: the Dutch Hunger Winter'. *International Journal of Epidemiology* 33 (4), pp. 831–836.

Steppuhn, Henriette, Ronny Kuhnert, and Christa Scheidt-Nave (2017) '12-month prevalence of asthma among adults in Germany'. *Journal of Health Monitoring* 2 (3).

Süß, Winfried (2003) *Der "Volkskörper" im Krieg. Gesundheitspolitik, Gesundheitsverhältnisse und Krankenmord im nationalsozialistischen Deutschland 1939-1945*. Oldenbourg.

Thom, Julia, Ronny Kuhnert, Sabine Born, and Ulfert Hapke (2017) '12-month prevalence of self-reported medical diagnoses of depression in Germany'. *Journal of Health Monitoring* 2 (3), pp. 68–76.

Torche, Florencia (2011) 'The Effect of Maternal Stress on Birth Outcomes: Exploiting a Natural Experiment'. *Demography* 48 (4), pp. 1473–1491.

University of Essex (Nov. 2016) *Understanding Society: Waves 1-6, 2009-2015 [computer file]*. SN: 6614. Colchester, Essex: Institute for Social, Economic Research, NatCen Social Research, and Kantar Public, [producers].

US Department of Health and Human Services et al. (2014) 'The health consequences of smoking—50 years of progress: a report of the Surgeon General'. *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health* 17.

Usborne, Cornelie (2011) 'Social Body, Racial Body, Woman's Body. Discourses, Policies, Practices from Wilhelmine to Nazi Germany, 1912-1945'. *Historical Social Research* 36 (2), pp. 140–161.

Uysal, S. Derya (2015) 'Doubly Robust Estimation of Causal Effects with Multivalued Treatments: An Application to the Returns to Schooling'. *Journal of Applied Econometrics* 30 (5), pp. 763–786.

Valeri, Linda and Tyler VanderWeele (2013) 'Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros.' *Psychological Methods* 18 (2), p. 137.

Van den Berg, Gerard J., Pia R. Pinger, and Johannes Schoch (2016) 'Instrumental Variable Estimation of the Causal Effect of Hunger Early in Life on Health Later in Life'. *Economic Journal* 126 (591), pp. 65–506.

Van Ewijk, Reyn and Maarten Lindeboom (2016) *Why People Born During World War II are Healthier*. Gutenberg School of Management and Economics & Research Unit "Interdisciplinary Public Policy", Discussion Paper Series No. 1619.

Van Hooren, SAH, AM Valentijn, H Bosma, RWHM Ponds, MPJ Van Boxtel, and J Jolles (2007) 'Cognitive functioning in healthy older adults aged 64–81: a cohort study into the effects of age, sex, and education'. *Aging, Neuropsychology, and Cognition* 14 (1), pp. 40–54.

Van Leeuwen, Marco HD and Ineke Maas (2011) *HISCLASS: A historical international social class scheme*. Leuven University Press.

Van Leeuwen, Marco HD, Ineke Maas, and Andrew Miles (2002) *HISCO: Historical international standard classification of occupations*. Leuven University Press.

VanderWeele, Tyler and Stijn Vansteelandt (2009) 'Conceptual issues concerning mediation, interventions and composition'. *Statistics and its Interface* 2 (4), pp. 457–468.

Vansteelandt, Stijn (2009) 'Estimating direct effects in cohort and case–control studies'. *Epidemiology* 20 (6), pp. 851–860.

Vansteelandt, Stijn and Marshall Joffe (2014) 'Structural nested models and g-estimation: The partially realized promise'. *Statistical Science* 29 (4), pp. 707–731.

Vella, Francis and Marno Verbeek (1999) 'Two-step estimation of panel data models with censored endogenous variables and selection bias'. *Journal of Econometrics* 90 (2), pp. 239–263.

Verdier, Valentin (2017) *Estimation and inference for linear models with two-way fixed effects and sparsely matched data*.

Verhaeghen, Paul and Timothy A Salthouse (1997) 'Meta-analyses of age–cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models.' *Psychological bulletin* 122 (3), p. 231.

Vogt, Tobias C (2013) 'How many years of life did the fall of the Berlin Wall add? A projection of East German life expectancy'. *Gerontology* 59 (3), pp. 276–282.

Wagner, Gert, Joachim Frick, and Jürgen Schupp (2007) 'The German Socio-Economic Panel study (SOEP)-evolution, scope and enhancements'. *Schmollers Jahrbuch* 127, pp. 139–169.

Walter, S. D. (1992) 'The Analysis of Regional Patterns in Health DataI.Distributional Considerations'. *American Journal of Epidemiology* 136 (6), pp. 730–741.

Weber, Hermann (2011) *Die DDR 1945-1990*. Berlin, Boston: De Gruyter.

West, Brady T and Kristen Olson (2010) 'How much of interviewer variance is really nonresponse error variance?' *Public Opinion Quarterly* 74 (5), pp. 1004–1026.

Whitley, Elise, Ian J. Deary, Stuart J. Ritchie, G. David Batty, Meena Kumari, and Michaela Benzeval (2016) 'Variations in cognitive abilities across the life course: Cross-sectional evidence from Understanding Society: The UK Household Longitudinal Study'. *Intelligence* 59, pp. 39–50.

WHO (2016) *World Health Statistics 2016: Monitoring Health for the SDGs Sustainable Development Goals*. World Health Organization.

WHO (2017) *Noncommunicable diseases: WHO factsheet*. URL: `http://www.who.int/mediacentre/factsheets/fs355/en/` (visited on 03/08/2018).

WHO (2018) *Prevalence of tobacco smoking*. URL: `http://www.who.int/gho/tobacco/use/en/` (visited on 03/08/2018).

Wooden, Mark (2013) *The measurement of cognitive ability in wave 12 of the hilda survey*. Melbourne Institute Working Paper 13.

Wooldridge, Jeffrey M (1995) ‘Selection corrections for panel data models under conditional mean independence assumptions’. *Journal of Econometrics* 68 (1), pp. 115–132.

World Health Organization, ed. (2006) *Neonatal and perinatal mortality: country, regional and global estimates*. World Health Organization.

World Health Organization. Reproductive Health, ed. (2003) *Managing Complications in Pregnancy and Childbirth: A guide for midwives and doctors*. World Health Organization.

Zander, Josef and Elisabeth Goetz (1986) ‘Hausgeburten und klinische Entbindung im Dritten Reich (Über eine Denkschrift der Deutschen Gesellschaft für Gynäkologie aus dem Jahre 1939)’. In: *Zur Geschichte der Gynäkologie und Geburtshilfe: Aus Anlaß des 100jährigen Bestehens der Deutschen Gesellschaft für Gynäkologie und Geburtshilfe*. Ed. by Lutwin Beck. Springer.

Ziebarth, Nicolas and Gert Wagner (2013) *How attitudes toward risky behavior are shaped: evidence from performance-enhancement drugs and divided Germany*.