



Network for Studies on Pensions, Aging and Retirement

Topics in Economics of Labor, Health, and Education

Yi Zhang

NETSPAR ACADEMIC SERIES

PhD 05/2020-006

Topics in economics of labor, health, and education

Zhang, Yi

Document version:

Publisher's PDF, also known as Version of record

Publication date:

2020

[Link to publication](#)

Citation for published version (APA):

Zhang, Y. (2020). *Topics in economics of labor, health, and education*. CentER, Center for Economic Research.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.



Topics in Economics of Labor, Health, and Education

YI ZHANG

Topics in Economics of Labor, Health, and Education

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. K. Sijtsma, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie aan Tilburg University op maandag 29 juni 2020 om 16.00 uur door

Yi Zhang

geboren op 30 oktober 1985 te Hubei, China

Promotiecommissie:

Promotor: Prof. dr. A.H.O. van Soest

Copromotor: Dr. M. Salm

Overige leden: Dr. J.R. de Bresser
Dr. N.M. Daysal
Prof. dr. M.G. Knoef
Prof. dr. ir. J.C. van Ours

Acknowledgements

This dissertation would not have been possible without the kind help and support from many people.

I would like to acknowledge my indebtedness and render my warmest thanks to my supervisors, Martin Salm and Arthur van Soest, for their invaluable guidance and generous help throughout all stages of the work. Martin has always been a great mentor and a good friend, unreservedly sharing experience on academic writing (“Write short sentences!”), making effective presentations (“No more than 4 lines on a page!”), doing research (“The first step of writing a good paper, is writing a bad one...”), and living wisely (“If you cannot reduce the consumption of sth., you’d better stop it completely...”; “It is important to finish things...”)...His experience turns out to be important both to PhD research and to life in general... Arthur, just like Professor Albus Dumbledore in *Harry Potter*, has been the role model of many of his PhD students. I, just like many other students, secretly pray that one day I could be a productive researcher as he is, a good teacher as he is, and a capable and benevolent person as he is. I would like to thank Martin and Arthur for always trusting me and giving me opportunities to try new things, and for always being encouraging whenever I feel upset or unconfident. What they taught me and did for me is more than I could ever give them credit for here. It has been very lucky, in every respect, for me to be their student.

I would also like to express my deep gratitude to the other members of my doctoral committee, Jochem de Bresser, Meltem Daysal, Marike Knoef, and Jan van Ours. Their insightful suggestions have contributed greatly to the improvement of the thesis. Besides, the earlier version of Chapter 4 benefits a lot from the extended discussions with Jochem. His advice on job market preparation is also super helpful. Meltem and Marike offered me great opportunities to present the preliminary work of Chapter 3 and Chapter 4 at their workshops where many valuable suggestions were collected. Jan gave extensive advice on the earlier version of Chapter 2, as well as his kind support on my job market.

I want to thank my collaborators, Jia He, Tunga Kantarcı, and Jan-Maarten van Sonsbeek for their tremendous efforts dedicated to the projects. I learned a lot through the cooperation with them. Thanks for bearing with my randomness and for

bringing so much fun to the research process.

I would like to thank SEG participants and other faculty members for their constructive feedback on my work. In particular, the dissertation benefits a lot from the advice from Jaap Abbring, Bart Bronnenberg, Tobias Klein, and Moritz Suppliet. Discussions on this thesis and a wider range of research topics with Lei Lei, Jan Kabátek, Ana Moura, Ittai Shacham, Suraj Upadhyay, and Mingjia Xie have always kept me inspired. Many thanks to Otilia Boldea, Bettina Siflinger, Christoph Walsh, and Bas Werker for their helpful suggestions both on my papers and on preparing for the job market.

I also want to take a moment and extend my gratitude to my friends and colleagues. Ruonan Fu, Wenqian Hu, Xue Xu, and Yuanyuan Xu, all the happy time with you has been an indispensable part of my ~~PhD~~ life. The distance can never separate our friendship (we have Wechat anyways...). Elisabeth Beusch and Laura Capera Romero, my lovely “Granny Office” mates, thanks for all the interesting research discussions, the emotional support, and many happy talks that kept us off from work. Thanks to my foodie friend Chen He, who had shown me so many nice restaurants in the Netherlands. I have passed your philosophy for the greater food on to the next cohorts... Many thanks to Xiuqi Lin, Manwei Liu, Junjie Zhang, and Sili Zhang for fighting together in the *Langrissser* world. Thanks for the support and help from my friends: Santiago Bohorquez Correa, Mirthe Boomsma, Thijs Brouwer, Shuai Chen, Kadircan Çakmak, Wentian Diao, Lenka Fiala, Rafael Greminger, Di Gong, Tao Han, Yi He, Dorothee Hillrichs, Yue Hu, Maciej Husiatyński, Ye Kong, Xu Lang, Jing Li, Xuan Li, Zihao Liu, Shuo Liu, Pintao Lv, Emanuel Marcu, Zilong Niu, Renata Rabovic, Lingbo Shen, Yi Sheng, Lei Shu, Chen Sun, Bas van Heiningen, Xiaoyu Wang, Takumin Wang, Xingang Wen, Oliver Wichert, Yan Xu, Jierui Yang, Yadi Yang, Yuxin Yao, Wencheng Yu, Yifan Yu, Da Zhang, Miao Zhang, Wanqing Zhang, Xiaoyue Zhang, Nan Zhao, Shuo Zhao, Kun Zheng, Trevor Zheng, Yeqiu Zheng and Bo Zhou.

I would like to extend my sincere gratitude to Cecile de Bruijn for her kind support on the job market, to Korine Bor for always standing in our shoes (and finding ways to cover my travel expenses), and to our secretaries of EOR: Anja Heijeriks, Anja Manders-Struijs, Monique Mauer, and Heidi van Veen for their excellent support (the continuous provision of Cola Light has been crucial to my productivity...).

Special thanks to Jens Prüfer and Sebastian Dengler. My journey in Tilburg would have stopped much earlier before the PhD stage without their helping hands in my emotionally difficult times.

Finally my deepest love goes to my parents and my husband Mi Zhou for their constant love and unconditional support.

Yi Zhang

May 4, 2020

Beijing, China

Contents

Chapter 1 Introduction.....	1
Chapter 2 The Effect of Training on Workers' Perceived Job Match Quality	4
2.1 Introduction	4
2.2 Data and Measurement	8
2.3 The Effect of Training on Job Match Quality	15
2.4 Mechanisms	19
2.5 Sensitivity Analysis	23
2.6 Conclusion	26
Appendix 2.A Tables	28
Appendix 2.B Alternative Way to Construct Job Match Quality	40
References for Chapter 2	42
Chapter 3 The Effect of Retirement on Healthcare Utilization: Evidence from China.....	45
3.1 Introduction	45
3.2 Institutional Background and International Comparison	48
3.3 Data	51
3.4 Empirical Strategy	58
3.5 Main Results	61
3.6 Mechanisms	64
3.7 Sensitivity Analysis and Specification Checks	70
3.8 Discussion.....	75
Appendix 3.A Tables	77
Appendix 3.B Figures.....	91
Appendix 3.C Variables Used in Further Analysis	103
References for Chapter 3	104
Chapter 4 The Impact of a Disability Insurance Reform on Work Resumption and Benefit Substitution in the Netherlands.....	107
4.1 Introduction	107
4.2 Institutional setting	111
4.3 Data	116
4.4 Descriptive statistics	119
4.5 Empirical strategy.....	126
4.6 Main results	129
4.7 Effect over time	132
4.8 Heterogeneous effects.....	133
4.9 Sensitivity analyses.....	140
4.10 Conclusion	151
Appendix 4.A The under-reporting of sickness cases shorter than 180 days	153
Appendix 4.B Descriptive plots with yearly data	156
Appendix 4.C A back-of-envelope calculation to decompose effects of WIA reform on the probability of claiming DI and the probability of working	158
References for Chapter 4	160
Chapter 5 Measuring Non-cognitive Skills Exploiting Log-files on Online Behavior	163
5.1 Introduction	163
5.2 Two Examples: Perseverance and Deep Learning	165
5.3 Discussion.....	173
Appendix 5.A Details on Sample Restrictions of Example 1	175
References for Chapter 5	176

Chapter 1

Introduction

This dissertation consists of four essays on topics in labor economics, health economics, and economics of education. It primarily investigates how policy-induced incentive changes influence individuals' labor market and health outcomes. It also explores new methods to construct measures for education-related variables.

The first paper in Chapter 2, “*The Effect of Training on Workers' Perceived Job Match Quality*”, studies how training improves the quality of a job match. The quality of a job match indicates how well the characteristics of a worker match those of a job. It receives increasing attention as low job match quality, or mismatch, is associated with wage penalties, absenteeism, high turnover, and other negative labor market outcomes. A possible measure to improve job match quality that has been frequently discussed is training. We study the causal effect of training on the quality of a job match using longitudinal data for a representative sample of the Dutch population. To account for the multi-dimensional nature of job match quality, we construct an index of workers' perceived job match quality from five survey questions on job satisfaction and on how a worker's education and skills match with the job. Based on a dynamic linear panel data model, which accounts for potential endogeneity of training, we find that training has significantly positive short- and long-term effects on job match quality. This is mainly driven by training for human capital accumulation. Further analysis incorporating job changes shows that training for job change purpose increases the probability to change jobs, but job changes immediately following this type of training do not significantly increase job match quality. On the other hand, those who change jobs one year after this training do tend to get a better-matched job.

The second paper in Chapter 3, “*The Effect of Retirement on Healthcare Utilization: Evidence from China*”, studies how retirement influences healthcare utilization among elderly individuals. This question is important for the understanding of the full consequences of retirement policies. We examine the causal effect of retirement on healthcare utilization in China using longitudinal data. We use a nonparametric fuzzy regression discontinuity design, exploiting the statutory retirement age in

urban China as a source of exogenous variation in retirement. In contrast to previous results for developed countries, we find that in China retirement increases healthcare utilization. This increase can be attributed to deteriorating health and in particular to the reduced opportunity cost of time after retirement. For the sample as a whole, income is not a dominating mechanism. People with low education, however, are more likely to forego recommended inpatient care after retirement. The fact that retirement increases healthcare use means that, at least in the short run, raising the statutory retirement ages would reduce expenditures on public health insurance in urban China. On the other hand, raising retirement ages might have negative effects on health if workers postpone necessary treatment due to time constraints. An increase in retirement ages should therefore go along with more facilitation of preventive care and more efforts to reduce employees' opportunity costs of seeking medical treatment. Moreover, policy makers should not ignore that high co-payments can imply financial barriers to medical care and can lead to more forgone inpatient care for the low socioeconomic status group.

The third paper in Chapter 4, "*The Impact of a Disability Insurance Reform on Work Resumption and Benefit Substitution in the Netherlands*", studies how disability insurance (DI) reform influences sick individuals' labor market participation and social benefit claiming. We evaluate a major DI reform introduced in 2006 in the Netherlands. This reform introduces a basket of changes including extending the waiting period before applying for disability insurance by one year, tightening the entrance criteria, and introducing work resumption incentives, etc. Using unique administrative data on agents who fall sick before and after the reform, and are hence subject to old and new DI systems, we analyze the overall effect of the reform on individuals' labor participation decisions and on the use of benefits from alternative social security programs. Difference-in-difference analyses find that the reform decreases the use of disability benefit substantially and persistently. It increases labor participation and the use of unemployment benefits to a sizable extent. The effects on labor participation and salary are persistent, but the spillover effect on the use of benefits from alternative benefit programs is non-lasting. The reform is least effective for unemployed and older individuals who struggle most to resume working. The reform even decreases the probability of working and the income for the unemployed individuals, compared with their counterparts insured under the old DI system. The worsening prospect of work resumption for the unemployed is possibly

due to a larger scarring effect and more human capital loss, as a result of spending more time waiting for DI due to the extension of the waiting period by an extra year. This raises inequality concerns of the reform for this vulnerable labor market group.

The fourth paper in Chapter 5, “*Measuring Non-cognitive Skills Exploiting Log-files on Online Behavior*”, proposes a new measure of non-cognitive skills. Conventional self-reported measures of non-cognitive skills suffer from self-presentation bias and reference group effects, which can produce paradoxical results, particularly in cross-country comparisons. We propose a new source of measures derived from computer-generated log files on the behavior of individuals taking an online test or respondents taking an online survey. We analyze measures of two desirable non-cognitive skills, perseverance and deep learning, constructed with log-file data from two large-scale educational surveys. We show that these measures have higher cross-country comparability as they predict the performance of tests consistently at both the individual and the country level. We also discuss the methodological implications of log-based behavioral measures, and encourage researchers to apply them in combination with conventional self-reported measures.

Chapter 2

The Effect of Training on Workers' Perceived Job Match Quality¹

2.1 Introduction

Job match quality is increasingly recognized as an important predictor not only of individuals' psychological, social, and economic well-being, but also of firm productivity, and even of economic growth. Individual-level analyses have shown that low job match quality, or mismatch, is closely associated with wage penalties, absenteeism, high turnover, and other negative labour market outcomes, even controlling for wages, working hours, and standard demographic and job characteristics (Vahey 2000, Dolton and Vignoles 2000, Allen and van der Velden 2001, Clark 2001, Green and Zhu 2010, Nordin et al. 2010, Mavromaras et al. 2013, Pecoraro 2014, Congregado et al. 2016). Firm-level meta-analysis finds that higher job match quality is related to higher employee engagement and firm profitability and productivity (Harter et al. 2002).

Moreover, "improving job match quality" is a strategic goal of the European Union: The 10-year strategy of Europe 2020 identifies "better matching labour supply and demand" and "developing skills throughout the lifecycle" as new engines to boost economic growth and to increase job quality.² The recent Strategic Framework for Health and Safety at Work 2014-2020 emphasizes the importance of improving job quality for the competitiveness and productivity of European companies.³

In spite of the acknowledged importance of job match quality, there is limited empirical research on how job match quality can be improved. A possible measure to improve job match quality that has been frequently discussed is training. However,

¹ This chapter is the same as the published version in *Empirical Economics*. It is coauthored with Martin Salm and Arthur van Soest. The authors would like to thank CentERdata of Tilburg University for providing the data. We are very grateful for many helpful comments of the anonymous referee, the associate editor, and the coordinating editor. We are grateful to Jeffrey Campbell, Tobias Klein, Jan van Ours, Loes Verstegen, and seminar participants at Tilburg University and the 29th Annual Conference of EALE for their helpful comments and suggestions.

² <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF>

³ <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014DC0332>. Strictly speaking, job quality is not the same as job match quality. Job quality focuses on job characteristics, and it is a part of job match quality

studies that aim to estimate the effect of training on job match quality face two challenges. First, there is no unanimous definition of job match quality, and second, it is difficult to identify a causal effect of training on job match quality.

The quality of a job match indicates how well the characteristics of a worker match those of a job. Job match quality can be defined either from the worker's perspective (Kalleberg and Vaisey 2005, Clark 2015), or from the firm's perspective (Jones et al. 2009), using objective measures (Gaure et al. 2012, Lachowska et al. 2016, Le Barbanchon 2016) or subjective measures (Gottschalk and Maloney 1985, Clark 2005, Ferreira and Taylor 2011).

In this study, we use a measure of workers' perceived job match quality that captures the multidimensional quality of a job match. Following the example of Ferreira and Taylor (2011), we use factor analysis to derive a continuous measure of job match quality from five job match-related questions. Our measure is a combination of educational match, skill match, and satisfaction with the job. The use of job satisfaction as a measure of job match quality goes back to Ferreira and Taylor (2011) and Barmby et al. (2012). Our measure is correlated with observed job characteristics and educational background, and it predicts on-the-job search even after controlling for observed job characteristics, education, and individual fixed effects.

We use a broad concept of training – any course or educational program important for work or profession. This includes training on the job as well as off the job. Our aim is to estimate the causal effect of training on job match quality. Training can affect job match quality in multiple ways. First, training can increase human capital and individual productivity, as implied by human capital theory (Becker 1962). This could improve the perceived job match quality through a wage increase. Even without a wage increase, training could still improve the perceived match quality by filling gaps in knowledge and skills and by making employees feel more competent at their jobs. Second, training may have an indirect impact on match quality through job changes. Training can influence the probability of matching to a new job with higher requirements and usually also with higher wages, as implied by the theory of

career mobility (Sicherman and Galor 1990),⁴ which could lead to a change in the perceived match quality after switching to a new job.⁵

We estimate a dynamic linear panel data model, using eight years of data from the Dutch LISS data (Longitudinal Internet Studies for the Social sciences). A challenge for estimating the causal effect of training on job match quality is that training can be endogenous to job match quality. One reason is the presence of time persistent unobserved factors that drive participation in training and job match quality in related ways. For example, more ambitious people may select themselves into training and may also be more likely to find a better matched job. In our panel data models, we account for this using fixed individual effects, allowed to be correlated with training and other regressors. Second, training can be directly affected by job match quality if, for example, training aims at making up for the lack of skills required for the current job (i.e., the imperfect nature of the job match) or if training is taken because employees are dissatisfied with their job and wish to improve their labour market opportunities. We address the potential endogeneity of training exploiting the timing of events, assuming that shocks in job match quality are not correlated to past training (or other events in the past). This implies that we can use lagged variables of training as instruments for current training and apply GMM.

We find positive effects of training on job match quality both in the short-run and long-run. This finding is robust to alternative definitions of job match quality and training. We then explore possible pathways. First, since training for different purposes has different content and possibly different effects on job match quality, we estimate a model that allows for heterogeneous effects of training. Our results indicate that the effect of training is largest for training aimed at human capital improvement, in accordance with human capital theory. We also find some evidence

⁴ Note that human capital theory and career mobility theory are not mutually exclusive. The key distinction is that the original human capital theory only considers wage as the return of human capital investment, whilst care mobility theory provides additional dimensions of return, i.e. inter- or intra-firms occupational upgrading.

⁵ Similar mechanisms of how training can influence job match quality can also be derived from a search and matching framework (see, e.g., Mortensen 1978, Jovanovic 1979, or Topel and Ward 1992). On the one hand, match quality is revealed to both workers and firms as time elapses. If training affects productivity through human capital build-up then it would affect the perceived quality of the work-firm match. On the other hand, training could increase the search ability of the worker (e.g. training on CV writing skills leads to higher success rate of matching with any new job), or increase the probability of matching to jobs with higher quality (e.g. certificates allow workers to send signals to more challenging and better paid jobs).

supporting the theory of career mobility: We find that training for the purpose of a job change immediately increases the probability to change jobs. However, these new jobs are as often better matches as worse matches. On the other hand, for those who do not change jobs immediately but in the next period, training for job change purposes significantly improves job match quality. Finally, we find no evidence that training for other purposes (mainly training referred to as “required for my job”) has any effects on either job match quality or the likelihood to switch jobs.

Our study makes two contributions to the literature. First, we provide evidence for a causal effect of training on job match quality based on a dynamic panel data model. Existing empirical studies either focus on the association between training and match quality (Chiang et al. 2005, Jones et al. 2009, Han et al. 2014),⁶ or they identify a causal effect using different identifying assumptions than we use. For example, some previous studies assume that training is exogenous after conditioning on a set of observed characteristics (Georgellis and Lange 2007, Messinis and Olekalns 2008, Burgard and Gurlitz 2014, Pagan-Rodriguez 2015). In contrast, our study addresses the endogeneity of training exploiting the timing of events based on a dynamic panel data model.

Furthermore, we examine the mechanisms underlying the effect of training on match quality. For this purpose, we provide evidence for a causal effect of training on job changes and on job match quality after a job change. A previous study by Dekker et al. (2002) examines how training influences upward and lateral job mobility. However, limited by the cross-sectional nature of their data, they are not able to address the potential endogeneity in training. In contrast, this endogeneity can be accounted for with our empirical strategy.

Our paper continues as follows: Section 2.2 describes the data and the measurement of main variables. Section 2.3 presents the main empirical analysis for the effect of training on job match quality. Section 2.4 presents a more detailed analysis aimed at identifying mechanisms that explain this effect. Section 2.5 lists sensitivity analyses. Section 2.6 concludes.

⁶ Studies on job assistance programs sometimes use objective one-dimensional measures, e.g. job duration (Blundell et al. 2004) and wage (Crépon et al. 2013) etc., to indicate the quality of a job match. This is not our focus. The job match quality of our interest is from a subjective and multi-dimensional perspective.

2.2 Data and Measurement

We use data from the LISS panel (Longitudinal Internet Studies for the Social Sciences) administered by CentERdata affiliated with Tilburg University, which provides a representative sample of approximately 5,000 Dutch households.⁷ We combine LISS Panel Background Information and the module of Working and Schooling Survey (2008-2015, eight waves). The latter is a longitudinal survey on labour market participation, job characteristics, pensions, schooling and training courses, etc.⁸ The dataset is ideal for our analysis. It provides information about the match between the job and an individual's education and skills, which is rare in other longitudinal household datasets. Furthermore, the panel is long enough to estimate dynamic linear panel data models.

We apply some restrictions to our sample. We only keep individuals doing paid work and drop logically inconsistent observations.⁹ We only keep workers appearing in the dataset for at least two consecutive years, since we need information on at least one lag in the econometric model.¹⁰ We drop observations with missing values in main explanatory variables, e.g. training and job changes. The remaining sample size is 4905 individuals and 21,992 individual-year observations. The structure of the resulting unbalanced panel is listed in the Appendix Table 2.A.1.

Perceived Job Match Quality

From a worker's perspective, a measure of job match quality should be able to capture how "good" the job match is. This is a multidimensional concept, not only determined by contracted job characteristics (e.g. wage, hours of work etc.) and education background of the worker, but also influenced by match-specific characteristics only perceived by the worker like stressfulness, working atmosphere,

⁷ See <https://www.lissdata.nl/lissdata/about-panel> for detailed information of the LISS Panel.

⁸ See [http://www.lissdata.nl/dataarchive/study units/view/1](http://www.lissdata.nl/dataarchive/study%20units/view/1) for more modules of the LISS Panel.

⁹ For example, for the variable "year when entering the current job" (used to construct an indicator for changing jobs) in 2008, an individual reported entering the current job in 2006. But in 2010, she reported entering the current job in 1984. Observations with such logically inconsistent answers are dropped.

¹⁰ We start with 12,328 individuals and 48,752 individual-year observations. Keeping individuals doing paid work reduces the sample to 29,016 observations (8,748 individuals). By dropping logically inconsistent observations, we are left with 25,850 observations (8,092 individuals). Keeping those who appear in the dataset for at least two consecutive years reduces the sample to 22,454 observations (5,023 individuals).

self-realization etc. To capture observed and unobserved aspects of job match quality, there are two ways to construct a measure of workers' perceived job match quality. First, workers could be asked many detailed questions on satisfaction with pay, hours of work, future prospects, work pressure, job content, interpersonal relationships etc., and all these aspects of the job could be aggregated into a single measure. The practical problem with this method is that researchers can hardly be exhaustive in including all the relevant job aspects. Usually in survey data, questions on satisfaction cover a limited number of job aspects. The alternative is to ask more general questions about how good the job match is, assuming that respondents will aggregate the more detailed observed and unobserved aspects of job match quality themselves.

Table 2.1: Five Job-Match Related Variables

Variable	Overall			In Year 2011		
	Obs	Mean	Std. Dev	Obs	Mean	Std. Dev
Educational match	21992	5.633	2.524	2462	5.763	2.425
Skill match	21992	6.407	1.865	2462	6.485	1.798
Satisfaction with type of work	21793	6.627	1.551	2433	6.601	1.589
Satisfaction with career	21732	6.292	1.528	2434	6.277	1.555
Satisfaction with current work	21800	6.448	1.511	2438	6.42	1.53

We choose the second method and use five job match-related questions: (1) one question about educational match: “how does your highest level of education suit the work that you now perform”. (2) One question about skill match: “how do your knowledge and skills suit the work you do”. (3-5) Three questions about overall satisfaction of the job match: “How satisfied are you with the type of work that you do / your career so far / your current work”.¹¹ Table 2.1 shows some sample statistics for these five variables on a scale from 0 to 9.¹² For each of the five variables, a higher value points at better job match quality. On average, the educational match (question 1) is evaluated substantially worse than the other four aspects of job match quality.

¹¹ In sensitivity analysis (Section 2.5) we use a larger set of survey questions to construct an index of job match quality. The main results remain the same.

¹² The original variables range from 0 (worst) to 10 (best) in all years except 2011, when they range from 0 to 9. For all years except 2011 we combined the two lowest scores (which are rarely reported) and recoded scores from 0 to 9, leading to very similar distributions in all years.

Table 2.2 shows that there are high positive correlations among the answers to the five questions; the high value of Cronbach’s alpha of $0.807 > 0.7$ confirms their high internal consistency, as an indicator of an underlying factor. Following Ferreira and Taylor (2011), we therefore use factor analysis to derive a continuous measure of job match quality from the five job match-related variables. The first latent factor (the linear combination of the five variables that explains the most variation in the pooled data) is a summary measure of the perceived quality of the job match.¹³ Table 2.3 lists the factor loadings, which indicate the relative importance of each of the five questions.¹⁴

Table 2.2: Correlations of Job-Match Related Variables

	Educational match	Skill match	Satisfaction with types of work	Satisfaction with career	Satisfaction with current work
Educational match	1.000				
Skill match	0.609	1.000			
Satisfaction of type of work	0.340	0.425	1.000		
Satisfaction of career	0.352	0.415	0.692	1.000	
Satisfaction of current work	0.286	0.374	0.825	0.736	1.000

Table 2.3: Factor Loadings

	Factor loadings
Educational match	0.104
Skill match	0.126
Satisfaction with type of work	0.364
Satisfaction with career	0.254
Satisfaction with current work	0.289

¹³ The factor extraction method is iterated principal factor (IPF); the maximum-likelihood factor method gave very similar results. We only retain the first factor because this is the only one with eigenvalue larger than 1 (the eigenvalues of the five factors are 2.669, 0.549, -0.051, -0.121, and -0.205.)

¹⁴ The highest factor loading in Table 2.3 is 0.364. The literature does not reach a consensus on the suggested criteria for factor loadings. For example, Hair et al. (1998) suggest a minimum factor loading of 0.3 for a sample of size 350. Stevens (1992) suggests 0.4 irrespective of the sample size. Child (2006) relaxes it to 0.2. These criteria are often used for selecting variables for extraction. In our study, we do not use a data-driven approach to select variables. Instead, we select variables based on economic intuition. In our case, the factor loadings show the relative importance of each variable to the latent factor.

The constructed index of job match quality gives higher weights to satisfaction with work and career than to self-reported level of educational and skill match. In sensitivity analysis, we also use a simple average of the five components of job match quality, and the main results are robust to the weighting.

With the loadings of the first factor, we predict the dependent variable for each individual and rescale it to obtain our final index of job match quality on a continuous scale from 0 to 1, with mean 0.711 and standard deviation 0.149. Figure 2.1 displays the distribution of the index, which is asymmetric and left-skewed, with a small minority of workers with very low job match quality. In a sensitivity check (Section 2.5), we will analyse the influence of these very low values on our results.

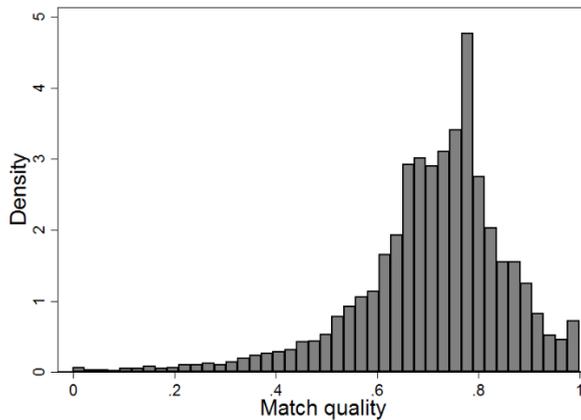


Figure 2.1: Distribution of Job Match Quality

To validate our constructed measure of job match quality, we first investigate how it relates to education and job characteristics (Appendix Table 2.A.2, Column 1). Controlling for individual fixed effects, job match quality is influenced by field of study, type of contract, income level,¹⁵ job sector, supervision level, working hours, tenure, and commuting time. Secondly, we check whether our index of job match quality is predictive of a worker’s on-the-job search behaviour (Table 2.A.2, Column 2). Controlling for individual unobserved heterogeneity and all the observed

¹⁵ Unfortunately, we do not have wage information in the data. Instead, we look at the level of net income. Income level is defined as: 0:”No income”; 1:”EUR 500 or less”; 2:”501 to 1000”; 3:”1001 to 1500”; 4:”1501 to 2000”; 5:”2001 to 2500”; 6:”2501 to 3000”; 7:”3001 to 3500”; 8:”3501 to 4000”; 9:”4001 to 4500”; 10:”4501 to 5000”; 11:”5001 to 7500”; 12:”More than 7500”.

education and job characteristics, we still find that higher job match quality predicts lower incidence of on-the-job search.¹⁶ This implies that our constructed measure of job match quality is able to capture at least part of the unobserved quality of a match from a worker's perspective. Moreover, it confirms that how people feel about job matches actually matters for their labour market behaviour.

Training

We use the following question to construct a dummy variable for training in our main estimation:

“Have you, in the past 12 months, followed any educational programs or courses or are you presently following one or more educational programs or courses? This concerns educational programs or courses that are important for your work or profession. (1 yes. 2 no.) ”

The definition of training here is broader than in the previous literature, and it includes both on-the-job training and off-the-job training. Due to the flexible Dutch education system, many people take part in formal education programs while working, e.g., a part-time vocational education program, or a night university program to get certificates for a profession etc. We allow for a variety of learning activities as long as they are considered as important for work or profession.

The status of receiving training or not is updated each year. There are 8096 individual-year observations taking training. A respondent could take multiple training programs. Respondents taking training were asked to report at most three training programs in the last 12 months. 3297 and 1091 out of 8096 observations reported to participate in a second and a third training program, respectively. In a sensitivity analysis, we constructed two mutually exclusive dummy variables “receiving a single training program” and “receiving multiple training programs”. We find that the effect of “multiple training” is not significantly different from that of “single training”. Therefore, we do not use the number of training programs in our analysis. As a measure of training intensity, we also constructed “total days of training per year”. It is the sum of the days spent in all three training programs (if applicable) per individual-year.¹⁷ Conditional on participating in training, the

¹⁶ The dummy variable for on-the-job search is constructed from the question: “*I perform paid work, but am looking for more or other work. (0 no. 1 yes.)*”

¹⁷ Days of training is calculated using the survey question “*What is the official duration of this program*

average total days of training per individual and year is 69 days. 70% of the observations have less than 30 days of training per year. Notably, about 19% of individuals who participated in training followed educational programs with a duration of at least 260 days (one year in our definition). An example for such a long-term educational program could be a part-time training program to become a yoga teacher. In Section 2.4 below, we also look at different types of training.

Job Change Incidence

Since there is no direct information about job changes in the survey, we construct a dummy variable for job change J_{it} , which indicates whether individual i starts a new job in year t , inferred from “the year when entering the current job”. J_{it} equals 1 if “the year when entering the current job” changes compared to the last period.

Table 2.4: Descriptive Statistics

Variables	Mean	Std. Dev
<i>Dependent variable</i>		
Job match quality	0.711	0.149
Job change incidence	0.061	0.240
<i>Treatment</i>		
Training	0.368	0.482
Total days of training per year (conditional on participating in training)	68.880	116.137
Single training	0.215	0.411
Multiple training	0.151	0.358
<i>Demographics</i>		
Female	0.515	0.500
Age in years	44.257	11.309
Disabled	0.011	0.103
Individuals	4,905	
Observations	21,992	

Table 2.4 presents the main descriptive statistics of our estimation sample. The sample size is 21,992 observations. About 52% of our sample are women. The

or course? __ (part-days/days/weeks/months/years).” On average, the days of the first, second and third training are 58, 21 and 15 days. For the total days of training, we add up the days of the three training programs. The maximum possible total days of training per year is $3 \times 260 = 780$ days.

average age is 44 years. Training is quite common in the Netherlands, with an incidence rate of 36.8% per year. On average, about 6.1% of all workers change jobs in a given year.

Figure 2.2 shows the average match quality over years relative to the year in which an individual receives training. If an individual received training in multiple years, each episode of training is considered separately. The plot on the left defines years between two trainings as “years after training”. For example, if we denote receiving training in a year as “1” and no training as “0” and an individual’s training sequence from year 2008 to 2015 is: “0, 1, 0, 0, 1, 1, 0, 0”, then for this sequence the year relative to training is “ $t-1, t, t+1, t+2, t, t, t+1, t+2$ ”. We compute average match quality for all observations at each year relative to training, for example at “ $t+1$ ”, “ $t+2$ ”, and so on. The plot on the right defines years between trainings as “years before training”. Then for the example above the training sequence “0, 1, 0, 0, 1, 1, 0, 0” corresponds to “ $t-1, t, t-2, t-1, t, t, t+1, t+2$ ”.

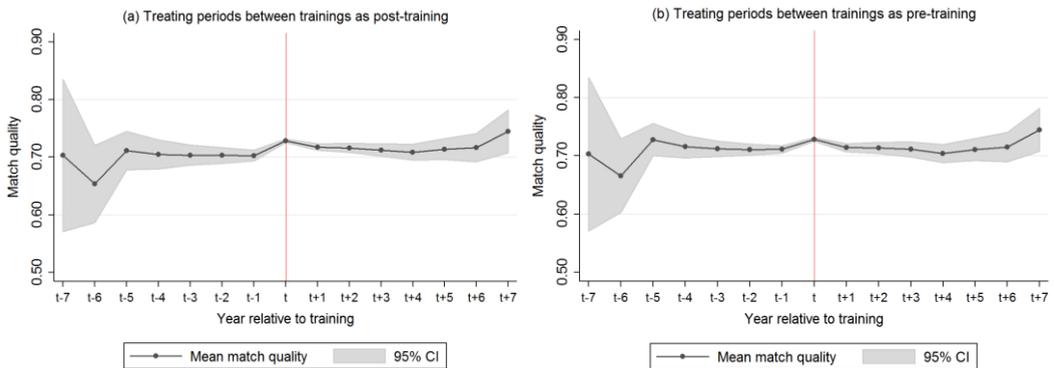


Figure 2.2: Match Quality Over Years Relative to Training

The two plots in Figure 2.2 show the same pattern. There is no dip in job match quality immediately before training. Job match quality is flat in the years before and after training, with an increase in the year of training only, suggesting that training only has a short-run effect on job match quality. This small and instantaneous increase in match quality does not necessarily reflect the causal impact of training, as we will further discuss in the next section.

2.3 The Effect of Training on Job Match Quality

The primary interest of this paper is to understand the causal effect of training on job match quality. As already discussed in Section 2.1, training may be endogenous. A first potential source of endogeneity is the presence of unobserved individual characteristics that simultaneously drive participation in training and perceived job match quality. Another potential source of endogeneity can be unobserved employer policy if, for example, some firms deliberately hire employees that first need to be trained before they are ready for the job. Moreover, there might be reverse causality between training and match quality - individuals who experience an unexpected change in the quality of their match may change their participation in training in the same year.

In order to address the above issues, we specify the following dynamic linear panel data model:

$$M_{it} = \sum_{j=1}^3 \theta_j M_{it-j} + \gamma T_{it} + X'_{it} \beta + \alpha_i + \varepsilon_{it} \quad (2.1)$$

For $t = 4$ to 8 and $i = 1 \dots N$.¹⁸

M_{it} is the constructed index of job match quality. The M_{it-j} are lagged dependent variables, with “state dependence” coefficients θ_j .¹⁹ They capture dynamics in M_{it} , e.g. due to partial adjustment. For example, a worker with low perceived job match quality last year (low M_{it-1}) who did not change jobs, is likely to still have low perceived quality this year (low M_{it}).

T_{it} is a dummy variable for training. Note that T_{it} measures the training participation in the time interval between $t-1$ and t (the past 12 months), while M_{it} is the job match quality measured at time t .²⁰ The parameter γ is the short run treatment effect of training on job match quality. Effects in the longer run also depend on the θ_j . A more general Local Average Treatment Effect (LATE) interpretation of γ does not seem possible, since there is no good reason why the monotonicity condition should be satisfied.²¹

¹⁸ Observations in years 2008 to 2011 are dropped due to the inclusion of lagged dependent variables.

¹⁹ Our choice of using three lags is based upon specification tests; specifications with one or two lags are rejected by the Arellano-Bond serial correlation test.

²⁰ In an alternative specification, we added lagged training T_{it-1} but this was not significant.

²¹ (The change in) past training is used as an instrument for (the change in) current training. The effect

X_{it} is a vector of control variables, including age, age squared, and dummy variables for work disability and calendar years. Since they cannot be chosen by the worker, it seems plausible to assume that they are strictly exogenous (i.e., independent of all $\varepsilon_{is}, s = 1, \dots, T$). We do not control for current education or job characteristics because they may change as the result of training. Past education and job characteristics are captured in M_{it-j} .

The α_i refer to individual fixed effects, capturing time invariant unobserved heterogeneity such as genetic traits and personality. The ε_{it} are idiosyncratic error terms, assumed to be independently and identically distributed. We assume that ε_{it} is an “innovation”, independent of everything that happened before time period t , including T_{i1}, \dots, T_{it-1} . This seems plausible, since individuals cannot make training decisions that anticipate unpredictable future shocks in job match quality. On the other hand, this assumption allows for an arbitrary correlation between ε_{it} and training in *current and future periods*. In other words, past or current shocks to match quality may influence training participation in the same time period or in later time periods. In this way, we exploit the timing of events for identification. The identifying assumptions are supported by the usual tests for misspecification: Both the Sargan test based upon over-identifying restrictions and the Arellano-Bond test for autocorrelation in the error terms lead to the conclusion that the assumption cannot be rejected.

The dynamic panel data model introduced above is estimated with system GMM estimation (Blundell and Bond, 1998) with finite sample correction for the variance of linear efficient two-step GMM estimators (Windmeijer, 2005).²² The instruments for the differenced equation are M_{it-2} to M_{it-7} , and ΔX_{it} for $t = 5$ to 8. Since the error term in the differenced equation is $\varepsilon_{it} - \varepsilon_{it-1}$, these instruments are valid if ε_{it} is indeed independent of everything before time period t . The instruments for the level equation are ΔT_{it-1} and ΔM_{it-1} for $t = 4$ to 8. Their use relies on auxiliary stationarity assumptions, see Blundell and Bond (1998). Table 2.A.4 in the Appendix shows how sensitive the estimated coefficient of training $\hat{\gamma}$ is when we vary the instruments of training and lagged dependent variables. For all GMM

of past training on current training may be positive in some cases (habit formation, advantages of continuous learning) but negative in others (firms may stimulate different workers for training each year).

²² Arellano-Bond GMM estimates (Arellano and Bond, 1991) give similar results.

Table 2.5: The Effect of Training on Job Match Quality

	(1)	(2)	(3)	(4)	(5)	(6)
Methods			OLS	FE-FD		
	OLS	FE-FD	Reduced	Reduced	GMM-SYS	GMM-SYS
	Full Sample	Full Sample	Sample	Sample		
Match quality (t-1)					0.398***	0.391***
					(0.043)	(0.044)
Match quality (t-2)					0.141***	0.137***
					(0.031)	(0.032)
Match quality (t-3)					0.045**	0.046**
					(0.023)	(0.023)
Training	0.032***	0.008***	0.028***	0.005*	0.007*	0.045**
	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)	(0.019)
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Other controls	Yes	Yes	Yes	Yes	Yes	Yes
Training endogenous	No	No	No	No	No	Yes
Specification tests					p-value	p-value
m2 test (p-value)					0.102	0.141
m3 test (p-value)					0.276	0.295
Sargan test (p-value)					0.081	0.186
Individuals	4896	4896	1900	1900	2336	2336
Observations	21677	21677	7077	7077	6890	6890

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are robust standard errors in columns (1) and (3), and WC-robust standard errors in columns (2), (4), (5) to (6) (Windmeijer, 2005). For the specification tests, the p values are reported. Column (1) and (2) are based on the whole sample. Column (1) shows the pooled OLS estimates. Column (2) shows the two-step GMM fixed effects (first difference) estimates. Column (3) and (4) show the same specifications as column (1) and (2) but for a sample that is comparable to columns (5) and (6). Columns (5) and (6) show system GMM two-step estimates. In column (5), training is taken as exogenous. The instruments for differenced equation are M_{it-2} to M_{it-7} , and the differenced exogenous variables (including training). The instruments for the level equation are ΔM_{it-1} . In column (6) training is treated as endogenous. The instruments for differenced equation are T_{it-2} , M_{it-2} to M_{it-7} , and the differenced exogenous variables. The instruments for the level equation are ΔT_{it-1} and ΔM_{it-1} . The specification tests m2 and m3 are the Arellano-Bond tests for 2nd and 3rd order autocorrelation in the differenced error terms. See Table 2.A.3 for the complete version of Table 2.5. Other controls include age, age squared, and dummy variables for work disability. Additionally, a dummy for female is included in other controls for Columns with OLS estimation.

estimation results, we report the p-values of the Sargan test and the Arellano-Bond autocorrelation test for the validity of the model assumptions. In all cases presented, the p-values imply that the model assumptions are not rejected using tests of size $\alpha = 0.05$.

We cannot directly compare this number 0.007 with 0.008 in Column 2 because the inclusion of lagged dependent variables results in a smaller sample. We therefore construct a comparable sample keeping individuals observed for at least 2 consecutive years for first-differencing, and dropping observations from year 2008 to 2011 for the lagged dependent variables.²³ Based on this reduced sample we run the same estimation as in Columns 1 and 2, which yields the results in Columns 3 and 4, respectively. The estimates for γ are slightly smaller than for the full sample (cf. Col. 1). Column 4 shows that the estimate of γ in a static fixed effects model on the smaller sample is only 0.005 (standard error 0.003), which is smaller than the 0.007 estimate (standard error 0.004) in Column 5. This suggests that omitting lagged dependent variables tends to bias the effect of training downwards.

Column 6 presents the estimates based on the assumptions, instruments and moments discussed in the previous section, allowing training to be endogenous and instrumenting training with past training. This is our preferred specification. The estimated effect of training increases to 0.045. A possible explanation for the large change compared to Col. 5 could be that workers who suffer from a poor current job match are inclined to take training in the same year to improve job match quality. This would make training contemporaneously endogenous. Ignoring this endogeneity (“reverse causality”) will lead to underestimation of the effect of training.²⁴

The significant lag terms show that there is positive state dependence in the dynamic adjustment process of job match quality. The short-term effect of training on job match quality is 0.045. The long-run effect is $\frac{\gamma}{1-(\theta_1+\theta_2+\theta_3)} = \frac{0.045}{1-(0.391+0.137+0.046)} = 0.106$, more than twice as large. This means that if an individual permanently changes from no training to training in each year (e.g. life-long learning), perceived job match quality improves by approximately 70% of one standard deviation of the index in the long run.

²³ The sample size for the comparable sample is not exactly the same as in Columns 5 and 6. This is due to differences in reporting sample sizes across STATA commands. For example, individuals with only one year of observation, which do not contribute to the first-differencing estimation results, are not automatically dropped from the sample. In spite of the slight difference, we show in Table 2.A.9 that the descriptive statistics for the reduced sample are similar to those in Table 2.4.

²⁴ A Hausman specification test on the training coefficient shows that the specification estimated in Column 5 is rejected against the more general specification in Column 6 (p-value 0.041).

The only control variable of substantive interest is the dummy variable “disabled” (see Table 2.A.3). Keeping other variables constant, having a disability significantly reduces the perceived match.

2.4 Mechanisms

As discussed in Section 2.1, there are many different types of training, and these may have different effects on job match quality. To understand why we found substantial positive overall effects of training on job match quality, we distinguish different types of training according to the purpose for which the training is taken. We consider training to improve human capital (Becker, 1962), training taken to improve labour market opportunities (the “career mobility theory” of Sicherman and Galor, 1990), and training taken for other purposes. We also investigate whether the types of training work through their intended mechanism.

We use the following survey question asked to all workers who took some training: “What was your main reason to start following this program or course?” The answers “1 to stay up-to-date in my profession (3773 observations)” and “2 to gain promotion (493 obs.)” are categorized as “training for human capital build-up”. The answer “3 to increase my chances of getting another job (657 obs.)” is categorized as “training for job change”. The remaining reasons are categorized as “other training”.²⁵ The majority participates in training for human capital build-up purpose (4266 out of 8044). Around 8% of all training is taken for the purpose of changing jobs.²⁶

To estimate separate effects of training for the three purposes, we replaced the training variable with three indicators for the three types of training, and estimated a model similar to the main estimation (Table 2.5, Column 6), allowing for the

²⁵ The remaining answers are “4 required by my job (2219 obs.)”, “5 required by CWI / UWV / Public Employment Service (15 obs.)” “6 required by municipality or social service (15 obs.)”, “7 am still of school age (34 obs.)”, “8 am still completing my school career (147 obs.)”, “9 for another reason (691 obs.)”. In an alternative estimation, we drop the 211 observations in answer 5, 6, 7 and 8. This makes hardly any difference for the results.

²⁶ If a respondent took several training programs in a given year, this classification is based on the first reported training. This assumes that the first training is the most important training, which is supported by the fact that the average duration of the first training per year is 58 days, compared to 21 days for the second and 15 days for the third training. In Table 2.A.5, we present the results of a sensitivity check with a non-mutually exclusive way of defining type of training, treating different training programs equally. In this case, for an individual who participates both in job change training and other training, both type-of-training variables will take value 1. The results retain the same pattern.

endogeneity of training and its purpose. We make essentially the same identifying assumptions as before (Table 2.5, Column 6) – shocks in job match quality are not related to past training (or purpose of training) or past job match quality. Results are summarized in Table 2.6, Column 1.²⁷

We find that training for human capital improvement has a large and significant effect, improving job match quality in the short run by 0.079 (53% of a standard deviation) and in the long-run by 0.171.²⁸ Since this is also the most common purpose of training, this finding implies that training for human capital improvement largely explains the positive training effects we found in Section 2.3. The effect of training for job change is not significantly different from zero, but also not significantly different from the effect of training for human capital improvement. The effect of other training (mainly training “required by my job”) is even smaller and also insignificant.

In the other columns of Table 2.6, we analyse whether training for a given purpose indeed affects job match quality through the intended mechanism, with a focus on changing jobs or not. We use the same type of dynamic panel data model as in our main estimation, accounting for the dynamics of job match quality, for fixed individual effects, and for endogeneity of the training variables (cf. the model in column 6 of Table 2.5). We do not include the second and third lags of job match quality because they are jointly insignificant.

Column 2 considers the intermediate step: it explains the likelihood of a job change, one possible pathway to improve job match quality. It shows that training for job change purposes tends to achieve its goal: it substantially increases the probability to switch jobs. The effect of other training on job change incidence is also positive and significant. But the training for human capital improvement has no such effects. This is in line with Cheng and Waldenberger (2013) who find that the effect of training on job change depends on the type of training, though their distinction between training types is different: They find that training for specific skills is associated with lower turnover intentions, while training for general skills is

²⁷ We also tried adding lags of the training variables but their coefficients were very small and jointly insignificant.

²⁸ Calculated as $0.079 / (1 - 0.367 - 0.127 - 0.044)$.

Table 2.6: Effect of Training by Purpose and Effect of Training through Job Change

GMM-SYS	(1)	(2)	(3)
Dependent variable	Match quality	Job Change Incidence	Job Change Outcome
Match quality (t-1)	0.367*** (0.087)	-0.225*** (0.056)	-0.334*** (0.105)
Match quality (t-2)	0.127*** (0.048)		
Match quality (t-3)	0.044* (0.023)		
Human capital build-up training	0.079*** (0.025)	-0.049 (0.051)	0.080 (0.069)
Human capital build-up training (t-1)			-0.006 (0.014)
Job change training	0.019 (0.050)	0.237* (0.122)	-0.057 (0.105)
Job change training (t-1)			0.060** (0.029)
Other training	0.014 (0.030)	0.124** (0.052)	0.081 (0.058)
Other training (t-1)			-0.007 (0.014)
Time fixed effects	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Training endogenous	Yes	Yes	Yes
Specification tests	p-value	p-value	p-value
m2 test	0.137	0.332	0.677
m3 test	0.243	0.554	0.661
Sargan test	0.592	0.201	0.154
Individuals	2,336	4867	4,823
Observations	6,890	15666	15,495

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are WC-robust standard errors. For the specification tests, the p values are reported. All columns are estimated with system GMM estimator. In column (1) and (2), the instruments are similar to those in column (6) of Table 2.5, except that all the instruments of training are replaced with instruments of three types of training: “human capital build-up training”, “job change training”, and “other training”. In column (3), T_{ijt-2} to T_{ijt-7} and M_{it-2} to M_{it-7} are used as instruments for differenced equation. In all columns, we use differenced exogenous variable as instruments for differenced equation. And we use ΔT_{ijt-1} and ΔM_{it-1} for the level equation. The specification tests m2 and m3 are the Arellano-Bond tests for 2nd and 3rd order autocorrelation in the differenced error terms. Other controls are the same as those in Table 2.5. In Column (3) we drop 315 observations with missing values for job match quality.

associated with higher turnover intentions. Column 2 also shows that a higher past job match quality significantly lowers the probability to change jobs.

The final column in Table 2.6 analyses whether training helps to find not only a new job but also a better match.²⁹ This question is related to the topic of Dekker et al. (2002), who study how training influences upward mobility (job-to-job moves that result in an increase in job level) and lateral mobility (moves without change of job level). They use cross-sectional data and include training participation as an exogenous variable, controlling for many other individual and job characteristics

Our dynamic linear panel data model is similar to the earlier models (e.g., column 6 in Table 2.5) and accounts for potential endogeneity in training in the same way. Moreover, based upon preliminary estimation results, we added past training participation as explanatory variables. The dependent variable is a constructed variable “job change outcome” O_{it} interacting the job change incidence dummy with the sign of the change of job match quality (-1 for a deterioration, 0 for no change, 1 for an improvement). It equals 0 when individual i does not change jobs in year t , or changes to a job but retains the same level of job match quality. It is -1 when the worker changes to a new job with lower job match quality and 1 when the worker changes to a job with higher match quality.³⁰ We therefore do not condition on job changes, but explain the joint outcome of whether someone changes jobs or not and if so, how this changes job match quality.

Most of the time, no job change takes place (20,357 observations). Of the 1320 observed job changes, 61.4% are changes to a better match (810 observations with $O_{it} = 1$) and 36.4% to a worse match (480 observations with $O_{it} = -1$), while the remaining 2.2% change to a new job with the same perceived match quality (30 observations).

The estimated coefficient of training in this model can be interpreted as the effect of training on the difference between the probabilities of changing to a job with better quality and changing to a job with worse quality.³¹ As expected, the estimated

²⁹ Previous studies on how job changes influence mismatch (Congregado et al. 2016) or job satisfaction (Zhou et al. 2017) give mixed results. Congregado et al. find hardly any effect, while Zhou et al. (2017) find a positive short-run effect on job satisfaction. These studies did not address the role of training.

³⁰ Here we drop 315 observations with missing values for job match quality.

³¹ Because in linear model, $E(O_{it}) = 1 \cdot P(O_{it} = 1) + 0 \cdot P(O_{it} = 0) - 1 \cdot P(O_{it} = -1) = P(O_{it} = 1) -$

coefficient of the lagged job match quality is negative, because the higher the match quality in the old job, the less likely is a change to a job with even higher match quality.

The immediate effect of job-change training is not significant, but job-change training does have a significant positive effect on the job change outcome one year later. The estimated coefficient means that participating in job-change training in the last period will increase the probability difference by 0.060, raising the probability of getting a new job with a better match, and/or reducing the probability of getting a worse matched new job. This finding is in line with career mobility theory. Training for other purposes than changing jobs (training for human capital improvement or other purposes) has no significant effect, as expected.³²

Combining results in columns 2 and 3 suggests that taking job-change training will increase the probability to change jobs immediately, but there is no evidence that these immediate changes tend to lead to better job matches. For those who take some time and change jobs in the next period, job change training tends to lead to a better-matched job.

2.5 Sensitivity Analysis

One concern might be the unbalanced panel structure caused by sample attrition. To investigate if our results are influenced by sample attrition, we further restrict the sample to observations that are in the data for at least five consecutive years. The resulting sample is more balanced but also more selective (see Table 2.A.8 in the Appendix).³³ We reconstructed (and rescaled) the dependent variable for this new sample and performed the same system-GMM estimation; see Column 1 of Table 2.7. The estimated effect of training is slightly smaller than the original estimate, possibly because the new sample leaves out individuals with more unstable employment status who may potentially benefit more from training. The same reasoning suggests that, due to attrition, the estimated effect of training according to

$P(O_{it} = -1)$. Separate estimates for the effects on improvement and deterioration through job change give less precise and insignificant results.

³² The coefficients of the four training dummy variables are jointly insignificant.

³³ The new sample has fewer workers engaged in temporary jobs, unskilled jobs, and self-employment.

our preferred estimates in Table 2.5, Column 6 also slightly underestimates the effect in the complete population.

Another concern is the skewed distribution of the dependent variable. In response to this concern, we truncate the distribution of job match quality at 0.42116.³⁴ We do the same estimation with the truncated sample. Column 2 of Table 2.7 shows that the effect of training remains positive and significant, though it is slightly reduced in magnitude. This makes sense since truncation removes the most mismatched workers, who may benefit most from training.

Third, we check if the results are sensitive to which components we include to construct our measure of job match quality. Besides the five variables used in the main body of the paper, we further include “satisfaction with earnings”, “satisfaction with working hours” and “satisfaction with the general atmosphere among your colleagues”. These three variables focus on satisfaction with specific job characteristics, rather than the overall perception of the quality of the match and seem less directly affected by training. Table 2.B.1 in the Appendix displays their summary statistics, showing that average satisfaction with wages or earnings is lower than other satisfaction averages. Table 2.B.2 shows that they are positively correlated among each other and with the other five variables used to construct the index, but the correlations tend to be somewhat weaker than those among the five original variables. Table 2.B.3 presents the new factor loadings, showing that the three new variables give positive but lower weights, indicating that they are conceptually farther away from the underlying perceived job match quality. Figure 2.B.1 shows that the new measure constructed with extra components has a similar distribution as in Figure 2.1. The main estimation results using the new job match quality index are in column 3 of Table 2.7. The short-term effect of training is slightly smaller than in Table 2.5, but remains significant.

Next, we check which of the 5 components that we use to construct job match quality are driving the results. Table 2.A.10 shows results of the same model as in the last column of Table 2.5, separately for each of the 5 variable that are included in our composite measure of job match quality. Note that the variables are scaled differently, so all the estimated coefficients are about ten times larger. Training has a large and

³⁴ This is calculated as $\text{mean} - (\text{max} - \text{mean})$ of the dependent variable: $0.71058 - (1 - 0.71058) = 0.42116$.

significant effect on satisfaction metrics, especially for satisfaction with current work and satisfaction with career, but training has also a large and significant effect on educational match. This could be related to the fact that quite a few training programs are formal education (e.g. one year part-time vocational education program).

Table 2.7: Sensitivity Analysis

GMM-SYS	(1)	(2)	(3)	(4)	(5)
Dependent variable	Match quality (5 consecutive yrs)	Match quality (Truncated)	Match quality (8 components)	Match quality (Alternative treatment)	
Match quality (t-1)	0.374*** (0.044)	0.327*** (0.074)	0.400*** (0.047)	0.396*** (0.081)	0.381*** (0.099)
Match quality (t-2)	0.130*** (0.031)	0.120** (0.050)	0.155*** (0.036)	0.142*** (0.046)	0.133** (0.060)
Match quality (t-3)	0.041* (0.022)	0.027 (0.027)	0.060** (0.025)	0.046* (0.024)	0.045* (0.024)
Training	0.037* (0.021)	0.034* (0.019)	0.039** (0.018)		
Days of training				0.0002** (0.0001)	
Single training					0.047 (0.035)
Multiple training					0.035 (0.032)
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Other controls	Yes	Yes	Yes	Yes	Yes
Training endogenous	Yes	Yes	Yes	Yes	Yes
Specification tests	p-value	p-value	p-value	p-value	p-value
m2 test	0.131	0.294	0.277	0.126	0.198
m3 test	0.270	0.162	0.272	0.228	0.311
Sargan test	0.120	0.175	0.418	0.179	0.393
Individuals	1600	2188	2061	2336	2336
Observations	6154	6326	5901	6890	6890

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are WC-robust standard errors. P-value for the specification tests. All the model specifications and instrument choices are the same as those in the main estimation; see notes to Table 2.5.

The fifth robustness check is to use alternative treatment variables. In column 4 of Table 2.7, the variable “days of training” is constructed as the total days of all training programs that a worker received in the last 12 months. We assume the effect of training is linear in days and additive across multiple programs. The estimated coefficient shows that one day of training significantly increases the job match quality by 0.0002. On average, workers taking training take about 69 days, giving a much smaller effect ($0.0002 \times 69 \approx 0.014$) than the main estimate of 0.045 in Table 2.5. In additional analysis (not presented), we find that training programs lasting no longer than seven days (the majority of cases) are the most effective ones, indicating that the effect of training might be concave in the duration of the training.

In column 5, we construct two mutually exclusive dummy variables “receiving a single training program” and “receiving multiple training programs”. The estimated coefficients on these two variables are not significantly different from each other, which indicates that it is the participation in training or not that really matters and not the number of training programs taken.

The final robustness checks use alternative definitions of type of training, job change outcomes, and job match quality. In Table 2.A.5, we utilize information on multiple training programs to construct a non-mutually exclusive classification of training: For an individual who participates in job change training as well as other training, both types of training variable will take the value 1. In Table 2.A.6, job change outcome (O_{it}) is a continuous variable defined as the interaction of the job change incidence J_{it} with the improvement of job match quality ΔM_{it} (instead of its sign). In Table 2.A.7, we take the simple average of the five job-match related variables to construct an alternative index “new job match quality” instead of the index using factor analysis.³⁵ In all cases, the results and patterns are quite similar to those in the main analysis.

2.6 Conclusion

In recent years, researchers and policy makers have increasingly become aware of the importance of job match quality. Previous studies find a positive relation between training and job match quality. We add to the literature by estimating the causal effect

³⁵ For the 315 observations who have one or several missing values in these five job-match related variables, we take the average of the available variables.

of training on job match quality using a different identification strategy, exploiting the timing of events in a dynamic panel data framework. We find that training has a positive short-run effect of approximately 0.045 (30% of a standard deviation of the job match quality index) and a long-run effect of 0.106 (71% of a standard deviation). This result is mainly driven by training programs aimed at building up additional human capital. To investigate the role of job changes in explaining the effect of training on job match quality, we explicitly incorporate job changes in the model, and find that training for job change purposes immediately increases the probability of changing a job, while training for other purposes has no such positive effect. These immediate job changes are equally likely to end up in a worse matched new job and in a better-matched new job. For those who change jobs one year later, however, training for job change purposes tends to lead to a job with a better match quality.

Our findings confirm that training helps to improve job match quality. This opens possibilities for policy makers who aim at improving job match quality to design policies with respect to training. Well-designed policies that encourage training could be instrumental to reach the European Union's strategic goal to "improve job match quality." Our findings suggest that training programs aimed at investing in workers' human capital are the most efficient way.

Appendix 2.A Tables

Table 2.A.1: Panel Structure

Years of observation	Individual records		Years of observation	Individual records	
	Frequencies	Percentage		Frequencies	Percentage
2008-2015	812	16.55	2008-2012/2014-2015	37	0.75
2008-2009	554	11.29	2012-2013/2015	34	0.69
2014-2015	553	11.27	2008-2010/2012	33	0.67
2008-2010	357	7.28	2009-2015	31	0.63
2012-2015	268	5.46	2008-2009/2011	30	0.61
2008-2011	181	3.69	2010-2013	29	0.59
2010-2015	135	2.75	2011-2015	28	0.57
2008-2014	121	2.47	2008-2012/2015	23	0.47
2008-2013	119	2.43	2010-2014	22	0.45
2008-2012	118	2.41	2008-2010/2012-2013	21	0.43
2010-2011	101	2.06	2008/2010-2011	20	0.41
2012-2013	89	1.81	2008/2010-2015	20	0.41
2008-2010/2012-2015	69	1.41	2008-2012/2014	20	0.41
2013-2015	67	1.37	2008-2009/2011-2012	18	0.37
2008-2013/2015	66	1.35	2012/2014-2015	17	0.35
2009-2010	60	1.22	2008-2011/2013-2014	17	0.35
2008-2011/2013-2015	57	1.16	2008-2011/2013	15	0.31
2012-2014	55	1.12	2013-2014	14	0.29
2010-2012	43	0.88	2011-2012	13	0.27
2008-2009/2011-2015	38	0.77	2008-2009/2011-2013	13	0.27
other 141 Trajectories	587	11.97			
Total				N = 4905	100

Table 2.A.2: Validation of Job Match Quality

Dependent variable	(1)	(2)
	FE-Within Match quality	FE-Within On-Job-Search
Match quality		-0.436*** (0.020)
Field of study: General	-0.002 (0.006)	-0.011 (0.015)
Field of study: Education	0.002 (0.010)	-0.020 (0.024)
Field of study: Humanity	-0.018* (0.011)	-0.020 (0.027)
Field of study: Social science	0.002 (0.004)	-0.003 (0.011)
Field of study: Science and Technology	-0.003 (0.010)	-0.030 (0.024)
Field of study: Engineering	-0.007 (0.007)	-0.007 (0.017)
Field of study: Agriculture	0.015 (0.014)	-0.071** (0.034)
Field of study: Health	0.012* (0.007)	0.002 (0.018)
Field of study: Service	0.008 (0.005)	-0.002 (0.013)
Level of education: Low/mid secondary	-0.014 (0.013)	0.030 (0.033)
Level of education: High secondary level	-0.009 (0.014)	0.054 (0.035)
Level of education: MBO	-0.012 (0.013)	0.086*** (0.033)
Level of education: HBO	-0.017 (0.014)	0.086** (0.033)
Level of education: WO	-0.016 (0.015)	0.101*** (0.036)
Age	-0.004*** (0.001)	-0.014*** (0.004)
Age ²	0.000*** (0.000)	0.000** (0.000)
Disabled	-0.013 (0.010)	0.063** (0.025)
Self-employed	0.043***	0.095***

	(0.008)	(0.019)
Temporary job	-0.009**	0.028***
	(0.004)	(0.009)
Income level	0.005***	-0.008**
	(0.001)	(0.003)
Public sector	-0.000	-0.008
	(0.006)	(0.014)
Job sector: Mining	-0.003	-0.002
	(0.010)	(0.023)
Job sector: Utility	-0.015	-0.111***
	(0.016)	(0.040)
Job sector: Construction	0.004	-0.002
	(0.015)	(0.036)
Job sector: Retail	-0.008	-0.034
	(0.008)	(0.021)
Job sector: Hospitality and Catering	0.000	0.076**
	(0.013)	(0.031)
Job sector: Transportation	-0.004	0.009
	(0.011)	(0.027)
Job sector: Finance	-0.014	0.007
	(0.011)	(0.027)
Job sector: Business	-0.007	-0.037*
	(0.008)	(0.020)
Job sector: Government	0.042***	-0.026
	(0.010)	(0.024)
Job sector: Education	0.058***	0.028
	(0.011)	(0.027)
Job sector: Health	0.009	0.070***
	(0.008)	(0.020)
Job sector: Entertainment	0.011	-0.050**
	(0.010)	(0.025)
Supervision level: High academic level	0.030**	-0.045
	(0.013)	(0.031)
Supervision level: High supervision level	0.036***	-0.023
	(0.012)	(0.029)
Supervision level: Mid academic level	0.013	0.002
	(0.011)	(0.026)
Supervision level: Mid supervision level	-0.014	0.006
	(0.011)	(0.027)
Supervision level: Mental level	-0.029***	0.017
	(0.011)	(0.027)
Supervision level: Skill manual level	-0.038***	0.039

	(0.013)	(0.032)
Supervision level: Semi-skill level	-0.091***	0.002
	(0.013)	(0.031)
Supervision level: Unskill manual level	-0.187***	0.043
	(0.013)	(0.032)
Hours of working	0.000***	-0.000*
	(0.000)	(0.000)
Job tenure	-0.002***	0.002***
	(0.000)	(0.001)
Commuting time	-0.000**	0.001***
	(0.000)	(0.000)
Constant	0.819***	0.711***
	(0.036)	(0.089)
<hr/>		
Individuals	4,679	4,679
Observations	20,380	20,380
R-squared	0.049	0.046
<hr/>		

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. Outcome variable “On job search” is a binary indicator for being employed, but looking for another job. Both specifications use within group individual fixed effects estimators.

Table 2.A.3: The Effect of Training on Job Match Quality
(Complete Version of Table 2.5)

Methods	(1) OLS	(2) FD-FE-	(3) OLS	(4) FD-FE-	(5) GMM-	(6) GMM-
Match quality (t-1)					0.398*** (0.043)	0.391*** (0.044)
Match quality (t-2)					0.141*** (0.031)	0.137*** (0.032)
Match quality (t-3)					0.045** (0.023)	0.046** (0.023)
Training	0.032*** (0.002)	0.008*** (0.002)	0.028*** (0.003)	0.005* (0.003)	0.007* (0.004)	0.045** (0.019)
Female	0.002 (0.002)		0.003 (0.003)			
Age	-0.002** (0.001)	-0.006* (0.003)	- (0.001)	-0.001 (0.004)	-0.004 (0.005)	-0.003 (0.005)
Age ²	0.000*** (0.000)	0.000* (0.000)	0.000*** (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Disabled	- (0.010)	-0.025* (0.014)	-0.041** (0.021)	-0.056*** (0.019)	-0.056* (0.030)	-0.053* (0.030)
Year 2009	0.005 (0.004)	0.007*** (0.002)				
Year 2010	-0.001 (0.004)	-0.004* (0.002)				
Year 2011	0.000 (0.004)	0.000 (0.003)				
Year 2012	0.001 (0.004)	0.000 (0.002)	0.007 (0.005)		0.001 (0.003)	0.001 (0.003)
Year 2013	0.003 (0.004)	0.001 (0.002)	0.009** (0.004)	0.002 (0.002)	0.006** (0.002)	0.004 (0.003)
Year 2014	-0.009** (0.004)	-0.009*** (0.002)		-0.008*** (0.002)		
Year 2015	-0.002 (0.004)		0.007 (0.005)		0.011*** (0.004)	0.008** (0.004)
Constant	0.696*** (0.015)	0.838*** (0.070)	0.759*** (0.027)	0.738*** (0.107)	0.449*** (0.138)	0.382*** (0.146)
Specification tests					p-value	p-value
m2 test					0.102	0.141
m3 test					0.276	0.295
Sargan test					0.081	0.186
Individuals	4,896	4,896	1,900	1,900	2,336	2,336
Observations	21,677	21,677	7,077	7,077	6,890	6,890
R-squared	0.024		0.020			

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are robust standard errors. For the specification tests, the p values are reported.

Table 2.A.4: Estimated Coefficient of Training with Alternative Instrument Choice

GMM-SYS		Instruments of Lagged Dep. Var.					
		2 to 2	2 to 3	2 to 4	2 to 5	2 to 6	2 to 7
Instruments of Lagged Training	2 to 2	0.035	0.038*	0.036*	0.042**	0.045**	0.045**
	2 to 3	0.030	0.030	0.030	0.035*	0.037**	0.036**
	2 to 4	0.031*	0.032*	0.031*	0.037**	0.038**	0.038**
	2 to 5	0.028	0.029*	0.029	0.034**	0.035**	0.035**
	2 to 6	0.029	0.029*	0.028	0.033*	0.033*	0.033**
	2 to 7	0.029	0.029*	0.028	0.034**	0.034**	0.033**

Note:*Significant at 10%; ** at 5%; *** at 1%. The varying instruments are for the differenced equation. The moments for the level equation are lagged differenced training and lagged differenced match quality all the time.

Table 2.A.5: Non-mutually Exclusive Classification of Training

	(1)	(2)	(3)
Dependent variable	Match Quality	Job Change Incidence	Job Change Outcome
Match quality (t-1)	0.378*** (0.075)	-0.230*** (0.051)	-0.330*** (0.055)
Match quality (t-2)	0.132*** (0.046)		
Match quality (t-3)	0.046** (0.023)		
Human capital build-up training	0.043* (0.025)	-0.068 (0.050)	0.015 (0.058)
Human capital build-up training (t-1)			-0.001 (0.011)
Job change training	0.007 (0.048)	0.216* (0.123)	-0.193* (0.104)
Job change training (t-1)			0.066** (0.028)
Other training	-0.009 (0.027)	0.120*** (0.046)	0.034 (0.049)
Other training (t-1)			-0.003 (0.012)
Time fixed effects	Yes	Yes	Yes
Other controls	Yes	Yes	Yes
Training endogenous	Yes	Yes	Yes
Specification tests	p-value	p-value	p-value
m2 test	0.131	0.345	0.702
m3 test	0.303	0.691	0.718
Sargan test	0.776	0.154	0.205
Individuals	2,336	4867	4,823
Observations	6,890	15666	15,495

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are WC-robust standard errors. For the specification tests, the p values are reported. The estimation strategy, other controls, and instrument choice are the same as those in Table 2.6 except that in column (1) M_{it-2} to M_{it-5} are used as instruments for the differenced equation.

Table 2.A.6: Alternative Definition of Job Change Outcome

Dependent variable	Job change outcome
Match quality (t-1)	-0.066*** (0.020)
Human capital build-up training	0.002 (0.008)
Human capital build-up training (t-1)	0.001 (0.002)
Job change training	-0.009 (0.017)
Job change training (t-1)	0.012** (0.005)
Other training	0.009 (0.009)
Other training (t-1)	0.000 (0.002)
Time fixed effects	Yes
Other controls	Yes
Training endogenous	Yes
Specification tests	p-value
m2 test	0.444
m3 test	0.344
Sargan test	0.544
Individuals	4,823
Observations	15,495

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are WC-robust standard errors. Job change outcome (O_{it}) is a continuous variable defined as the interaction of an indicator for changing jobs J_{it} and the improvement of job match quality ΔM_{it} . For the specification tests, the p values are reported. The estimation strategy, other controls, and instrument choice are the same as those in column (3) of Table 2.6.

Table 2.A.7: Constructing Job Match Quality with Simple Average of Components

	(1)	(2)	(3)	(4)	(5)
Dependent variable	New Match Quality	New Match Quality	Job Change Incidence	Job Change Outcome	New Job Change
Match quality (t-1)	0.374*** (0.043)	0.365*** (0.043)	-0.024*** (0.006)	-0.035*** (0.009)	-0.053*** (0.018)
Match quality (t-2)	0.139*** (0.030)	0.136*** (0.030)			
Match quality (t-3)	0.055*** (0.021)	0.053** (0.021)			
Training	0.389** (0.166)				
Human capital build-up training		0.589*** (0.209)	-0.080 (0.051)	-0.027 (0.067)	-0.045 (0.075)
Human capital build-up training (t-1)				0.008 (0.014)	0.017 (0.015)
Job change training		0.243 (0.409)	0.221* (0.124)	-0.150 (0.114)	-0.087 (0.130)
Job change training (t-1)				0.086*** (0.030)	0.079* (0.041)
Other training		0.260 (0.220)	0.102* (0.054)	0.038 (0.074)	-0.002 (0.091)
Other training (t-1)				0.010 (0.015)	0.019 (0.017)
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Other controls	Yes	Yes	Yes	Yes	Yes
Training endogenous	Yes	Yes	Yes	Yes	Yes
Specification tests	p-value	p-value	p-value	p-value	p-value
m2 test	0.440	0.418	0.282	0.454	0.952
m3 test	0.569	0.518	0.385	0.396	0.452
Sargan test	0.339	0.789	0.218	0.123	0.319
Individuals	2,404	2,404	4,905	4,905	4,905
Observations	7,131	7,131	15,865	15,865	15,865

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are WC-robust standard errors. For the specification tests, the p values are reported. The newly constructed dependent variable job match quality ranges from 0 to 10. The estimation strategy, other controls, and instrument choice are the same as those in main estimations. Column (5) is using job change incidence interacting with the improvement of job match quality ΔM_{it} to construct job change outcome O_{it} .

Table 2.A.8: The Panel Structure with Alternative Sample Restriction

Years of observation	Individual records		Years of observation	Individual records	
	Frequencies	Percentage		Frequencies	Percentage
2008-2015	812	49.66	2008-2012/2015	23	1.41
2010-2015	135	8.26	2010-2014	22	1.35
2008-2014	121	7.40	2008/2010-2015	20	1.22
2008-2013	119	7.28	2008-2012/2014	20	1.22
2008-2012	118	7.22	2009-2014	12	0.73
2008-2013/2015	66	4.04	2009/2011-2015	8	0.49
2008-2009/2011-2015	38	2.32	2008/2011-2015	7	0.43
2008-2012/2014-2015	37	2.26	2009-2013	6	0.37
2009-2015	31	1.90	2009-2013/2015	6	0.37
2011-2015	28	1.71	2008/2010-2014	6	0.37
Total				N = 1635	100

Table 2.A.9: Descriptive Statistics for the SYS-GMM Sample

Variables	Mean	Std. Dev
<i>Dependent variable</i>		
Job match quality	0.717	0.135
Job change incidence	0.049	0.215
<i>Treatment</i>		
Training	0.341	0.474
Total days of training per year (conditional on participating in training)	59.423	112.416
Single training	0.204	0.403
Multiple training	0.136	0.343
<i>Demographics</i>		
Female	0.502	0.500
Age in years	46.237	10.755
Disabled	0.007	0.086
Individuals	1900	
Observations	7077	

Table 2.A.10: Effect of Training on Each Match-Related Variable

	(1)	(2)	(3)	(4)	(5)
Dependent variable	Educational Match	Skill Match	Satisfaction with type of work	Satisfaction with career	Satisfaction with current work
DV (t-1)	0.205*** (0.054)	0.167*** (0.048)	0.345*** (0.041)	0.298*** (0.044)	0.314*** (0.041)
DV (t-2)	0.105*** (0.038)	0.051 (0.034)	0.126*** (0.033)	0.138*** (0.035)	0.118*** (0.032)
DV (t-3)	0.065*** (0.023)	0.055** (0.027)	0.047* (0.024)	0.083*** (0.024)	0.044* (0.025)
Training	0.644** (0.278)	-0.008 (0.277)	0.369 (0.258)	0.414* (0.222)	0.600*** (0.231)
Time fixed effects	Yes				
Other controls	Yes				
Training endogenous	Yes				
Specification tests	P-value				
M2 test (p-value)	0.842	0.342	0.066	0.561	0.765
M3 test (p-value)	0.702	0.373	0.337	0.209	0.443
Sargan test (p-value)	0.380	0.082	0.146	0.061	0.416
Individuals	2,404	2,404	2,358	2,351	2,363
Observations	7,131	7,131	6,968	6,936	6,989

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are WC-robust standard errors. Specifications and estimation method are the same as Table 2.5 Column (6) except that the dependent variable are the original match-related variables.

Appendix 2.B Alternative Way to Construct Job Match Quality

Table 2.B.1: Extra Job Satisfaction Variables (Scale: 0-9)

Variable (Corrected)	Obs	Mean	Std. Dev
Satisfaction with wage	21573	5.738	1.808
Satisfaction with working hours	21741	6.543	1.620
Satisfaction with atmosphere	19879	6.616	1.452

Table 2.B.2: Correlations of Job-Match Related Variables

	Education	Skill	Type of work	Career	Current work	Wage	Working hours	Atmosphere
Education	1.000							
Skill	0.604	1.000						
Type of work	0.348	0.431	1.000					
Career	0.357	0.418	0.691	1.000				
Current work	0.291	0.379	0.822	0.736	1.000			
Wage	0.218	0.245	0.396	0.496	0.454	1.000		
Working hours	0.145	0.197	0.418	0.401	0.456	0.373	1.000	
Atmosphere	0.166	0.233	0.487	0.493	0.572	0.279	0.339	1.000

Table 2.B.3: Factor Loadings

	Factor loadings
Educational match	0.071
Skill match	0.084
Satisfaction of type of work	0.222
Satisfaction with career	0.258
Satisfaction with current work	0.369
Satisfaction with wage	0.070
Satisfaction with working hours	0.068
Satisfaction with atmosphere	0.057

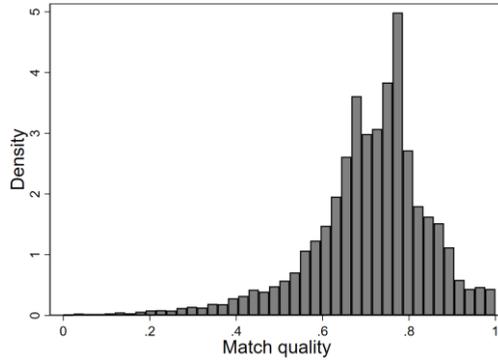


Figure 2.B.1: Distribution of Alternative Job Match Quality Index

References for Chapter 2

- Allen, Jim, and Rolf Van der Velden. "Educational Mismatches versus Skill Mismatches: Effects on Wages, Job Satisfaction, and On-the-Job Search." *Oxford Economic Papers* 53.3 (2001): 434-452.
- Arellano, Manuel, and Stephen Bond. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58.2 (1991): 277-297.
- Barmby, Tim, Alex Bryson, and Barbara Eberth. "Human Capital, Matching and Job Satisfaction." *Economics Letters* 117.3 (2012): 548-551.
- Becker, Gary S. "Investment in Human Capital: A Theoretical Analysis." *The Journal of Political Economy* 70.5 (1962): 9-49.
- Blundell, Richard, and Stephen Bond. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87.1 (1998): 115-143.
- Blundell, Richard, Monica Costa Dias, Costas Meghir, and John Van Reenen. "Evaluating the Employment Impact of a Mandatory Job Search Program." *Journal of the European Economic Association* 2.4 (2004): 569-606.
- Burgard, Claudia, and Katja Görlitz. "Continuous Training, Job Satisfaction and Gender: An Empirical Analysis Using German Panel Data." *Evidence-Based HRM: a Global Forum for Empirical Scholarship*. Vol. 2. No. 2. Emerald Group Publishing Limited, 2014.
- Cheng, Ying, and Franz Waldenberger. "Does Training Affect Individuals' Turnover Intention? Evidence from China." *Journal of Chinese Human Resources Management* 4.1 (2013): 16-38.
- Chiang, Chun-Fang, Ki-Joon Back, and Deborah D. Canter. "The Impact of Employee Training on Job Satisfaction and Intention to Stay in the Hotel Industry." *Journal of Human Resources in Hospitality & Tourism* 4.2 (2005): 99-118.
- Child, Dennis. *The Essentials of Factor Analysis*. 3rd ed. London: Continuum, 2006.
- Clark, Andrew E. "What Really Matters in a Job? Hedonic Measurement Using Quit Data." *Labour Economics* 8.2 (2001): 223-242.
- Clark, Andrew E. "Your Money or Your life: Changing Job Quality in OECD Countries." *British Journal of Industrial Relations* 43.3 (2005): 377-400.
- Clark, Andrew E. "What Makes a Good Job? Job Quality and Job Satisfaction." *IZA World of Labor*, Article 215 (2015).
- Congregado, Emilio, Jesús Iglesias, José María Millán, and Concepción Román. "Incidence, Effects, Dynamics and Routes out of Overqualification in Europe: a Comprehensive Analysis Distinguishing by Employment Status." *Applied Economics* 48.5 (2016): 411-445.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *The Quarterly Journal of Economics* 128.2 (2013): 531-580.

Dekker, Ron, Andries De Grip, and Hans Heijke. "The Effects of Training and Overeducation on Career Mobility in a Segmented Labour Market." *International Journal of Manpower* 23.2 (2002): 106-125.

Dolton, Peter, and Anna Vignoles. "The Incidence and Effects of Overeducation in the UK Graduate Labour Market." *Economics of Education Review* 19.2 (2000): 179-198.

Ferreira, Priscila, and Mark Taylor. "Measuring Match Quality Using Subjective Data." *Economics Letters* 113.3 (2011): 304-306.

Gaure, Simen, Knut Røed and Lars Westlie, "Job Search Incentives and Job Match Quality." *Labour Economics* 19.3: 438-450.

Georgellis, Yannis, and Thomas Lange. "Participation in Continuous, On-The-Job Training and the Impact on Job Satisfaction: Longitudinal Evidence from the German Labour Market." *The International Journal of Human Resource Management* 18.6 (2007): 969-985.

Gottschalk, Peter, and Tim Maloney. "Involuntary Terminations, Unemployment, and Job Matching: A Test of Job Search Theory." *Journal of Labor Economics* 3.2 (1985): 109-123.

Green, Francis, and Yu Zhu. "Overqualification, Job Dissatisfaction, and Increasing Dispersion in the Returns to Graduate Education." *Oxford Economic Papers* 62.4 (2010): 740-763.

Hair, Joseph F., Ronald L. Tatham, Rolph E. Anderson, and William C. Black. *Multivariate Data Analysis*. 5th [rev.] Ed., International ed. London: Prentice-Hall International, 1998.

Harter, James K., Frank L. Schmidt, and Theodore L. Hayes. "Business-Unit-Level Relationship between Employee Satisfaction, Employee Engagement, and Business Outcomes: a Meta-Analysis." *Journal of Applied Psychology* 87.2 (2002): 268.

Han, Kihye, Alison M. Trinkoff, Carla L. Storr, Nancy Lerner, Meg Johantgen, and Kyungsook Gartrell. "Associations between State Regulations, Training Length, Perceived Quality and Job Satisfaction among Certified Nursing Assistants: Cross-sectional Secondary Data Analysis." *International Journal of Nursing Studies* 51.8 (2014): 1135-1141.

Jones, Melanie K., Richard J. Jones, Paul L. Latreille, and Peter J. Sloane. "Training, Job Satisfaction, and Workplace Performance in Britain: Evidence from WERS 2004." *Labour* 23.s1 (2009): 139-175.

Jovanovic, Boyan. "Job Matching and the Theory of Turnover." *Journal of Political Economy* 87.5, Part 1 (1979): 972-990.

Kalleberg, Arne L., and Stephen Vaisey. "Pathways to a Good Job: Perceived Work Quality among the Machinists in North America." *British Journal of Industrial Relations* 43.3 (2005): 431-454.

- Lachowska, Marta, Merve Meral and Stephen A. Woodbury. "Effects of the Unemployment Insurance Work Test on Long-Term Employment Outcomes." *Labour Economics* 41.3 (2016): 246-265.
- Le Barbanchon, Thomas. "The Effect of the Duration of Unemployment Benefits on Unemployment Exits to Work and Match Quality in France." *Labour Economics* 42.3 (2016): 16-29.
- Mavromaras, Kostas, Seamus McGuinness, Nigel O'Leary, Peter Sloane, and Zhang Wei. "Job Mismatches and Labour Market Outcomes: Panel Evidence on University Graduates." *Economic Record* 89.286 (2013): 382-395.
- Messinis, George, and Nilss Olekalns. "Returns to Training and Skill Mismatch: Evidence from Australia." *CSES Working Paper* No. 40. Victoria: Victoria University, 2008.
- Mortensen, Dale T. "Specific Capital and Labor Turnover." *The Bell Journal of Economics* 9.2 (1978): 572-586.
- Nordin, Martin, Inga Persson, and Dan-Olof Rooth. "Education-Occupation Mismatch: Is There an Income Penalty?" *Economics of Education Review* 29.6 (2010): 1047-1059.
- Pagán-Rodríguez, Ricardo. "Disability, Training and Job Satisfaction." *Social Indicators Research* 122.3 (2015): 865-885.
- Pecoraro, Marco. "Is There Still a Wage Penalty for Being Overeducated But Well-Matched in Skills? A Panel Data Analysis of a Swiss Graduate Cohort." *Labour* 28.3 (2014): 309-337.
- Sicherman, Nachum and Oded Galor. "A Theory of Career Mobility." *Journal of Political Economy* 98.1 (1990): 169-192.
- Stevens, James P. *Applied Multivariate Statistics for the Social Sciences*. 2nd ed. Hillsdale, NJ: Erlbaum, 1992.
- Topel, Robert H., and Michael P. Ward. "Job Mobility and the Careers of Young Men." *The Quarterly Journal of Economics* 107.2 (1992): 439-479.
- Vahey, Shaun P. "The Great Canadian Training Robbery: Evidence on the Returns to Educational Mismatch." *Economics of Education Review* 19.2 (2000): 219-227.
- Windmeijer, Frank. "A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators." *Journal of Econometrics* 126.1 (2005): 25-51.
- Zhou, Ying, Min Zou, Mark Williams, and Vurain Tabvuma. "Is the Grass Greener on the Other Side? A Longitudinal Study of the Impact of Employer Change and Occupational Change on Job Satisfaction." *Journal of Vocational Behavior* 99 (2017): 66-78.

Chapter 3

The Effect of Retirement on Healthcare Utilization: Evidence from China³⁶

3.1 Introduction

China is aging rapidly. The number of persons above age 65 grew from 100 million in 2005 to 143 million in 2015 (National Bureau of Statistics of China 2016). At the same time, the Chinese statutory retirement ages (SRAs) of 60 years for men and 50 or 55 years for women are among the lowest in the world. The increasing number of retired people imposes large costs on public and private budgets. This has led to an ongoing debate about increasing retirement ages. In order to understand the full consequences of retirement policies, we have to take into account the effect of retirement on healthcare utilization. For example, if retirement increases healthcare utilization then this further adds to the costs of retirement.

In this study, we investigate the effect of retirement on healthcare utilization in urban China. The direction of this effect is far from obvious. Economic theory makes ambiguous predictions (Galama et al. 2013, Kuhn et al. 2015). On the one hand, retirement might increase healthcare utilization because of reduced time cost of using healthcare services or a negative effect of retirement on health (e.g. cognitive decline, obesity). On the other hand, retirement might decrease healthcare utilization because of lower income after retirement in the presence of high co-payments, or because people switch to a healthier lifestyle.³⁷

³⁶ This chapter is the same as the published version in *Journal of Health Economics*. It is coauthored with Martin Salm and Arthur van Soest. The authors thank the China Health and Retirement Longitudinal Study (CHARLS) team for providing the data. We are very grateful for many helpful comments of the editor and two anonymous referees. We also thank Vellore Arthi, Bart Bronnenberg, Meltem Daysal, Max Groneck, Jan Kabátek, Peter Eibich, Tunga Kantarci, Tobias Klein, Lei Lei, Carol Propper, Renata Rabovič, Hans Henrik Sievertsen, Moritz Suppliet, Agne Suziedelyte, Yan Xu, seminar participants at Tilburg University, and conference participants at the 4th SDU Workshop on Applied Microeconomics, the 31st Annual Conference of ESPE, Netspar Pension Day 2017, and the 12th European Conference on Health Economics (EuHEA) for their helpful comments and suggestions.

³⁷ A growing literature discusses the causal effect of retirement on physical and mental health (e.g. Mein et al. 2003, Neuman 2008, Lei et al. 2011, Coe and Zamarro 2011, Behncke 2012, Bonsang et al. 2012, Coe et al. 2012, Hernaes et al. 2013, van der Heide 2013, Atalay and Barrett 2014, Insler 2014, Iparraguirre 2014, Eibich 2015, Che and Li 2018, Hagen 2018, Shai 2018), as well as the effect

The empirical evidence for the effect of retirement on healthcare utilization is also mixed. Previous studies based on data from developed countries find either negative effects (Bejarano et al. 2014, Eibich 2015, Coe and Zamarro 2015a, Grötting and Lillebø 2017, Shai 2018), or no significant effects (Laaksonen et al. 2012, Fé and Hollingsworth 2012, Hagen 2018).

A fundamental challenge in estimating the causal effect of retirement on healthcare utilization is that retirement can be endogenous to health, and therefore also to healthcare utilization. People in poor health might be more likely to retire early. One approach to address this endogeneity problem is to use SRAs as a source of exogenous variation in retirement (Neuman 2008, Lei et al. 2011, Bonsang et al. 2012, Insler 2014, Eibich 2015, Coe and Zamarro 2015b, Godard 2016). This approach can be implemented with a fuzzy regression discontinuity design, comparing individuals of ages just below and just above their SRA.

We employ this method to examine the effect of retirement on healthcare utilization based on the 2011 and 2013 waves of the “China Health and Retirement Longitudinal Study” (CHARLS). We exploit the discontinuity in retirement rates at the SRA in urban China. At the SRA, the probability of being retired increases by around 30 percentage points. We can exclude the possibility that changes in healthcare utilization at the SRA are due to factors other than retirement, such as changes in health insurance. Therefore, the assumptions of a fuzzy regression discontinuity approach are met allowing us to estimate the causal effect of retirement on healthcare utilization.

We find that retirement increases healthcare utilization. Specifically, retirement significantly increases the number of doctor visits, the number of hospital stays, yearly out of pocket expenditures for inpatient care, and monthly out of pocket expenditures for self-treatment. This finding is robust to alternative specifications such as different parametric functional forms of age or different age bandwidths for choosing the sample. For men, we also find a marginally significant positive effect of retirement on the incidence of outpatient care and a strong and significant positive effect on out-of-pocket inpatient cost.

of retirement on health behaviors (e.g. Lang et al. 2007, Zantinge et al. 2014, Bertoni et al. 2016, Coe and Zamarro 2015b, Kim et al. 2016, Godard 2016). Retirement can have health-preserving and health-damaging effects and the evidence on which effect dominates is inconclusive.

To better understand our findings, we explore three possible mechanisms. The first possible explanation is deteriorating health after retirement. We find negative effects on objective measures of physical functioning and an increase in self-reported incidence of chronic diseases. A second explanation could be the reduced opportunity cost of time after retirement. We find that the increase in inpatient care use is mainly driven by retirees who previously worked in the private sector, where it might be more difficult to take time off for medical care. This relates to China's institutional feature that employment protection practices are much less generous compared to developed countries, leading to higher opportunity costs of time. A third potential mechanism relates to income, but we do not find a significant income drop at retirement.

For the low educated, there is no income drop at retirement either, but we do find that retirement leads to a significant increase in the likelihood of foregoing inpatient care of 20%-points, even though it was suggested by a physician. Our interpretation is that for this group, retirement releases the time constraint but financial barriers, in particular the high copayments that form another institutional feature typical for China, still prevents many people from using care. This finding implies that policymakers should pay attention to keeping medical care affordable for retirees with lower socio-economic status.

Our study contributes to the growing literature on the effects of retirement on health and healthcare utilization. To the best of our knowledge, it provides the first evidence of an effect of retirement on healthcare utilization from a developing country.³⁸ Our results strongly differ from existing evidence for developed countries. The analysis of the mechanisms underlying the effect of retirement on healthcare utilization in relation to institutional characteristics will also be informative for other developing countries with low employment protection and high co-payments.

The remainder of the paper is organized as follows: Section 3.2 introduces the institutional background. Section 3.3 describes the data. Section 3.4 presents the empirical strategy. Section 3.5 shows our main results, and Section 3.6 explores the possible mechanisms. Section 3.7 provides sensitivity analyses, while Section 3.8 discusses the role of institutions and policy implications.

³⁸ Lei et al. (2011) examine the effect of retirement on health in China.

3.2 Institutional Background and International Comparison

Statutory Retirement Ages in Urban China

The statutory (full) retirement age in China is 60 years for men, 55 years for female civil servants, and 50 years for other female employees. China has the lowest retirement ages in the world, even though its population is aging fast as a result of birth control policies and increasing life expectancy. For historical reasons, SRAs only apply to urban China.³⁹ Retirement arrangements were introduced to protect urban employees in the 1950s when the only employers were either the government or state-owned companies and institutions. Private sector and self-employment entered after the economic reforms in the 1980s. Retirement arrangements were adapted to cover urban workers in these “new” sectors, but still do not apply to rural China. Farmers usually continue working as long as their health permits. In this study, we therefore restrict our analysis to urban residents.

In principle, employees are required to retire at their SRA, but deviations are possible: (1) Employees are allowed to retire five years earlier than the full retirement age if their jobs are dangerous or harmful to health, or if a medical exam proves that they are too ill to continue working.⁴⁰ (2) Retirement at the SRA is not as strictly enforced in the private sector, self-employment, and temporary employment as in the public sector and state-owned companies. Therefore, “compliance” with the SRA is not perfect: a substantial number of people still works for pay after reaching the SRA. However, as we show later in this study, we do observe a discontinuity in the retirement rate at this age.

Pension and “Processed Retirement”

Urban employees are required to participate in pension programs. This policy is strictly enforced in the public sector, state-owned enterprises, and big companies in the private sector. Deviations exist in small private companies and in informal employment.

Employees are eligible to claim a pension when they reach their SRA and “process” retirement. The pension income varies in amount and composition, depending on

³⁹ In 2011 the labor force in China included 359 million people in urban areas and 405 million people in rural areas (National Bureau of Statistics of China 2016).

⁴⁰ In our sample, around 22% of retirees took early retirement.

pension program, years of contribution, and occupation. The actual pension income can be lower or higher than the pre-retirement wage.

“Processed retirement” means that an employee reaching the SRA leaves the current job after going through all the formalities with employer and local government. A difference from many other countries is that people can still continue working after “processing retirement”. They can work for a new employer or even for the former employer with a temporary contract, while at the same time claiming pension (and health insurance benefits) from the former employer. This fact complicates the definition of retirement, which we will further discuss in Section 3.3.

Health Insurance System

Health insurance in urban China is organized independently from retirement arrangements: Eligibility for public health insurance does not depend on retirement status or pension claiming. Public health insurance programs cover more than 95% of the overall population. Private health insurance programs are much less prevalent.⁴¹ There are different types of public health insurance programs, as seen in Figure 3.B.1.⁴² An individual’s type depends on occupation and residential status, and it is not easy to switch types.

Health insurance programs differ in generosity, but generally require high patient copayments. The two major urban insurance programs, covering 75% of our sample, are the Urban Employee’s Basic Medical Insurance (UEBMI) for urban employees and the Urban Resident Basic Medical Insurance (URBMI) for urban residents without formal employment. These plans have copayments of at least 35% and at least 45% for inpatient care, respectively.

Despite differences across insurance programs, the covered benefits and the copayment rates within the same health insurance program remain the same before and after retirement. Moreover, individuals do not change program when they retire.

International Comparison

⁴¹ In our sample, coverage by private health insurance programs is about 5%. Such programs usually provide supplementary insurance to public health insurance plans. Eligibility and benefits are independent of retirement status.

⁴² Figures 3.B.1 – 3.B.10 are in the Appendix.

Table 3.1 compares labor market characteristics in urban China with selected developed countries (US, Germany) and less developed countries (Malaysia and Brazil). In urban China, sometimes referred to as “the world’s factory”, the share of employment in mining, manufacturing, utilities and construction sectors was above 50% in 2012, higher than in any of the other countries. Compared to the US and Germany, working hours in urban China are substantially longer, indicating that visiting a clinic or hospital will be difficult without absence from work, creating a potential impediment to seeking healthcare before retirement.

Table 3.1: International Comparison on Labor Market Characteristics

	Employment in sectors of mining, manufacturing, utilities, and construction	Hours of Work	Monthly earnings	Share of out-of-pocket expenditure	Share of out-of-pocket health expenditure per capita
	%	Weekly	Constant 2011 PPP \$	% of total health expenditure	% of monthly earnings
Urban China	52.75	46	1106	55.3	6.4
The United States	18.50	36	4417(in 2010)	11.7	1.8 (in 2010)
Germany	30.20	36	4762	11.9	1.1
Malaysia	28.57	46	1317	32.7	1.9
Brazil	23.01	39	907	30.6 (in 2009)	3.3
Year	2012	2012	2012	2010	2012

Notes: Years are chosen as close as possible to 2012, given data availability. Figures are from International Labour Organization (2018) unless mentioned otherwise. In column (4), “Share of out-of-pocket expenditure” refers to out-of-pocket expenditure as a percentage of total health expenditure. Monthly earnings in US dollars are converted using 2011 purchasing power parities (PPPs). For urban China, employment in sectors of mining, manufacturing, utilities, and construction is computed using figures of “Number of Employed Persons in Urban Units” and “Number of Engaged Persons in Urban Private Enterprises and Self-employed Individuals” in 2012 from National Bureau of Statistics of China (2016). In column (5), “share of out-of-pocket health expenditure per capita” refers to out-of-pocket health expenditure per capita as a percentage of monthly earnings. For urban China, this figure is computed using figures of monthly “per capita health expenditure in urban areas (yuan)” in 2012 from National Bureau of Statistics of China (2016), divided by “monthly earning in urban China (yuan)” in 2012 from International Labour Organization (2018). For the other countries, “share of out-of-pocket health expenditure per capita” is computed using “out-of-pocket expenditure as percentage of current health expenditure (CHE)” and monthly “current health expenditure (CHE) per capita in US\$” from World Health Organization (2018), together with “monthly earnings in in US\$” from International Labour Organization (2018).

Wages in urban China are similar to those in Malaysia and Brazil, but much lower than in the US or Germany. After adjusting for purchasing power differences,

monthly earnings are around one fourth of those in the US and Germany. Due to high copayments, the share of healthcare expenditures that patients pay out of pocket is high, about 5 times larger than in the US and Germany and almost twice that in Malaysia or Brazil. Accordingly, the share of out-of-pocket expenditure as a percentage of monthly earnings in urban China is also much higher than in the other countries. This may be another reason for not seeking health care, both before and after retirement.

3.3 Data

The data we use come from “The China Health and Retirement Longitudinal Study” (CHARLS).⁴³ This dataset is ideal for our study because it collects rich information about retirement, healthcare utilization, health, income and expenditures, health insurance status, and demographic characteristics, from a nationally representative sample of 17,500 individuals aged 45 or older and their spouses. Interviews are repeated every two years. We use information from the first two waves of the survey in 2011 and 2013.

We restrict the sample to urban residents aged between 40 and 75.⁴⁴ This reduces the sample size from 36,338 to 7,286 individual-year observations. We further exclude the self-employed since they do not process retirement, and we exclude those who neither work nor report “processed retirement”.⁴⁵ ⁴⁶ This reduces the sample size to 5,438 observations. Finally, we exclude observations with missing information on retirement status, age, or gender, leaving us with a sample of 5,178 individual-year observations for 3,511 individuals. Around half of the individuals in our sample are observed in both waves (see Table 3.2). In Section 3.7 we test whether

⁴³ CHARLS is harmonized with the Health and Retirement Study (HRS), the English Longitudinal Study of Aging (ELSA), and the Survey of Health, Aging and Retirement in Europe (SHARE). For more details, see <http://charls.ccer.edu.cn/en/page/about/CHARLS>.

⁴⁴ We define urban and rural using “Hukou”, a household registration system that distinguishes “Urban Hukou” and “Rural Hukou”.

⁴⁵ “Working” refers to either engaging in agricultural work (including farming, forestry, fishing, and husbandry for one’s own family or others) for more than 10 days per year, or performing any of the following activities for at least one hour per week: earning a wage, running one’s own business or helping in a family business, etc. “Working” does not include activities without pay, such as housework or voluntary work.

⁴⁶ Those who neither work nor report “processed retirement” account for around 1/6 of the sample.

the results are sensitive to sample attrition by excluding those who only appear in the first wave.

Table 3.2: Panel Structure

Years of observation	Individual records	
	Frequencies	Percentage
2011 and 2013	1667	47.48
2011	825	23.50
2013	1019	29.02
Total	N=3511	100

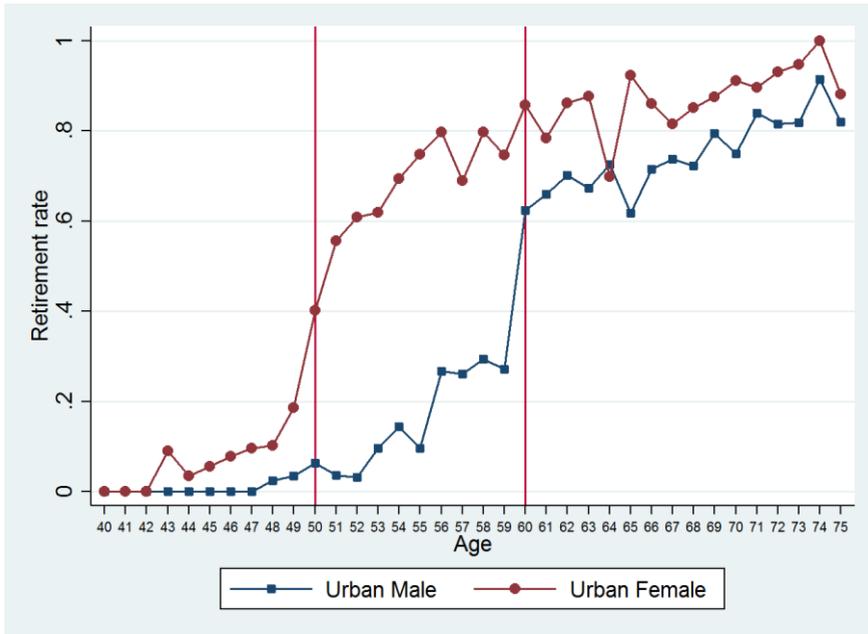
Retirement

There are three common ways to define retirement: (1) self-reported “processed retirement”. This is not ideal because many people still work for pay after processing retirement from their “career jobs” (20% of individuals in our sample continue working after processing retirement). (2) Neither working for pay nor searching for a paid job. This is not ideal either, because some people have never had a paid job at all and did not “retire” from work. (3) Both “processed retirement” and not working for a paid job anymore. We adopt this third definition because we consider “retirement” as a change from working to non-working. In additional analyses, we show how alternative definitions of retirement influence the results.

Normalized Age

Figure 3.1 shows the “retirement rate” (the sample fraction of individuals in retirement) by age and gender. It shows a clear discontinuity at age 50 for women and at age 60 for men, the SRAs. There are also substantial increases in retirement rates at other ages (e.g. 51 for women and 56 for men), but these are not related to formal retirement rules.

For female civil servants, the SRA is 55. In Figure 3.1 we do not see a clear discontinuity in retirement rates for women at age 55, since civil servants are a relatively small group, and the distinction between civil servants and other public sector employees is poorly measured in our data. We therefore proxy the SRA to age 50 for all women.



Note: The vertical lines at ages 50 and 60 are the SRAs for female and male workers.

Figure 3.1: Retirement Rate by Age

Workers can retire (at most) 5 years before the SRA if their jobs are dangerous, health-damaging, or extremely onerous. Therefore early retirement starts at age 45 for female workers and 55 for male workers. While we do not use early retirement ages in our main analysis, we add a dummy for having reached the early retirement age as an additional instrument as a sensitivity check. This also allows us to test the validity of our instruments.

We define the normalized age a as the actual age minus the corresponding SRA for each gender: $a = \text{age} - 60$ for men, and $a = \text{age} - 50$ for women. This is the assignment variable for the fuzzy regression discontinuity design in our main analysis.

Outcome Variables

We consider the following outcome variables in our main analysis: (1) *Outpatient incidence*: a dummy indicating whether the respondent used outpatient care in the

past month or not.⁴⁷ (2) # *Doctor visits*: number of doctor visits in the past month.⁴⁸ (3) *Outpatient cost*: the out-of-pocket expenditure for outpatient care in the past month. The expenditures in tables and graphs are in RMB or CNY (both mean Chinese Yuan). In the text, we translate them into US dollars using the exchange rate: 1 USD= 6.5 CNY. (4) *Inpatient incidence*: a dummy indicating whether the respondent received inpatient care in the past year or not. (5) # *Hospital stays*: the number of times the respondent received inpatient care during the past year. (6) *Inpatient cost*: the out-of-pocket expenditure for inpatient care in the past year.⁴⁹ (7) *Self-treatment incidence*: a dummy indicating whether the respondent treated herself in the past month or not.⁵⁰ (8) *Self-treatment cost*: the out-of-pocket expenditure for self-treatment in the past month. (9) *Health check incidence*: a dummy indicating whether the respondent did any health check in the past 2 years or not. (10) *Forgone outpatient incidence*: a dummy indicating whether the respondent was sick but did not seek outpatient care in the past month or not. (11) *Forgone inpatient incidence*: dummy indicating whether or not a doctor had suggested that the respondent needed inpatient care but the respondent was not hospitalized in the past year. (12) *Self-reported health*: self-reported health status on a scale from 1 to 5, with 1: excellent; 2: very good; 3: good; 4: fair; 5: poor.⁵¹

Control Variables

⁴⁷ There are no general practitioners in China. To see a doctor, one usually needs to go to a hospital or clinic. According to the survey question, outpatient care refers to visiting a public hospital, private hospital, public health center, clinic, or health worker's or doctor's practice, or to home visits by a health worker or doctor.

⁴⁸ More precisely: the total number of visits to general hospitals, specialized hospitals, Chinese medicine hospitals ("Zhongyi"), community healthcare centers, township hospitals, healthcare posts, private clinics, and other healthcare organizations.

⁴⁹ Inpatient expenditures include fees paid to the hospital, including ward fees but excluding wages paid to a hired nurse, transportation costs, and accommodation costs for the respondent herself or for family members.

⁵⁰ Self-treatment refers to treatment without resorting to professional medical care, such as over-the-counter drugs, traditional herbs or medication, tonic/health supplement, and the use of healthcare equipment.

⁵¹ We do not look at dentist visits as outcome variable in our main analysis. Summary statistics and estimation results of dentist visits are available in Tables 3.A.12 and 3.A.14. We do not find a significant effect of retirement on dentist visits. Questions about dentist visits are only included in the second wave. Dentists work in outpatient care departments in hospitals. As for other outpatient care, patients do not have to make an appointment, and waiting times tend to be short.

We control for predetermined variables such as individual's gender, a third-degree age polynomial, living with or without a partner,⁵² and education level. Age is constructed from information about birth year and month. Education level is categorized as low (at most elementary education),⁵³ middle (finished middle school, high school, or vocational school), or high ("two-/three-year college/associate degree", "four-year college/bachelor degree", master degree, or doctoral degree). In additional analysis and sensitivity analysis, we also study mechanisms related to variables like BMI, chronic disease, etc.

Summary Statistics

Table 3.3 presents summary statistics. The sample consists of 5,178 observations. About 55% of them are men. The average age is 59 years old. 52.8% have both processed retirement and stopped working. The vast majority (86.6%) have a partner. The middle education group is the largest (61.1%), while 28.1% of the sample belongs to the low education group and only a small minority are highly educated. The average individual yearly income is about \$4000 with a standard deviation of \$2700; 94.9% of individuals in our sample are covered by health insurance, and 61.6% are enrolled in a pension plan.⁵⁴

The probability of having used outpatient care last month is close to 20%, while the probability for having used inpatient care last year is only 12.8%. On average, people visit a doctor once every two months and stay in hospital once every five years. Average out-of-pocket expenditures on outpatient care (monthly) and inpatient care (yearly) are \$8 and \$41, respectively.⁵⁵ Self-treatment is very common: Almost 60%

⁵² In our sample, 95.62% of respondents have no further household members other than their spouse. Results are essentially unchanged if we add the number of household members other than the spouse as additional control.

⁵³ The low education group includes illiterates, those who did not finish primary school but are capable of reading and/or writing, those who have been to home school, and those who finished elementary school.

⁵⁴ The survey question in 2011 asks respondents whether they are claiming a pension. In 2013, the question changed to "whether they are either claiming or accumulating a pension". We try to use other questions in 2011 to retrieve the information about whether the respondent is accumulating a pension in government/institution/new rural/other pension programs. However, we cannot exclude the possibility of some measurement error if contributors to some pension programs are left out. Adding pension claiming as a control variable does not influence the estimation results.

⁵⁵ These average amounts are unconditional. The averages conditional on being positive are: out-of-pocket expenditure on outpatient care (monthly) \$117; inpatient care (yearly) \$1405; monthly out-of-pocket cost of self-treatment \$40.

Table 3.3: Summary Statistics

Variable	Obs.	Mean	Std. dev.	Variable	Obs.	Mean	Std. dev.
<u>Outcome Variables</u>				<u>Other Variables</u>			
Outpatient incidence	5162	0.193	0.395	Enrolled in pension plan	5151	0.616	0.486
# Doctor visits	5162	0.431	1.61	Medical Insurance	5178	0.949	0.219
Outpatient cost	5178	51.882	434.556	<u>Mechanism Variables</u>			
Inpatient incidence	5176	0.128	0.334	Mental health	4616	11.951	4.032
# Hospital stays	5173	0.184	0.622	Life Satisfaction	4572	2.861	0.645
Inpatient cost	5178	266.37	2655.498	Individual income	4293	25444.61	17857.99
Self-treatment incidence	5163	0.58	0.494	Chronic disease	5178	0.657	0.475
Self-treatment cost	5178	114.524	342.017	Smoking	4650	0.25	0.433
Health check incidence	5178	0.62	0.485	<u>Physical Functioning Variables</u>			
Forgone outpatient incidence	5178	0.073	0.26	BMI	3314	24.692	3.980
Forgone inpatient incidence	5178	0.043	0.203	Systolic blood pressure	3340	131.055	20.926
Self-reported health	4454	3.674	0.891	Diastolic blood pressure	3339	77.272	13.016
<u>Treatment Variable</u>				Diabetes	4972	0.109	0.311
Retirement	5178	0.528	0.499	Cancer	4982	0.012	0.108
<u>Control Variables</u>				Stomach disease	4993	0.169	0.375
Male	5178	0.551	0.497	<u>Instrument Variable</u>			
Age in years	5178	58.871	8.438	Age ≥ 60 (or 50)	5178	0.651	0.477
Has a partner	5178	0.866	0.341				
Low education	5178	0.281	0.449				
Middle education	5178	0.611	0.488				
High education	5178	0.107	0.309				

Notes: Mental health: index for mental health problems, ranging from 8 to 32. Life satisfaction: scale from 1 (completely satisfied) to 5 (not at all satisfied). See text and Appendix 3.C for detailed variable definitions.

of respondents report that they have used it last month. The average monthly out-of-pocket cost of self-treatment is about \$18. Approximately 62% had a health check in the last two years. During the past year, 7.3% of respondents have foregone outpatient care and 4.3% have foregone inpatient care even though such care was recommended by a physician.

In Table 3.A.1⁵⁶ we show additional summary statistics for individuals who are just below and just above the SRAs. Retired individuals just above the SRA use substantially more health care than people just below SRA or those just above SRA who have not yet retired. For example, they have substantially more hospital stays and inpatient care incidences and much higher outpatient and inpatient costs. Whether these differences reflect causal effects of retirement on healthcare use is what we will analyze in the next section, but the data already suggest that some mechanisms are more likely than others. For example, individual income is slightly higher among the non-retirees above the age cut-off than among retirees, while health expenditures are much higher for the retirees, suggesting that income is not a major channel. On the other hand, non-retirees above the age cut-off more often forego outpatient and inpatient care than retirees, suggesting that, controlling for diagnosis, current workers less often use (or postpone) health care, possibly since their higher opportunity cost of time.⁵⁷

Table 3.A.1 also shows summary statistics for individuals who were below the SRA and working in the first wave and above SRA and retired in the second wave (columns 5 and 6). For this (small) sample of compliers, socio-economic characteristics such as education level do not differ markedly from the general population in the full sample, and their healthcare utilization before retirement is generally not higher than for others of the same age. This suggests that compliers did not already use more medical treatment before they retired.

⁵⁶ Tables 3.A.1 – 3.A.16 are in the Appendix.

⁵⁷ Note that the age difference between columns (2) and (3) is substantial. This is due to the large difference between SRA for men and women and the difference in the fractions of retirees just above the age cut-off among men and women.

3.4 Empirical Strategy

Our aim is to estimate the causal effect of retirement on healthcare utilization. We start with a linear model:

$$H_{it} = \tau R_{it} + X'_{it}\beta + \varepsilon_{it} \quad (3.1)$$

H_{it} is one of the 12 outcome variables measuring healthcare utilization (or health) of individual i in wave t . R_{it} is the binary variable for retirement and τ is the causal effect of retirement on the outcome, the main parameter of interest. X_{it} is a vector of predetermined variables including gender, age, age², age³, living with a partner, education level, and a constant.

If we assume that ε_{it} is an idiosyncratic shock that is uncorrelated with R_{it} and X_{it} , then OLS gives a consistent estimate of τ . But this assumption may not be valid, since retirement might be endogenous to healthcare use. For example, both retirement decisions and healthcare utilization might be influenced by an unobserved component of health or another unobserved factor.

To correct for potential endogeneity of retirement, instrumental variables estimation (Coe and Zamarro 2015a) and a fuzzy regression discontinuity design (Fé and Hollingsworth 2012, Eibich 2015) are two frequently used methods in the existing literature. We use a nonparametric fuzzy regression discontinuity (RD) design for our main analysis, avoiding restrictive assumptions on functional form. As a robustness check, we present results with linear IV regressions (Section 3.7), using a binary variable Z_{it} for being at or above the SRA (60 for men and 50 for women) as an instrument for retirement.

Fuzzy Regression Discontinuity Design

The regression discontinuity design exploits the SRA as a source of exogenous variation in retirement status. Since not all individuals retire exactly at their SRA, this RD framework is fuzzy (Lee and Lemieux 2010). The treatment effect can be estimated as the ratio of the jump in the outcome variable H and the jump in the probability of being retired at the SRA, as shown in equation (3.2):

$$\tau_{FRD} = \frac{\lim_{\varepsilon \downarrow 0} E[H|a=0+\varepsilon] - \lim_{\varepsilon \uparrow 0} E[H|a=0+\varepsilon]}{\lim_{\varepsilon \downarrow 0} E[R|a=0+\varepsilon] - \lim_{\varepsilon \uparrow 0} E[R|a=0+\varepsilon]} \quad (3.2)$$

Here a is the normalized age (as defined in Section 3.3) which is zero at the cutoff point. τ_{FRD} is the local average treatment effect (LATE), the effect on compliers at the cutoff point. In our context, it is the average change in healthcare utilization for those who retire exactly at the SRA.

A valid fuzzy RD design relies on two main assumptions (Imbens and Lemieux 2008). The first assumption requires a discontinuity in the probability of treatment at the cutoff point:

$$\lim_{\varepsilon \downarrow 0} \Pr(R = 1 | a = 0 + \varepsilon) \neq \lim_{\varepsilon \uparrow 0} \Pr(R = 1 | a = 0 + \varepsilon)$$

This assumption is verified in Figure 3.2, which shows how retirement rates vary with age. We can see a discontinuity at the SRA ($a = 0$) where the probability of being retired increases by around 30 percentage points.

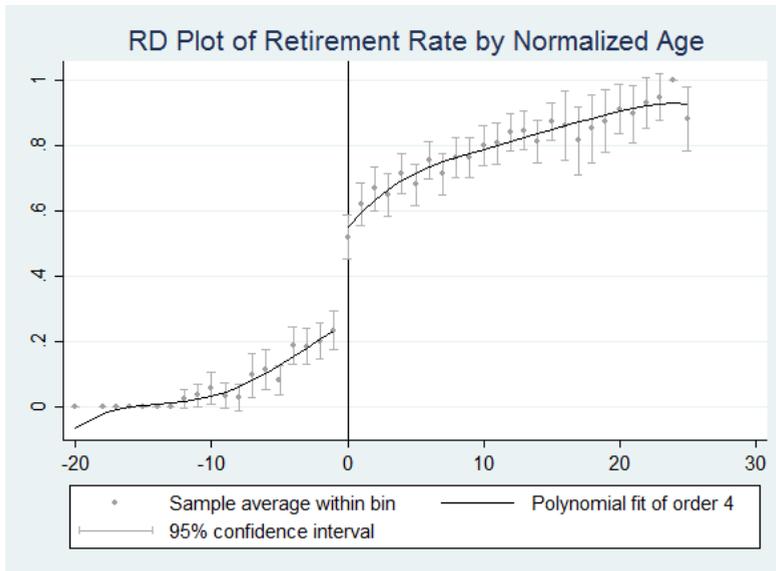


Figure 3.2: Retirement Rate by Normalized Age

The second assumption requires continuity in potential outcomes as a function of the assignment variable around the cutoff point. This implies that in the absence of retirement, healthcare utilization should not change at the cutoff point. In other words, “all other factors” driving healthcare utilization must be continuous at the cutoff point (see, e.g., Hahn, Todd, and van der Klaauw, 2001). Though we cannot test this

assumption directly, we check in Section 3.7 whether relevant variables change significantly at the age cutoff, considering, e.g. participation in pension and health insurance programs and switches between health insurance programs. The untreated group in our study includes individuals who process retirement, but continue working. We test whether income, working hours, and healthcare utilizations change significantly at the full retirement age for those who continue working, and find no effect. We also examine whether healthcare utilization changes at the SRAs in rural areas. This is a falsification test since the SRAs do not apply to rural China. In addition, in a sensitivity analysis we employ a different definition of retirement which includes all people who have processed retirement. We find that results are similar compared to the baseline estimates.

Nonparametric Estimation

The LATE parameter τ_{FRD} can be estimated parametrically or non-parametrically. In our main analysis, we choose nonparametric estimation to avoid assuming a particular functional form of the assignment variable. We present parametric estimates as a robustness check in Section 3.7.

The nonparametric estimation uses local linear regressions (Fan 1992) to estimate the elements of equation (3.2). The estimate of τ_{FRD} is then computed as in equation (3.3):

$$\hat{\tau}_{FRD}(b) = \frac{\hat{\mu}_{H+}(b) - \hat{\mu}_{H-}(b)}{\hat{\mu}_{R+}(b) - \hat{\mu}_{R-}(b)} \quad (3.3)$$

Here $\hat{\mu}_{H+}(b)$ is the estimate of $\lim_{\varepsilon \downarrow 0} E[H|a = 0 + \varepsilon]$, the healthcare utilization just above the cutoff point. Similarly, $\hat{\mu}_{H-}(b)$, $\hat{\mu}_{R+}(b)$, and $\hat{\mu}_{R-}(b)$ are estimates of the corresponding terms in equation (3.2). For a given bandwidth b , $\hat{\mu}_{H+}(b)$ is computed using a triangular kernel-weighted linear regression of H using observations to the right of the age cutoff. The intercept of this local linear regression is $\hat{\mu}_{H+}(b)$. The other three terms in equation (3.3) are computed similarly.

Following Lee and Lemieux (2010) we “residualize” the outcome variables. We regress outcome variables on age polynomials (age, age², and age³) and other control variables, and then conduct the nonparametric RD analysis described above based

on the residuals.⁵⁸ Residualizing is not necessary for consistency, but reduces the variance. In Section 3.7 we show that estimation results are very similar without residualizing. Figure 3.B.2 displays how the “residualized” outcome variables change with normalized age. It suggests that several types of healthcare use increase abruptly at retirement.

In order to choose the bandwidth b of the kernel function, we use a data driven method suggested by Calonico et al. (2014; 2016). We select b based on two separate MSE-optimal bandwidth selectors (below and above the cutoff). We use a robust variance estimator clustered at the individual level in order to account for the correlation of error terms across waves for the same individual.⁵⁹

We report two types of estimates: “conventional” estimates using conventional coefficient and variance estimators, and “bias-corrected” estimates using a bias-corrected coefficient estimator and a robust variance estimator. According to Calonico et al. (2014; 2016), the latter corrects for the possible bias of $\hat{\tau}_{FRD}(b)$ caused by potential misspecification of the local linear regression with a limited sample size.

3.5 Main Results

The OLS estimates for τ (Table 3.4 Column 1) indicate that retirement is associated with a deterioration of self-reported health (a positive effect means a health decline) and an increase in the utilization of most types of healthcare.

Columns (2) and (3) of Table 3.4 report conventional and bias-corrected RD estimates, respectively. These estimates are close to each other. The remaining columns show the RD results by gender, which are sometimes very imprecise, due to limited sample sizes.⁶⁰ Furthermore, coefficients for men and women are not significantly different from each other (see test statistics in Table 3.A.7). We will therefore mainly focus on columns (2) and (3).

⁵⁸ We also tried age and age². The results are very similar.

⁵⁹ We use a cluster-robust nearest neighbor variance estimator with three nearest neighbors. We also tried alternative variance estimators like heteroskedasticity-robust nearest neighbor variance estimator; the results are very similar.

⁶⁰ Figure 3.B.6 shows plots of retirement rates for men and women separately. In both plots, there is a clear discontinuity in the probability of being retired at the cutoff point.

The overall conclusion is that almost all RD estimates imply a positive effect of healthcare utilization at retirement. The point estimates are typically larger than with OLS, but also much less precise. This applies in particular to the estimated effects on costs (for inpatient, outpatient and self-treatment). The most robust finding is the significant positive effect on the number of doctor visits, for men as well as women. Retirement increases the number of doctor visits by almost one per month for women and more than one per month for men. Since the mean number of doctor visits in the sample is only 0.43, these are quite large effects. Retirement raises the number of hospital stays per year by around 0.4 (60% of one standard deviation). Retirement also leads to an increase in the out-of-pocket amount spent on self-treatment (informal healthcare use) by about 177 RMB/month (\$27/month) - 52% of a standard deviation and around 8% of annual income. These two effects are marginally significant in the pooled sample.

For men, we find a marginally significant positive effect of retirement on the incidence of outpatient care, while this effect is insignificant and even negative for women. Perhaps this is because men tend to work more hours than women, raising their opportunity cost of time when still working. For men, we also find a strong and significant effect on out-of-pocket inpatient cost, which amounts to about 6.8% of average individual annual income. This effect is smaller and insignificant for women. The other effects are less robust or insignificant. In spite of this, the fact that almost all point estimates are positive suggests that there is a general pattern that health care use increases at retirement.

In China, visiting a doctor by itself is not expensive, with a “regular outpatient registration fee” of only about 7 to 15 RMB (\$1 to \$2). This again suggests that particularly for the number of doctor visits the main mechanism may not relate to the monetary costs of health care; it is more likely that the large opportunity cost of time before retirement plays a decisive role. On the other hand, we find essentially no effect of retirement on the probability of a health check. Although retirees should have more time to invest in preventive care, health checks are rather costly for retirees as they are usually not covered by health insurance.

Table 3.4: Main Results

Dependent Variable	OLS (1)	RD		Male		Female	
		Conventional (2)	Robust (3)	Conventional (4)	Robust (5)	Conventional (6)	Robust (7)
Outpatient incidence	0.032** (0.014)	0.104 (0.149)	0.103 (0.183)	0.380* (0.201)	0.392* (0.227)	-0.298 (0.506)	-0.526 (0.712)
# Doctor visits	0.059 (0.066)	0.816* (0.433)	0.887* (0.508)	1.162** (0.532)	1.140* (0.592)	0.829*** (0.176)	0.900*** (0.207)
Outpatient cost	48.688*** (17.365)	0.289 (231.81)	-19.655 (292.76)	529.88* (299.13)	578.11 (358.5)	-1523 (1410.3)	-2469.7 (1834)
Inpatient incidence	0.065*** (0.012)	0.163 (0.131)	0.208 (0.159)	0.103 (0.174)	0.142 (0.214)	0.204* (0.111)	0.158 (0.120)
# Hospital stays	0.112*** (0.020)	0.409* (0.225)	0.491* (0.261)	0.411 (0.283)	0.437 (0.330)	0.416 (0.485)	0.493 (0.659)
Inpatient cost	277.027*** (90.850)	1517.6** (610.66)	1660.4** (690.43)	1925.7** (805.45)	1741.9* (897.69)	824.61 (709.59)	1227.7 (838.04)
Self-treatment incidence	0.052*** (0.018)	0.096 (0.198)	0.102 (0.241)	0.048 (0.303)	-0.035 (0.353)	1.531 (1.650)	2.562 (2.213)
Self-treatment cost	59.403*** (11.383)	177.29** (90.01)	170.06 (107.36)	75.08 (92.86)	68.069 (108.09)	481.8 (399.95)	607.12 (480.68)
Health check incidence	0.030* (0.018)	-0.051 (0.187)	-0.040 (0.229)	0.150 (0.253)	0.093 (0.317)	-0.167 (0.127)	-0.126 (0.139)
Forgone outpatient incidence	-0.011 (0.009)	0.004 (0.110)	0.010 (0.135)	0.022 (0.115)	0.011 (0.145)	0.049 (0.075)	0.049 (0.075)
Forgone inpatient incidence	0.001 (0.008)	0.079 (0.068)	0.068 (0.083)	0.028 (0.071)	0.045 (0.082)	0.192 (0.195)	0.292 (0.238)
Self-reported health	0.143*** (0.035)	0.069 (0.395)	0.060 (0.491)	0.104 (0.499)	0.087 (0.623)	0.245 (0.237)	0.448** (0.227)
Covariates	Yes	Residualized		Residualized		Residualized	
Observations	5,178	5,178		2,851		2,327	

Notes:*Significant at 10%; ** at 5%; *** at 1%. In column (1), numbers in parentheses show robust standard errors clustered at the person level. Table 3.A.2 shows full results. In columns (2) and (3), numbers in parentheses show robust standard errors clustered at the person level. “Conventional” refers to estimates using conventional coefficient and variance estimators, and “Robust” refers to estimates using bias-corrected coefficient estimators and robust variance estimators. Full results with first stage results and information on bandwidth are shown in Table 3.A.3. The number of observations varies slightly across outcome variables due to missing values. Numbers of observations for each outcome variable are reported in Tables 3.A.2 and 3.A.3. Columns (4) to (7) are the same nonparametric RD estimates but separately for men and women. For details and full table, see Table 3.A.7. Covariates refer to age polynomials (age, age², and age³), binary variables for male, having a partner, having mid-education, and having high education. For “residualized” outcome variables, we regress outcome variables on the covariates, and then conduct the nonparametric RD analysis described above based on the residuals.

3.6 Mechanisms

We consider three mechanisms that may explain how retirement affects healthcare utilization: health, time, and income. We also look at possible longer term effects of retirement.

Health and Health Behavior

It is not clear a priori how retirement influences health. On the one hand, retirement can have a negative effect on health if retirees are more often physically inactive, or if retirement comes with the loss of valuable social contacts, identity, and self-esteem. On the other hand, the effect can be positive if retirees use their additional time for health enhancing activities or if retirement ends strenuous or dangerous working conditions. Our data concern several measures of health, including self-reported overall health. Since self-reported health may systematically differ before and after retirement due to, e.g., justification bias (Currie and Madrian 1999, McGarry 2004), we also use objective health measures (e.g. biomarkers for BMI and blood pressure).

Table 3.5 shows RD estimates of the effect of retirement on several measures of health, health behavior, and well-being. We find significant negative effects of retirement on several measures of health: a significant increase in systolic blood pressure, which can be associated with hypertension among the elderly; significant increases in self-reported incidence of diabetes and stomach disease, and even a marginally significant positive effect on the incidence of cancer. These effects could reflect causal effects of retirement on genuine health, but it could also be that retirement makes it more likely that health problems are diagnosed, since, as we saw before, retirement increases the probability of visiting a doctor or hospital. The increase in the incidence of self-reported chronic diseases could be explained by bad health conditions and habits already developed before retirement but without being diagnosed, and not so much by retirement per se changing the disease occurrence. Furthermore, we find that retirement significantly increases mortality (defined as the probability of death between the two waves) by 2% (see Table 3.A.14), even though this is driven by a small number of deaths.

The effects on other health variables are insignificant. Retirement has an insignificant effect on an indicator whether patients have any of a broader range of chronic diseases (e.g. arthritis, stroke, chronic lung diseases etc.). We also find an

insignificant deterioration of self-reported health (Table 3.4) and life satisfaction (Table 3.5). Based on a larger sample, Lei et al. (2011) found that retirement leads to significantly worse self-reported health and subjective well-being; perhaps our insignificant findings are due to the smaller sample and less precise estimates.

Table 3.5: The Effect of Retirement on Mechanism Variables: RD estimates

Dependent variable	Mechanisms	
	Conventional	Robust
log(1+annual income)	-0.044 (0.341)	-0.126 (0.411)
Mental Health	-1.220 (1.535)	-1.637 (1.885)
Life Satisfaction	0.152 (0.319)	0.119 (0.402)
Smoking	-0.096 (0.256)	-0.163 (0.306)
Chronic disease	0.184 (0.198)	0.139 (0.241)
BMI	4.346** (1.884)	5.637** (2.384)
Systolic blood pressure	21.021** (10.446)	24.963* (13.157)
Diastolic blood pressure	7.556 (6.618)	8.560 (8.144)
Diabetes	0.229** (0.099)	0.263** (0.117)
Cancer	0.062 (0.039)	0.083* (0.046)
Stomach disease	0.391** (0.167)	0.461** (0.199)
Residualized	No	
Observations	5,178	

Notes: See Table 3.3, Section 3.3 and Appendix 3.C for variable definitions. *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are robust standard errors clustered at the person level. “Conventional:” estimates using conventional coefficient and variance estimators; “Robust:” estimates using bias-corrected coefficient estimators and robust variance estimators.

We find that retirement does not improve health behaviors: it has no significant effect on the incidence of smoking, and BMI even increases with retirement (Table 3.5). This could be the result of the sudden cessation of intensive work while keeping the same eating habits. As shown in Figure 3.B.10, with a very stable level of food consumption and physical activities, retirees did not substitute work with more calorie-burning exercising. Physical inactivity after retirement could lead to health deterioration and hence more healthcare use.

In summary, our results on the effect of retirement on health should be interpreted with caution. While we find an effect of retirement on high blood pressure and on BMI, it is not clear whether this would immediately translate into higher health care expenditures. Also, the effect of retirement on self-reported chronic diseases could potentially be explained by more physician visits rather than by an effect of retirement per se on chronic diseases.

Time

A second mechanism could be time. The opportunity cost of the time spent on medical care is reduced after retirement, which provides incentives to increase healthcare utilization. Before retirement, taking sick leave could imply a loss of income or even a risk of job loss. According to the Labor Law of the People's Republic of China, a worker with a non-work-related disease is entitled to sick leave for a period between 3 to 24 months, depending on years of employment. A longer sick leave gives the employer the right to terminate the contract. Even within the sick leave period, the salary can be reduced substantially, to 80% of the local minimum wage (Mayer Brown JSM 2008).⁶¹ Taking sick leave can also have a negative impact on variable wage components such as an end-of-month bonus. This is different from the European context, where income and job loss are less of a concern when taking sick leave.

The opportunity cost of sick leave can be especially high in small private companies and for temporary employment, where the law is less strictly enforced and employers are less cooperative with respect to sick leave. This gives workers incentives to postpone medical care, especially time-consuming inpatient care. After retirement,

⁶¹ 80% of the local minimum wage varied across provinces from about \$100 to \$200 per month in the year 2013.

when workers no longer have to worry about income and job loss, they can spend more time on queuing for an available ward and staying in hospital.

This mechanism is not directly testable, but we provide some indirect evidence. We would expect that opportunity costs of sick leave are higher for retirees from small private companies or temporary jobs than for those who work for government, public institutions or state-owned companies. If the time mechanism is relevant, we expect a larger increase in inpatient care use at retirement for employees outside the public sector and in less protected jobs. Separate estimates in Table 3.6 indeed show that this is the case.⁶²

The reduced opportunity cost of time after retirement can potentially also explain the increase in self-reported incidence of diabetes and stomach disease: It is possible that a worker's health has deteriorated already before retirement, but a formal diagnosis is only given after retirement when the worker has time to seek treatment. However, this explanation does not apply to our results for biomarkers which are not self-reported.

Income

A third possible mechanism relates to income. If income falls at retirement, this may have a negative effect on healthcare utilization due to high co-payments. However, we do not find a significant change in income at retirement (Table 3.5 and Figure 3.B.3), and retirees actually use more instead of less healthcare after retirement. This suggests that income is not the main mechanism.

Income does matter, however, to individuals with lower socio-economic status for whom the income constraint can be binding. We split the sample into groups with a low level of education and a middle or high level of education.⁶³ People born before the 1970s with a middle education would be able to get a skilled job and earn an

⁶² We have 2020 observations for workers who currently work for or are retired from government, public institutions or state-owned enterprises ("public") and 3158 other observations ("private"). The first stage plots in Figure B.4 show that the discontinuity in the public sector is larger than in the private sector, in line with the fact that retirement at the SRA is more common in the public than in the private sector.

⁶³ Figure 3.B.4 shows the first stage plot by education groups. We observe clear discontinuities in all three groups. There are 1454 observations in the low education group, 3165 in middle education group and 553 observations in the high education group. Because most sample respondents are born before the Cultural Revolution, high education is rare. We therefore merge middle and high levels of education.

average wage. In contrast, the low educated usually earn a lower income in a strenuous unskilled job.⁶⁴ Table 3.6 presents the estimation results for the two education groups, showing that the effect of retirement on healthcare utilization, especially hospital stays and out-of-pocket inpatient expenditure is larger for people with low education than for people with high or middle education. At the same time, the incidence of forgoing inpatient care increases by more than 20%-points at retirement for the low education group, and by only about 6 or 7%-points for the middle and high education group.

Longer Term Effects

Our regression discontinuity estimates measure the immediate effect of retirement on health care utilization at the SRA. To analyze the effect of retirement on health care costs in the longer run, we examine the effects on health investment behavior and consider the possibility that individuals postpone healthcare use from (shortly) before until after retirement (“demand-shifting”).

Higher health care utilization after retirement can reduce health care expenditures in the long term if it is aimed at preventive care. This is not what we find: Only 10% of outpatient care is preventive care (Table 3.A.12), and there is no significant effect of retirement on preventive outpatient care. Furthermore, higher self-treatment expenditures after retirement are largely explained by a higher consumption of “over the counter medicines” which commonly does not aim at health prevention (Table 3.A.13, Table 3.A.14, and Figure 3.B.11). In addition, the incidence of health checks does not significantly increase at retirement (Table 3.4), and health behaviors such as BMI or smoking do not improve at retirement (Table 3.5).

The increase in health care utilization after retirement can be short lived if it is explained by pent-up demand or by shifting treatment from before to after retirement, due to the high opportunity cost of time while still working. In Figure 3.B.2 we show health care utilization at different ages around the SRA. For inpatient care costs, we see that expenditures decrease for ages just below the SRA, increase at the SRA, and then decrease again, suggesting that there could indeed be demand shifting. Yet, we do not see such a pattern for other health care utilization variables.

⁶⁴ The correlation between the level of education and individual income is around 0.4.

Table 3.6: The Effect of Retirement on Healthcare Use by Groups

Dependent variable	Public sector		Private sector		Low education		Mid & high edu.	
	Convnt. (1)	Robust (2)	Convnt. (3)	Robust (4)	Convnt. (5)	Robust (6)	Convnt. (7)	Robust (8)
Outpatient incidence	0.115 (0.144)	0.109 (0.180)	0.052 (0.213)	0.064 (0.257)	0.317 (0.310)	0.296 (0.398)	0.046 (0.157)	0.034 (0.190)
# Doctor visits	0.626* (0.346)	0.611 (0.405)	0.325 (0.536)	0.456 (0.637)	1.264 (0.844)	1.408 (1.016)	0.267 (0.377)	0.294 (0.435)
Outpatient cost	-77.545 (126.33)	-112.56 (154.17)	75.733 (411.43)	108.22 (516.87)	530.77 (411.88)	732.37 (583.32)	-233.54 (283.94)	-342.43 (352.58)
Inpatient incidence	0.042 (0.136)	0.075 (0.171)	0.315* (0.175)	0.356* (0.211)	0.379 (0.291)	0.434 (0.369)	0.099 (0.149)	0.144 (0.183)
# Hospital stays	0.063 (0.166)	0.099 (0.208)	0.836** (0.368)	0.973** (0.429)	0.908* (0.501)	1.042* (0.621)	0.275 (0.203)	0.339 (0.238)
Inpatient cost	495.33 (347.96)	527.51 (371.45)	2492.2* (1398.8)	2965* (1562.2)	4510.4* (2622.9)	4962.1* (2785.6)	1084** (506.63)	1297.1** (543.05)
Self-treatment incidence	0.051 (0.226)	0.063 (0.281)	0.158 (0.303)	0.154 (0.374)	0.483 (0.470)	0.575 (0.594)	0.030 (0.216)	0.004 (0.257)
Self-treatment cost	87.616 (93.067)	102.92 (112.36)	234.65 (142.93)	232.48 (163.08)	180.15 (149.54)	197.82 (179.02)	169.37* (98.453)	155.38 (117.21)
Health check incidence	-0.132 (0.205)	-0.157 (0.257)	0.220 (0.330)	0.319 (0.394)	-0.038 (0.484)	-0.187 (0.591)	0.016 (0.199)	0.084 (0.235)
Forgone outpatient incidence	0.033 (0.109)	0.053 (0.133)	0.034 (0.145)	0.040 (0.178)	0.0323 (0.227)	-0.012 (0.307)	-0.009 (0.098)	0.003 (0.119)
Forgone inpatient incidence	0.074 (0.062)	0.072 (0.077)	0.066 (0.115)	0.073 (0.138)	0.168 (0.103)	0.209* (0.123)	0.050 (0.066)	0.062 (0.079)
Self-reported health	-0.294 (0.340)	-0.425 (0.411)	0.580 (0.758)	0.640 (0.917)	0.013 (0.829)	-0.001 (1.062)	0.134 (0.422)	0.118 (0.510)
Residualized	Yes		Yes		Yes		Yes	
Observations	2,020		3,158		1,454		3,718	

Notes:*Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are robust standard errors clustered at the person level. “Convnt.” refers to estimates using conventional coefficient and variance estimators, and “Robust” refers to estimates using bias-corrected coefficient estimators and robust variance estimators. We choose separate bandwidths for the region to the left and to the right of the cutoff point for all outcome variables except for “Inpatient cost” for the low education group. For the latter, due to a small number of observations to the left of the cutoff point, we use a common bandwidth on both sides of the cutoff point. For “residualized” outcome variables, we regress outcome variables on age polynomials (age, age², and age³), binary variables for male, having a partner, having mid-education, and having high education, and then conduct the nonparametric RD analysis described above based on the residuals.

As a formal test for demand shifting we apply a donut-hole regression discontinuity design (see Shigeoka 2014), which boils down to not using the years just before and after SRA. Table 3.A.11 shows how parametric and nonparametric estimates change. For inpatient care costs, the positive effect of retirement disappears, but for other variables, the results remain qualitatively unchanged. Especially for self-treatment cost and incidence of forgone inpatient care, the effects are very stable as we drop more years of observations around the SRA cutoff point. This suggests that while we do find some evidence of demand shifting for inpatient care costs, the increased health care use after retirement cannot be attributed solely to demand shifting or pent-up demand.

3.7 Sensitivity Analysis and Specification Checks

Functional Form Assumptions

In addition to the nonparametric fuzzy RD estimation approach presented above we also used alternative parametric estimation approaches. These methods use Z_{it} , the indicator of individual i being at or above the SRA at time t , as instrumental variable for retirement R_{it} . We restrict the sample to 10 years below and above the SRAs to reduce the impact of observations far away from the age cutoff.⁶⁵

In our first robustness check we control for a cubic function of age that is the same function on both sides of the cutoff point as shown in equation (3.1). Column (1) of Table 3.7 presents the results. They are similar to the nonparametric RD estimates in Table 3.4.

In column (2) of Table 3.7, we present IV estimates for the alternative specification in equation (3.4):

$$H_{it} = \delta_1 a_{it} + \delta_2 a_{it}^2 + \gamma_1 a_{it} \cdot R_{it} + \gamma_2 a_{it}^2 \cdot R_{it} + \tau R_{it} + X'_{it} \beta + \varepsilon_{it} \quad (3.4)$$

Here $a_{it} \cdot R_{it}$, $a_{it}^2 \cdot R_{it}$, and R_{it} are instrumented by $a_{it} \cdot Z_{it}$, $a_{it}^2 \cdot Z_{it}$, and Z_{it} respectively. This model assumes a quadratic form of normalized age a_i and allows for different age trends below and above the age cutoff, in contrast to the

⁶⁵ For the parametric estimations in Section 3.7, we also estimated specifications which restricted the sample to 5 years below and above the age cutoff. The qualitative results do not change.

specification in column (1). Column (2) shows that estimation results are similar to the main results.

Column (3) shows non-parametric RD estimation results without residualizing the covariates. These results are very close to the main estimates in column (2) of Table 3.4.

Alternative Estimation Methods

We estimated a specification with fixed effects instrumental variables estimation (see columns 1 and 2 in appendix Table 3.A.15). The coefficients point in the same direction as in the baseline specification and are even larger in absolute size. But the standard errors are also very large. This is not surprising because the identification of this model relies on a small number of individuals who are below the SRA and working in the first wave and above the SRA and retired in the second wave.

In addition, we estimated a model which includes additional instrumental variables for being above the early retirement age (45 years for females and 55 years for males). Results shown in column (3) of Table 3.A.15 are almost identical to the baseline estimates in column (1) of Table 3.7. Adding additional instrumental variables allows us to test the joint validity of the instruments using over-identifying restrictions. The null hypothesis that the instruments are valid is never rejected.

Columns (4) and (5) of Table 3.A.15 show separate estimates for men and women. They are qualitative similar to the non-parametric RD estimates, but much less precise due to the smaller number of observations.

Attrition

We examine to what extent our estimation results are affected by sample attrition. It is possible that some individuals left the sample in the second wave for reasons that are related to health or healthcare use. This motivates two robustness checks: 1) excluding individuals who are only observed in the first wave, keeping those observed in both waves or in the second wave only. 2) using the balanced sample of individuals observed in both waves. We use the same nonparametric estimation approach as in our main specification. The results in columns (5) to (8) of Table 3.7 are very similar to those in the main estimation in Table 3.4. The main finding that retirement increases inpatient care utilization remains.

Table 3.7: Alternative Estimation Approaches and Sub-Samples

Dependent variable	Parametric RD 1	Parametric RD 2	Nonparametric RD		Both-wavers + only in the second wave		Only both-wavers	
	±10 years (1)	±10 years (2)	Conv. (3)	Robust (4)	Conv. (5)	Robust (6)	Conv. (7)	Robust (8)
Outpatient incidence	0.095 (0.081)	0.036 (0.121)	0.104 (0.148)	0.109 (0.183)	0.158 (0.182)	0.195 (0.224)	0.114 (0.213)	0.173 (0.265)
# Doctor visits	0.341 (0.287)	0.242 (0.372)	0.767* (0.411)	0.839* (0.486)	1.221** (0.591)	1.412** (0.676)	1.867* (0.992)	2.387** (1.138)
Outpatient cost	153.283 (111.622)	122.235 (220.788)	-1.005 (234.18)	-20.508 (294.64)	35.312 (242.6)	28.911 (303.3)	251.67 (384.9)	343.58 (484.77)
Inpatient incidence	0.127* (0.066)	0.119 (0.099)	0.165 (0.124)	0.195 (0.152)	0.167 (0.157)	0.206 (0.194)	0.376* (0.218)	0.507** (0.258)
# Hospital stays	0.238** (0.105)	0.346** (0.157)	0.398* (0.223)	0.460* (0.257)	0.445** (0.226)	0.539** (0.266)	0.778** (0.387)	1.011** (0.445)
Inpatient cost	556.356 (451.243)	1,425.696** (693.242)	1513** (598.34)	1592.4** (677.91)	1557.7** (613.19)	1699.9** (687.01)	2515** (1092.3)	3172.1*** (1201.3)
Self-treatment incidence	0.254** (0.110)	0.119 (0.174)	0.098 (0.200)	0.093 (0.243)	0.121 (0.183)	0.130 (0.222)	-0.002 (0.263)	-0.027 (0.318)
Self-treatment cost	178.762*** (62.585)	214.480*** (79.955)	175.55** (84.241)	166.5 (102.67)	161.1 (108.83)	165.92 (134.22)	192.13 (131.78)	195.6 (156.57)
Health check incidence	-0.026 (0.105)	-0.028 (0.160)	-0.049 (0.198)	-0.017 (0.241)	-0.070 (0.247)	-0.158 (0.301)	0.323 (0.265)	0.438 (0.313)
Forgone OP incidence	-0.007 (0.056)	0.017 (0.087)	0.005 (0.111)	0.011 (0.136)	0.026 (0.112)	0.046 (0.140)	-0.027 (0.150)	-0.029 (0.183)
Forgone IP incidence	0.084** (0.044)	0.083 (0.067)	0.081 (0.069)	0.070 (0.084)	0.093 (0.079)	0.099 (0.093)	0.108 (0.119)	0.120 (0.142)
Self-reported health	0.380* (0.205)	0.166 (0.305)	0.054 (0.398)	0.002 (0.507)	-0.154 (0.408)	-0.168 (0.513)	-0.488 (0.437)	-0.663 (0.524)
Covariates	Yes	Yes	No		Residualized		Residualized	
Age poly.	3	2	No		3 (residualized)		3 (residualized)	
Observations	3,542	3,542	5,178		4,353		3,334	

Notes:*Significant at 10%; ** at 5%; *** at 1%. In columns (1) and (2) numbers in parentheses show robust standard errors clustered at the person level. First stage results are shown in Table 3.A.4; full results for Column (1) are shown in Table 3.A.5 and for Column (2) in Table 3.A.6. In columns (3) to (8) numbers in parentheses show robust standard errors clustered at the person level. “Conv:” estimates using conventional coefficient and variance estimator; “Robust:” estimates using bias-corrected coefficient estimators and robust variance estimators. Covariates refer to age polynomials (age, age², and age³), binary variables for male, having a partner, having mid-education, and having high education. “OP” refers to outpatient care. “IP” refers to inpatient care. “Age poly.” Refers to the order of age polynomials. For “residualized” outcome variables, we regress outcome variables on the covariates, and then conduct the nonparametric RD analysis described above based on the residuals.

Alternative Definitions of Retirement

In order to check whether estimation results are sensitive to the definition of retirement, we use two alternative definitions of retirement. Firstly, following Coe and Zamarro (2015a) we define a person as retired if she no longer works in a paid job. This definition is less restrictive than before since it does not require that the respondent has “processed” retirement. It adds the self-employed and those who neither work nor report “processed retirement” to the estimation sample and excludes those who report that they never worked for at least three months during their lifetime.⁶⁶ The resulting sample size is 7021 observations. Figure 3.B.5 (A) shows the first stage plot for the new definition of retirement. The discontinuity at the age cutoff is still noticeable although its magnitude is reduced to around 0.2. Columns (1) and (2) of Table 3.8 display results using the new retirement variable. The magnitude of the effect of retirement is even larger than in Table 3.4. This could be explained by the composition of the new sample which includes a higher share of lower educated people for whom the effect of retirement on health care utilization is larger (cf. Section 3.6).

Secondly, we define retirement based on “processed retirement” only, regardless whether an individual continues working or not. Figure 3.B.5 (B) plots the first stage for this definition. The discontinuity in retirement rates at the age cutoff is around 0.4. Columns (3) and (4) of Table 3.8 show that the estimated effects are slightly smaller, but similar to the results in the main analysis.

Specifications Checks

We check the requirement that other variables do not change discontinuously at the cutoff point. Table 3.A.8 and Figure 3.B.7 confirm that participation rates in pension or health insurance plans do not change significantly at the age cutoff, and the number of switches between different types of health insurance programs does not increase either. Table 3.A.9 shows that for individuals who continue working after their SRA, there are no discontinuities in income, working hours, or healthcare utilizations at the SRA. These results also address the concern that reaching certain ages may have a direct psychological impact, which may lead to an increase in

⁶⁶ Work includes agricultural work, paid employment, self-employment, and unpaid work in family businesses.

Table 3.8: Alternative Retirement Definitions and Placebo Tests

Dependent variable	Retirement definition 1 (Stop working)		Retirement def. 2 (Processed retirement)		Placebo tests for different age cutoffs (Nonparametric RD-Conventional)			
	Convent.	Robust	Convent.	Robust	Cutoff: -1	+1	-5	+5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Outpatient incidence	-0.031 (0.192)	-0.088 (0.231)	0.136 (0.153)	0.142 (0.185)	-0.367 (0.461)	-0.069 (0.337)	-0.754 (1.092)	-0.584 (1.467)
# Doctor visits	0.735 (0.667)	0.721 (0.806)	0.852* (0.466)	0.953* (0.561)	-3.768 (3.476)	-0.053 (0.967)	-1.496 (3.113)	-4.099 (4.422)
Outpatient cost	29.756 (314.75)	21.968 (404.04)	-7.159 (180.72)	-29.312 (233.12)	778.55 (632.97)	4619.7 (8052.4)	1501.8 (3103.3)	-2486.3 (3762.9)
Inpatient incidence	0.407 (0.248)	0.496* (0.291)	0.125 (0.085)	0.146 (0.101)	0.256 (0.364)	0.224 (0.257)	-1.640 (2.669)	-0.902 (1.633)
# Hospital stays	1.091** (0.513)	1.335** (0.592)	0.309** (0.139)	0.366** (0.163)	0.541 (0.504)	0.977 (1.166)	-1.459 (3.566)	-1.646 (1.794)
Inpatient cost	3903.8** (1835.8)	4706.9** (2036.6)	1064.9** (418.38)	1182.7** (504.95)	996.72 (2458.4)	-1250.5 (1664.9)	18596 (42007)	-9123.6 (11549)
Self-treatment incidence	0.254 (0.292)	0.297 (0.354)	0.109 (0.160)	0.138 (0.201)	-0.426 (0.744)	0.438 (0.541)	-2.107 (2.971)	-2.442 (3.524)
Self-treatment cost	66.514 (182.28)	52.949 (229.32)	141.64* (80.455)	145.02 (95.385)	-64.284 (423.35)	-35.004 (225.2)	1223.7 (3553.9)	-1055.9 (1496.9)
Health check incidence	-0.098 (0.241)	-0.105 (0.296)	-0.021 (0.122)	-0.002 (0.148)	0.207 (0.429)	0.327 (0.839)	-2.368 (2.875)	1.809 (2.478)
Forgone OP incidence	0.077 (0.150)	0.093 (0.185)	0.003 (0.081)	0.008 (0.101)	0.084 (0.390)	-0.362 (1.069)	0.504 (1.582)	0.414 (0.895)
Forgone IP incidence	0.080 (0.094)	0.065 (0.114)	0.070 (0.060)	0.058 (0.073)	0.069 (0.260)	0.389 (0.464)	1.853 (5.419)	-0.134 (0.498)
Self-reported health	0.633 (0.532)	0.713 (0.634)	0.063 (0.301)	0.052 (0.371)	1.360 (1.137)	-10.472 (50.922)	2.087 (5.820)	1.202 (1.304)
Residualized Observations	Yes 7,021		Yes 5,178		Yes 5,178			

Notes:*Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses show robust standard errors clustered at the person level. “Convent.,” estimates using conventional coefficient and variance estimators; “Robust,” estimates using bias-corrected coefficient estimators and robust variance estimators. “OP” refers to outpatient care. “IP” refers to inpatient care. For “residualized” outcome variables, we regress outcome variables on age polynomials (age, age², and age³), binary variables for male, having a partner, having mid-education, and having high education, and then conduct the nonparametric RD analysis described above based on the residuals.

healthcare utilization irrespective of employment status (Behncki 2012). Reaching such milestones is unlikely to explain our results since there is no change in healthcare utilization at the SRA for individuals who do not retire. Columns (5) to (8) of Table 3.8 show placebo tests at other nearby cutoff points, -1, +1, -5 and +5. As expected, there is no effect at other cutoff points. We also performed a falsification test based on a sample of rural residents (20,657 observations). Figures

3.B.8 and 3.B.9 show retirement rates and healthcare utilization by normalized age.⁶⁷ As expected, in the absence of a statutory retirement age, we observe no discontinuity in either the probability of being retired or the outcome variables.

As a final check we examine how estimates change if we perform the same nonparametric estimation as in our main analysis but use alternative bandwidths (assuming the same bandwidth for observations below and above the cutoff). The results are robust to bandwidth choice (Table 3.A.10).

3.8 Discussion

This paper studies the causal effect of retirement on healthcare utilization in China. We find that retirement increases healthcare utilization. The size of this effect is substantial. For example, the effect of retirement on inpatient care costs for men is equivalent to around 6.8% of average individual annual income. One possible mechanism is deteriorating health. We find evidence of this in objective measures of physical functioning such as high blood pressure and BMI. Moreover, we find a higher incidence of self-reported diseases after retirement, possibly because retirement makes it more likely that health problems are diagnosed. Arguably, the main mechanism explaining our findings is more time available for medical care after retirement. For the sample as a whole, income is not the dominating channel, yet people with low education are more likely to forego inpatient care recommended by a physician after retirement.

Our findings contrast with previous studies using data from developed countries. They tend to find that retirement reduces the use of outpatient care and has no significant effect on the use of inpatient care. The difference can be explained by differences in institutional characteristics. Labor market institutions and economic conditions can influence healthcare utilization. In a labor market where employment protection is weaker, the opportunity cost of healthcare utilization can be high. Moreover, in a developing country, a larger proportion of people tends to be engaged in arduous or unskilled jobs where their employers are often less cooperative with medical leave because they can easily find a substitute if a worker stays away from his work for a considerable time. These institutional characteristics may well

⁶⁷ Retirement is now defined as having stopped with work, since “processed retirement” only applies to urban areas.

translate into underinvestment in health before retirement, and an increase in healthcare utilization after retirement when time constraints are relaxed.

The characteristics of the Chinese healthcare system also contribute to the different results. One possible explanation is the absence of primary care physicians as gatekeepers. Coe and Zamarro (2015a) point out that healthcare systems with primary care physicians as gatekeepers can be effective at decreasing healthcare utilization after retirement. In China, the main constraint on healthcare utilization are high copayments. They push up out-of-pocket healthcare expenditure and constrain inpatient care use of individuals with low socio-economic status.

Our results relate to the policy debate on whether and when to raise the retirement age in China. With an increasing life expectancy and the one-child policy, China is quickly depleting its demographic dividend and facing overwhelming pressure on the social security and medical care systems. In spite of that, China still has the world's youngest retirement ages. Our findings imply that retirement increases healthcare use. At least in the short run, this would mean that raising the SRAs would reduce expenditures on the public health insurance in urban China. On the other hand, raising retirement ages might have negative effects on health if workers postpone necessary treatment due to time constraints. An increase in retirement ages should therefore go along with more facilitation of preventive care and more efforts to reduce employees' opportunity costs of seeking medical treatment. Moreover, policy makers should not ignore that high co-payments can imply financial barriers to medical care and lead to more forgone inpatient care for the low socioeconomic status group. Last but not least, our findings on the incidence of outpatient care and the cost of inpatient care stem mainly from men. One has to be careful when interpreting the results and making policy suggestions for outpatient care for women.

Our findings may be relevant for other developing countries with a rapidly increasing urban population engaged in arduous jobs in industrial sectors, where employment protection are relatively weak and health insurance is not generous.

For future research, one interesting next step would be to examine the long-run impact of retirement on healthcare utilization in more depth. Another direction of future research could be to investigate how the effect of retirement on healthcare utilization changes as the working population becomes more educated and better paid in the future.

Appendix 3.A Tables

Table 3.A.1: Additional Summary Statistics

Variable	Below age cutoffs: [-2, 0)		Above age cutoffs: Not retire [0, 2)		Above age cutoffs: Retire [0, 2)		Compliers below age cutoffs		Compliers above age cutoffs	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Outpatient incidence	0.154	0.361	0.158	0.366	0.186	0.390	0.118	0.327	0.206	0.410
# Doctor visits	0.310	1.222	0.333	0.969	0.394	1.282	0.176	0.521	0.294	0.629
Outpatient cost	52.790	411.390	49.497	286.059	125.475	1044.241	10.794	44.014	60.588	342.840
Inpatient incidence	0.082	0.275	0.095	0.294	0.169	0.375	0.059	0.239	0.176	0.387
# Hospital stays	0.092	0.329	0.117	0.400	0.274	0.768	0.059	0.239	0.382	1.129
Inpatient cost	34.653	511.687	50.279	477.279	815.527	4604.855	0	0	1926.471	7885.672
Self-treatment incidence	0.552	0.498	0.592	0.493	0.608	0.489	0.676	0.475	0.706	0.462
Self-treatment cost	65.905	169.341	76.531	162.527	145.979	340.948	93.724	129.145	117.794	245.743
Health check incidence	0.599	0.491	0.570	0.496	0.608	0.489	0.794	0.410	0.618	0.493
Forgone outpatient incidence	0.074	0.263	0.089	0.286	0.063	0.244	0.088	0.288	0.088	0.288
Forgone inpatient incidence	0.027	0.163	0.056	0.230	0.046	0.211	0	0	0.088	0.288
Self-reported health	3.573	0.888	3.689	0.850	3.779	0.845	3.778	0.801	3.875	0.660
Retirement	0.218	0.413	0	0	1	0	0	0	1	0
Male	0.535	0.499	0.475	0.501	0.646	0.479	0.676	0.475	0.676	0.475
Age	53.854	5.041	55.196	5.099	57.004	4.829	55.353	4.867	57.382	4.948
Partner	0.906	0.292	0.916	0.278	0.895	0.308	0.912	0.288	0.912	0.288
Low education	0.198	0.399	0.307	0.463	0.165	0.372	0.147	0.359	0.147	0.359
Middle education	0.698	0.460	0.615	0.488	0.772	0.420	0.853	0.359	0.853	0.359
High education	0.101	0.302	0.078	0.269	0.063	0.244	0	0	0	0
Pension	0.634	0.482	0.629	0.484	0.626	0.485	0.588	0.500	0.912	0.288
Medical Insurance	0.958	0.201	0.944	0.230	0.945	0.228	0.941	0.239	1	0
Mental health	11.772	3.864	11.944	3.584	11.281	3.453	11.613	4.112	10.548	2.644
Life Satisfaction	2.885	0.632	2.987	0.703	2.868	0.601	2.621	0.775	2.839	0.523
Individual income	25808.21	17933.66	26520.03	16372.82	25634.26	14790.84	25537.38	12398.23	36183.03	14294.31
Chronic disease	0.550	0.498	0.575	0.496	0.646	0.479	0.5	0.508	0.559	0.504
Smoking	0.253	0.435	0.232	0.423	0.271	0.445	0.185	0.396	0.185	0.396
BMI	24.437	3.905	24.753	4.106	25.173	4.460	26.046	5.487	26.259	3.091
Systolic blood pressure	125.294	20.244	126.801	20.861	129.803	19.999	128.258	18.587	136.246	21.404
Diastolic blood pressure	77.491	12.452	76.720	12.208	79.598	12.007	79.485	10.520	84.652	10.981
Diabetes	0.062	0.242	0.076	0.265	0.128	0.335	0.176	0.387	0.206	0.410
Cancer	0.005	0.072	0.017	0.131	0.026	0.161	0.029	0.171	0.059	0.239
Stomach disease	0.151	0.359	0.253	0.436	0.162	0.369	0.088	0.288	0.088	0.288
Observations	404		179		237		34		34	

Notes: “Below age cutoffs: [-2, 0)” refers to those who are 2 years below the statutory full retirement ages. “Above age cutoffs: Not retire [0, 2)” refers to those who are 2 years above the statutory full retirement ages but not retired (retire defined as processed retirement and stop working). “Above age cutoffs: Retire [0, 2)” refers to those who are 2 years above the statutory full retirement ages and retired. In last four columns, “Compliers” refers to those individuals who were below retirement age and working in the first wave and above retirement age and retired in the second wave. For the compliers, the sample mean of variable “Pension” changes substantially across waves. This is not because of retirement (And we verified that controlling for pension or not does not influence the IV fixed effects estimation results), but because of the measurement error caused by the change of survey question used to construct this variable. See footnote 54 for the details.

Table 3.A.2: Full Version of Column (1) Table 3.4 (OLS Estimates)

Dependent variable	(1) Outpatient incidence	(2) #Dr. visits	(3) Outpatient cost	(4) Inpatient incidence	(5) # Hospital stays	(6) Inpatient cost	(7) Self-treatment incidence	(8) Self-treatment cost	(9) Health check incidence	(10) Forgone outpatient incidence	(11) Forgone inpatient incidence	(12) Self-reported health
Retirement	0.032** (0.014)	0.059 (0.066)	48.688*** (17.365)	0.065*** (0.012)	0.112*** (0.020)	277.027*** (90.850)	0.052*** (0.018)	59.403*** (11.383)	0.030 (0.018)	-0.011 (0.009)	0.001 (0.008)	0.143*** (0.035)
Age	-0.054 (0.098)	-0.316 (0.332)	94.132 (78.008)	0.131* (0.078)	0.210 (0.156)	143.994 (503.128)	-0.261** (0.124)	27.139 (67.710)	-0.205* (0.117)	0.042 (0.060)	-0.000 (0.043)	0.494** (0.248)
Age ²	0.001 (0.002)	0.006 (0.006)	-1.545 (1.332)	-0.002* (0.001)	-0.004 (0.003)	-2.122 (8.753)	0.005** (0.002)	-0.485 (1.171)	0.003* (0.002)	-0.001 (0.001)	0.000 (0.001)	-0.008* (0.004)
Age ³	-0.000 (0.000)	-0.000 (0.000)	0.008 (0.007)	0.000* (0.000)	0.000 (0.000)	0.010 (0.050)	-0.000** (0.000)	0.003 (0.007)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	0.000* (0.000)
Male	-0.063*** (0.013)	- (0.057)	-17.377 (13.036)	0.026** (0.011)	0.043** (0.020)	139.935* (80.205)	- (0.016)	-20.743* (10.803)	- (0.015)	-0.004 (0.008)	-0.011 (0.007)	-0.025 (0.031)
Partner	0.012 (0.017)	0.029 (0.066)	-10.746 (15.627)	0.030** (0.014)	0.070*** (0.023)	129.789 (98.268)	0.036* (0.021)	20.263 (12.486)	0.056*** (0.021)	-0.010 (0.011)	-0.000 (0.008)	0.047 (0.042)
Middle education	0.011 (0.014)	0.002 (0.064)	-11.502 (16.242)	-0.005 (0.012)	-0.022 (0.024)	-37.462 (101.177)	0.009 (0.018)	22.200* (11.954)	0.105*** (0.017)	-0.018* (0.009)	-0.001 (0.007)	- (0.035)
High education	0.006 (0.021)	-0.041 (0.073)	-17.119 (17.550)	-0.022 (0.018)	-0.067** (0.028)	-202.541** (84.043)	-0.000 (0.027)	42.183** (19.196)	0.292*** (0.024)	-0.029** (0.013)	-0.015 (0.010)	- (0.055)
Constant	1.203 (1.892)	5.824 (6.210)	-1,806.230 (1,499.083)	-2.398 (1.476)	-3.794 (2.905)	-3,262.560 (9,490.451)	5.273** (2.401)	-483.612 (1,284.186)	4.512** (2.266)	-0.709 (1.162)	-0.064 (0.809)	-6.643 (4.810)
Observations	5,162	5,162	5,178	5,176	5,173	5,178	5,163	5,178	5,178	5,178	5,178	4,454
R-squared	0.016	0.009	0.004	0.030	0.030	0.005	0.018	0.017	0.035	0.002	0.004	0.042

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses show robust standard errors clustered at the person level.

Table 3.A.3: Full Version of Column (2) Table 3.4 (Nonparametric Fuzzy RD Estimates)

	First-stage	Treatment	Cutoff c = 0	Left of c	Right of c		First-stage	Treatment	Cutoff c = 0	Left of c	Right of c
Outcome	Retirement	Outpatient Incidence	# obs	1803	3359	Outcome	Retirement	# Doctor visits	# obs	1803	3359
Convnt.	0.296*** (0.052)	0.104 (0.149)	BW loc. poly. (h)	3.643	6.428	Convnt.	0.286*** (0.057)	0.816* (0.433)	BW loc. poly. (h)	3.245	5.175
Robust	0.287*** (0.062)	0.103 (0.183)	BW bias (b)	7.503	8.453	Robust	0.281*** (0.065)	0.887* (0.508)	BW bias (b)	7.611	7.835
			rho (h/b)	0.486	0.760				rho (h/b)	0.426	0.661
			# clusters	861	1323				# clusters	861	1215
Outcome	Retirement	Treatment Outpatient Cost	# obs	1809	3369	Outcome	Retirement	Inpatient incidence	# obs	1808	3368
Convnt.	0.307*** (0.050)	0.289 (231.81)	BW loc. poly. (h)	3.808	9.113	Convnt.	0.297*** (0.055)	0.163 (0.131)	BW loc. poly. (h)	3.232	7.954
Robust	0.292*** (0.060)	-19.655 (292.76)	BW bias (b)	7.456	10.489	Robust	0.284*** (0.066)	0.208 (0.159)	BW bias (b)	7.211	9.406
			rho (h/b)	0.511	0.869				rho (h/b)	0.448	0.846
			# clusters	863	1551				# clusters	862	1457
Outcome	Retirement	Treatment #Hospital stays	# obs	1808	3365	Outcome	Retirement	Inpatient incidence	# obs	1809	3369
Convnt.	0.285*** (0.071)	0.409* (0.225)	BW loc. poly. (h)	2.709	6.836	Convnt.	0.291*** (0.070)	1517.6** (610.66)	BW loc. poly. (h)	2.238	7.362
Robust	0.273*** (0.080)	0.491* (0.261)	BW bias (b)	6.542	8.215	Robust	0.276*** (0.079)	1660.4** (690.43)	BW bias (b)	6.900	8.303
			rho (h/b)	0.414	0.832				rho (h/b)	0.324	0.887
			# clusters	809	1328				# clusters	810	1328
Outcome	Retirement	Treatment Self-treatment incidence	# obs	1800	3363	Outcome	Retirement	Self-treatment cost	# obs	1809	3369
Convnt.	0.303*** (0.050)	0.096 (0.198)	BW loc. poly. (h)	3.760	8.486	Convnt.	0.285*** (0.057)	177.29** (90.01)	BW loc. poly. (h)	3.210	5.443
Robust	0.289*** (0.060)	0.102 (0.241)	BW bias (b)	7.862	10.938	Robust	0.280*** (0.066)	170.06 (107.36)	BW bias (b)	7.306	8.209
			rho (h/b)	0.478	0.776				rho (h/b)	0.439	0.663
			# clusters	861	1548				# clusters	863	1328
Outcome	Retirement	Treatment Forgone outpatient incidence	# obs	1809	3369	Outcome	Retirement	Forgone outpatient incidence	# obs	1809	3369
Convnt.	0.309*** (0.050)	-0.051 (0.187)	BW loc. poly. (h)	3.737	9.594	Convnt.	0.295*** (0.051)	0.004 (0.110)	BW loc. poly. (h)	3.654	6.934
Robust	0.296*** (0.060)	-0.040 (0.229)	BW bias (b)	7.568	11.922	Robust	0.284*** (0.061)	0.010 (0.135)	BW bias (b)	7.528	8.707
			rho (h/b)	0.494	0.805				rho (h/b)	0.485	0.796
			# clusters	863	1644				# clusters	863	1328
Outcome	Retirement	Treatment Forgone inpatient incidence	# obs	1809	3369	Outcome	Retirement	Treatment Self-reported health	# obs	1583	2871
Convnt.	0.305*** (0.051)	0.079 (0.068)	BW loc. poly. (h)	3.631	8.932	Convnt.	0.294*** (0.057)	0.069 (0.395)	BW loc. poly. (h)	3.467	6.484
Robust	0.292*** (0.061)	0.068 (0.083)	BW bias (b)	7.610	10.744	Robust	0.290*** (0.069)	0.060 (0.491)	BW bias (b)	7.227	8.858
			rho (h/b)	0.477	0.831				rho (h/b)	0.480	0.732
			# clusters	863	1551				# clusters	774	1226

Notes: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses show robust standard errors clustered at the person level. "Convnt." refers to estimates using conventional coefficient and variance estimators, and "Robust" refers to estimates using bias-corrected coefficient estimators and robust variance estimators.

Table 3.A.4: First Stage Results for Linear-IV Regression with Restricted Sample (± 10 yrs)

Dependent variable	Retirement
Age ≥ 60 (or 50)	0.302*** (0.030)
Age	-0.413*** (0.159)
Age ²	0.009*** (0.003)
Age ³	-0.000*** (0.000)
Male	-0.212*** (0.027)
Partner	-0.019 (0.024)
Middle education	0.142*** (0.021)
High education	0.091*** (0.028)
F-Statistic	249.42
P-value	(0.000)
Observations	3528

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses show robust standard errors clustered at the person level.

Table 3.A.5 Linear-IV Estimates with Restricted Sample (± 10 yrs)

Outcome variable	(1) Outpatient incidence	(2) #Dr. visits	(3) Outpatient cost	(4) Inpatient incidence	(5) # Hospital stays	(6) Inpatient cost	(7) Self-treatment incidence	(8) Self-treatment cost	(9) Health check incidence	(10) Forgone outpatient incidence	(11) Forgone inpatient incidence	(12) Self-reported health
Retire-ment	0.095 (0.081)	0.341 (0.287)	153.283 (110.622)	0.127* (0.066)	0.238** (0.105)	556.356 (451.243)	0.254** (0.110)	178.762*** (62.585)	-0.026 (0.105)	-0.007 (0.056)	0.084* (0.044)	0.380* (0.205)
Age	0.052 (0.170)	-0.463 (0.494)	76.474 (119.049)	0.108 (0.133)	0.225 (0.212)	-372.360 (642.504)	0.163 (0.229)	462.955*** (152.032)	-0.172 (0.211)	0.116 (0.115)	-0.006 (0.076)	0.990** (0.446)
Age ²	-0.001 (0.003)	0.008 (0.009)	-1.368 (2.125)	-0.002 (0.002)	-0.005 (0.004)	6.614 (11.675)	-0.003 (0.004)	-8.591*** (2.798)	0.003 (0.004)	-0.002 (0.002)	0.000 (0.001)	-0.017** (0.008)
Age ³	0.000 (0.000)	-0.000 (0.000)	0.007 (0.012)	0.000 (0.000)	0.000 (0.000)	-0.041 (0.070)	0.000 (0.000)	0.052*** (0.017)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	0.000** (0.000)
Male	-0.079** (0.040)	-0.145 (0.159)	15.421 (55.569)	0.059* (0.031)	0.125** (0.049)	326.950 (238.652)	-0.001 (0.052)	27.261 (28.696)	-0.025 (0.050)	-0.007 (0.027)	0.013 (0.021)	-0.016 (0.099)
Partner	0.013 (0.021)	0.142*** (0.049)	18.042 (21.558)	0.009 (0.017)	0.017 (0.027)	2.166 (149.531)	0.009 (0.027)	-2.325 (15.863)	0.052* (0.028)	-0.015 (0.015)	0.005 (0.011)	-0.021 (0.053)
Middle Education	-0.018 (0.020)	-0.177** (0.088)	-37.807 (31.651)	-0.025 (0.018)	-0.056* (0.031)	-208.508 (159.223)	-0.020 (0.027)	-4.203 (16.035)	0.109*** (0.026)	-0.022 (0.014)	-0.015 (0.011)	-0.137*** (0.051)
High Education	-0.028 (0.026)	-0.255*** (0.084)	-46.692* (27.474)	-0.043** (0.022)	-0.087** (0.035)	-296.568** (133.579)	-0.053 (0.036)	9.297 (24.334)	0.301*** (0.032)	-0.042** (0.018)	-0.025** (0.012)	-0.367*** (0.068)
Constant	-0.829 (3.099)	8.606 (8.969)	-1,314.050 (2,210.439)	-1.732 (2.401)	-3.576 (3.855)	7,200.517 (11,681.496)	-1.925 (4.210)	-8,128.286*** (2,726.937)	3.950 (3.850)	-1.940 (2.095)	0.167 (1.359)	-15.439* (8.181)
Obs.	3,528	3,528	3,542	3,541	3,541	3,542	3,530	3,542	3,542	3,542	3,542	3,078
R-squared	0.014	0.007		0.020	0.018	0.003			0.028	0.003		0.028

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses show robust standard errors clustered at the person level.

Table 3.A.6: Parametric RD Estimation with Restricted Sample (± 10 yrs)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent variable	Outpatient incidence	#Dr. visits	Outpatient cost	Inpatient incidence	# Hospital stays	Inpatient cost	Self-treatment incidence	Self-treatment cost	Health check incidence	Forgone outpatient incidence	Forgone inpatient incidence	Self-reported health
Retirement	0.036 (0.121)	0.242 (0.372)	122.235 (220.788)	0.119 (0.099)	0.346** (0.157)	1,425.696** (693.242)	0.119 (0.174)	214.480*** (79.955)	-0.028 (0.160)	0.017 (0.087)	0.083 (0.067)	0.166 (0.305)
Retirement × Normalized age	-0.011 (0.038)	0.007 (0.106)	-77.074 (59.463)	-0.003 (0.031)	0.021 (0.052)	332.754 (253.434)	-0.122** (0.056)	-22.786 (27.882)	0.025 (0.052)	-0.003 (0.028)	0.020 (0.020)	- 0.235** (0.119)
Retirement × (Normalized age) ²	-0.001 (0.002)	-0.003 (0.006)	-0.490 (3.321)	0.000 (0.002)	0.004 (0.003)	20.833 (13.478)	-0.002 (0.003)	1.552 (1.644)	0.000 (0.003)	0.001 (0.002)	0.000 (0.001)	-0.005 (0.005)
Normalized age	0.014 (0.021)	0.013 (0.051)	30.172 (25.581)	-0.000 (0.017)	-0.028 (0.025)	-258.201** (125.746)	0.063* (0.033)	-0.072 (14.113)	-0.008 (0.031)	-0.002 (0.017)	-0.011 (0.011)	0.128* (0.065)
(Normalized age) ²	0.001 (0.002)	0.001 (0.004)	2.765 (1.989)	-0.000 (0.001)	-0.002 (0.002)	-21.049* (10.861)	0.006** (0.003)	0.475 (1.247)	-0.001 (0.003)	0.000 (0.001)	-0.001 (0.001)	0.011* (0.006)
Male	-0.048*** (0.018)	-0.105* (0.058)	-47.027** (19.898)	0.045*** (0.014)	0.082*** (0.026)	129.735 (127.743)	-0.029 (0.027)	-15.675 (13.866)	0.007 (0.023)	-0.014 (0.012)	-0.004 (0.011)	0.019 (0.050)
Partner	0.012 (0.021)	0.140*** (0.049)	18.876 (24.045)	0.009 (0.017)	0.020 (0.029)	3.358 (169.000)	0.016 (0.032)	0.311 (16.471)	0.054** (0.027)	-0.013 (0.015)	0.004 (0.011)	-0.007 (0.066)
Middle education	0.003 (0.041)	-0.157 (0.121)	14.472 (56.309)	-0.023 (0.033)	-0.099* (0.052)	-617.584** (252.994)	0.081 (0.062)	-0.761 (26.476)	0.092 (0.056)	-0.026 (0.031)	-0.026 (0.021)	0.031 (0.107)
High education	-0.007 (0.047)	-0.246* (0.128)	24.597 (51.349)	-0.037 (0.039)	-0.120* (0.062)	-743.349** (299.639)	0.083 (0.073)	26.109 (36.562)	0.283*** (0.065)	-0.043 (0.037)	-0.043* (0.023)	-0.159 (0.130) 3.626**
Constant	0.173*** (0.041)	0.326** (0.127)	32.507 (64.135)	0.045 (0.032)	0.011 (0.052)	-75.654 (191.036)	0.487*** (0.059)	14.147 (31.702)	0.461*** (0.052)	0.108*** (0.028)	0.021 (0.021)	* (0.113)
Observations	3,528	3,528	3,542	3,541	3,541	3,542	3,530	3,542	3,542	3,542	3,542	3,078

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses show robust standard errors clustered at the person level.

Table 3.A.7: Nonparametric Fuzzy RD Estimates of Retirement Effect $\hat{\tau}$ by

Gender

Dependent variable	Male		Female		Test if coefficients are equal	
	Convent.	Robust	Convent.	Robust	H ₀ : (1)=(3)	H ₀ : (2)=(4)
	(1)	(2)	(3)	(4)	P-value	P-value
Outpatient incidence	0.380* (0.201)	0.392* (0.227)	-0.298 (0.506)	-0.526 (0.712)	0.213	0.219
# Doctor visits	1.162** (0.532)	1.140* (0.592)	0.829*** (0.176)	0.900*** (0.207)		
Outpatient cost	529.88* (299.13)	578.11 (358.5)	-1523 (1410.3)	-2469.7 (1834)	0.154	0.103
Inpatient incidence	0.103 (0.174)	0.142 (0.214)	0.204* (0.111)	0.158 (0.120)	0.625	0.948
# Hospital stays	0.411 (0.283)	0.437 (0.330)	0.416 (0.485)	0.493 (0.659)	0.993	0.939
Inpatient cost	1925.7** (805.45)	1741.9* (897.69)	824.61 (709.59)	1227.7 (838.04)	0.305	0.675
Self-treatment incidence	0.048 (0.303)	-0.035 (0.353)	1.531 (1.650)	2.562 (2.213)	0.377	0.259
Self-treatment cost	75.08 (92.86)	68.069 (108.09)	481.8 (399.95)	607.12 (480.68)	0.322	0.274
Health check incidence	0.150 (0.253)	0.093 (0.317)	-0.167 (0.127)	-0.126 (0.139)	0.263	0.527
Forgone outpatient incidence	0.022 (0.115)	0.011 (0.145)	0.049 (0.075)	0.049 (0.075)	0.844	0.816
Forgone inpatient incidence	0.028 (0.071)	0.045 (0.082)	0.192 (0.195)	0.292 (0.238)	0.429	0.326
Self-reported health	0.104 (0.499)	0.087 (0.623)	0.245 (0.237)	0.448** (0.227)	0.799	0.586
Residualized Observations	Yes 2,851		Yes 2,327			

Notes:*Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses show robust standard errors clustered at the person level. “Convent.” refers to estimates using conventional coefficient and variance estimators, and “Robust” refers to estimates using bias-corrected coefficient estimators and robust variance estimators. In columns (3) and (4), we use a common bandwidth for the regions on both sides of the cutoff point in the estimation of “# hospital stays”, “self-treatment incidence”, “inpatient incidence”, and “forgone outpatient incidence”. We do not residualize outcome variable in the estimation of “# doctor visits” and “inpatient incidence”. And we use heteroskedasticity-robust nearest neighbor variance estimator in the estimation of “forgone outpatient incidence”. Columns (5) and (6) report the one-sided P-value from X^2 distribution with degree of freedom one. For “residualized” outcome variables, we regress outcome variables on age polynomials (age, age2, and age3), binary variables for male, having a partner, having mid-education, and having high education, and then conduct the nonparametric RD analysis described above based on the residuals.

Table 3.A.8: Potential Discontinuities in Other Variables

Dependent variable	Nonparametric RD	
	Conventional	Robust
Pension	-0.168 (0.205)	-0.241 (0.259)
Health insurance	-0.064 (0.088)	-0.074 (0.110)
Urban employee medical insurance	-0.016 (0.198)	-0.020 (0.240)
Urban resident medical insurance	0.080 (0.145)	0.130 (0.175)
New cooperative medical insurance	-0.082 (0.120)	-0.108 (0.146)
Urban & rural resident medical insurance	0.024 (0.046)	0.026 (0.056)
Government medical insurance	0.127 (0.082)	0.130 (0.099)
Medical aid	-0.023 (0.038)	-0.027 (0.046)
Private medical insurance (employer)	-0.019 (0.057)	-0.018 (0.071)
Private medical insurance (employee)	-0.011 (0.090)	0.013 (0.112)
Urban non-employed health insurance	0.032 (0.021)	0.038 (0.024)
Other medical insurance	-0.026 (0.043)	-0.032 (0.053)
No insurance	0.090 (0.085)	0.110 (0.106)
Residualized	No	
Observations	5,178	

Notes: Numbers in parentheses show robust standard errors clustered at the person level. “Conventional” refers to estimates using conventional coefficient and variance estimators, and “Robust” refers to estimates using bias-corrected coefficient estimators and robust variance estimators.

Table 3.A.9: Potential Discontinuities at SRAs for Working Individuals

Dependent variable	Working individuals	
	Conventional	Robust
Outpatient incidence	0.046 (0.068)	0.059 (0.083)
# Doctor visits	0.241 (0.192)	0.303 (0.236)
Outpatient cost	-12.371 (41.461)	-13.544 (51.635)
Inpatient incidence	0.003 (0.047)	0.017 (0.058)
# Hospital stays	0.017 (0.062)	0.040 (0.076)
Inpatient cost	78.509 (58.455)	90.81 (101.44)
Self-treatment incidence	0.034 (0.069)	0.036 (0.081)
Self-treatment cost	0.767 (25.871)	-5.108 (31.369)
Health check incidence	-0.066 (0.072)	-0.068 (0.089)
Forgone outpatient incidence	0.050 (0.037)	0.052 (0.044)
Forgone inpatient incidence	0.045* (0.027)	0.050 (0.033)
Self-reported health	-0.180 (0.126)	-0.203 (0.150)
log(1+annual income)	-0.149 (0.142)	-0.181 (0.168)
Hours of working / year	74.22 (198.33)	159.16 (234)
Residualized	Yes	
Observations	2,445	

Notes: Numbers in parentheses show robust standard errors clustered at the person level. “Conventional” refers to estimates using conventional coefficient and variance estimators, and “Robust” refers to estimates using bias-corrected coefficient estimators and robust variance estimators. For “residualized” outcome variables, we regress outcome variables on age polynomials (age, age2, and age3), binary variables for male, having a partner, having mid-education, and having high education, and then conduct the nonparametric RD analysis described above based on the residuals.

Table 3.A.10: Nonparametric Estimates with Different Bandwidths

Dependent variable	b = 8		b = 6		b = 4	
	Convent.	Robust	Convent.	Robust	Convent.	Robust
	(1)	(2)	(3)	(4)	(5)	(6)
Outpatient incidence	0.046 (0.100)	0.045 (0.159)	0.041 (0.117)	0.099 (0.206)	0.083 (0.160)	0.334 (0.402)
# Doctor visits	0.281 (0.312)	0.425 (0.366)	0.284 (0.325)	0.720 (0.501)	0.534 (0.382)	2.089 (1.396)
Outpatient cost	87.946 (178.26)	54.834 (290.81)	84.166 (222)	65.89 (362.1)	107.2 (308.65)	-23.151 (522.74)
Inpatient incidence	0.116 (0.082)	0.208 (0.133)	0.154 (0.098)	0.233 (0.173)	0.188 (0.137)	0.109 (0.323)
# Hospital stays	0.312** (0.128)	0.502*** (0.187)	0.393*** (0.148)	0.508** (0.253)	0.470** (0.204)	0.330 (0.459)
Inpatient cost	1312** (520.35)	1651.1*** (614.38)	1463.8*** (524.29)	1494* (841.73)	1720.7*** (646.92)	985.32 (1301.5)
Self-treatment incidence	0.109 (0.141)	0.069 (0.225)	0.104 (0.166)	0.158 (0.289)	0.088 (0.226)	0.879 (0.585)
Self-treatment cost	176.23*** (60.907)	148.43 (94.142)	156.9** (69.279)	176.29 (109.43)	137.18 (87.292)	205.42 (213.41)
Health check incidence	0.020 (0.137)	0.064 (0.217)	0.059 (0.161)	-0.124 (0.279)	-0.019 (0.215)	0.036 (0.549)
Forgone outpatient incidence	0.026 (0.072)	0.021 (0.120)	0.028 (0.086)	0.018 (0.153)	0.004 (0.119)	0.059 (0.296)
Forgone inpatient incidence	0.088* (0.053)	0.087 (0.079)	0.088 (0.061)	0.081 (0.101)	0.098 (0.081)	0.210 (0.193)
Self-reported health	0.134 (0.257)	0.056 (0.416)	0.120 (0.296)	0.116 (0.531)	0.192 (0.415)	-0.823 (1.061)
Residualized	Yes		Yes		Yes	
Observations	5,178		5,178		5,178	

Notes:*Significant at 10%; ** at 5%; *** at 1%. The model and estimation method are the same as in columns (2) and (3) of Table 3.4. We use a common bandwidth b for the regions on both sides of the cutoff point. For “residualized” outcome variables, we regress outcome variables on age polynomials (age, age2, and age3), binary variables for male, having a partner, having mid-education, and having high education, and then conduct the nonparametric RD analysis described above based on the residuals.

Table 3.A.11: “Donut Hole” Regression Discontinuity Design

Dependent variable	Parametric RD			Nonparametric RD	
	Omitted ages: a=0	Omitted ages: a=-1, 0, 1	Omitted ages: a=-2, -1, 0, 1, 2	Omitted ages: a = 0	
	(1)	(2)	(3)	Conventional (4)	Robust (5)
Outpatient incidence	0.089 (0.086)	0.071 (0.111)	0.107 (0.144)	0.127 (0.148)	0.122 (0.180)
# Doctor visits	0.438 (0.318)	0.356 (0.452)	0.382 (0.426)	0.852* (0.478)	0.897 (0.565)
Outpatient cost	48.536 (87.272)	174.881* (94.601)	239.549* (137.399)	-199.33 (193.26)	-266.74 (231.4)
Inpatient incidence	0.121* (0.069)	0.064 (0.083)	0.071 (0.113)	0.194 (0.125)	0.229 (0.153)
# Hospital stays	0.246** (0.119)	0.099 (0.149)	0.030 (0.207)	0.566** (0.270)	0.667** (0.314)
Inpatient cost	844.174* (495.762)	-236.441 (476.386)	-645.628 (821.740)	3421** (1607.5)	3956.9** (1837.3)
Self-treatment incidence	0.132 (0.111)	0.140 (0.132)	0.313 (0.197)	0.115 (0.197)	0.118 (0.238)
Self-treatment cost	197.571*** (64.967)	180.039** (85.634)	241.556* (130.252)	222.31** (92.793)	244.36** (110.32)
Health check incidence	-0.091 (0.107)	-0.128 (0.131)	-0.185 (0.180)	-0.074 (0.184)	-0.071 (0.225)
Forgone outpatient incidence	-0.032 (0.057)	-0.052 (0.069)	-0.105 (0.091)	-0.0003 (0.098)	-0.001 (0.119)
Forgone inpatient incidence	0.094** (0.042)	0.113** (0.051)	0.157** (0.067)	0.040 (0.068)	0.021 (0.081)
Self-reported health	0.143 (0.223)	0.284 (0.277)	0.431 (0.410)	-0.019 (0.362)	-0.053 (0.447)
Covariates		Yes		Residualized	
Observations	4,972	4,557	4,165	4,972	

Notes: *Significant at 10%; ** at 5%; *** at 1%. The “omitted ages” is the “donut hole” which specifies the region of observations that we drop for a donut hole RD design. For example, “a= -1, 0, 1” means that we drop observations with normalized age -1, 0, and 1. In columns (1) to (3), we use the same model specification as in column (2) of Table 3.7. Numbers in parentheses show robust standard errors clustered at the person level. In columns (4) and (5), the model specification is the same as in columns (2) and (3) of Table 3.4. Covariates refer to age polynomials (age, age2, and age3), binary variables for male, having a partner, having mid-education, and having high education. For “residualized” outcome variables, we regress outcome variables on the covariates, and then conduct the nonparametric RD analysis described above based on the residuals.

Table 3.A.12: Summary Statistics on Additional Variables

Variable	Obs.	Mean	Std. dev.
Dentist visit incidence	2682	0.224	0.417
# Dentist visits	2670	0.570	1.675
Dental cost	2596	165.129	696.123
Mortality	5178	0.007	0.081
Preventive outpatient care incidence	5178	0.018	0.134

Table 3.A.13: Summary Statistics of Reasons for Self-treatment

N=5178 Std. dev in “()”	Average cost (yuan)	Average cost (yuan) conditional on cost>0	Obs. of positive cost
1. Consumed over-the-counter modern medicines	60.243	217.682 (410.367)	1433
2. Consumed prescription medicines	42.028	334.291 (509.761)	651
3. Consumed traditional herbs or traditional medicines as treatment	41.996	453.980 (669.756)	479
4. Tonic/Health supplement	36.458	477.919 (738.355)	395
5. Use health care equipment	13.927	707.005 (846.543)	102
6. Other	41.968	462.139 (557.684)	18

Table 3.A.14: The Effect of Retirement on Additional Variables

Dependent variable	Mechanisms	
	Conventional	Robust
Dentist visit incidence	-0.081 (0.204)	-0.060 (0.251)
# Dentist visits	-0.194 (0.707)	0.119 (0.859)
Dental cost	-621.09 (478.86)	-749.12 (582.5)
Mortality	0.022** (0.010)	0.025** (0.011)
Preventive outpatient care incidence	0.029 (0.041)	0.029 (0.049)
Self-treatment cost – OTC medicine	137.97** (62.431)	142.9* (75.472)
Self-treatment cost – Prescribed medicine	-21.276 (49.703)	-28.707 (60.868)
Self-treatment cost – Tradition herbs	54.585 (64.941)	52.686 (80.103)
Self-treatment cost – Supplement	96.941* (55.471)	100.82 (64.924)
Self-treatment cost – Equipment	70.966 (52.271)	85.381 (62.165)
Self-treatment cost – Other	-8.593 (6.504)	-15.345* (8.440)
Residualized Observations	No 5,178 (2,682 for dental variables)	

Notes:*Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are robust standard errors clustered at the person level. “Conventional” refers to estimates using conventional coefficient and variance estimators, and “Robust” refers to estimates using bias-corrected coefficient estimators and robust variance estimators.

Table 3.A.15: IV Fixed Effects Estimation and Using Early Retirement Ages as

IV

Dependent variable	IV-FE		Parametric	Over-identifying	Parametric	Parametric
	(1)	(2)	RD 1	test	RD 1 - Male	RD 1 - Female
			(3)	P-value	(4)	(5)
Outpatient incidence	0.245 (0.313)	0.235 (0.324)	0.094 (0.081)	0.751	0.059 (0.119)	0.077 (0.229)
# Doctor visits	1.205 (0.918)	1.240 (0.970)	0.339 (0.283)	0.875	0.521 (0.463)	-0.140 (0.709)
Outpatient cost	392.290 (502.342)	382.587 (525.170)	151.095 (109.275)	0.361	422.845 (284.235)	-271.792 (288.600)
Inpatient incidence	0.407 (0.273)	0.407 (0.281)	0.125* (0.066)	0.364	0.090 (0.120)	0.144 (0.148)
# Hospital stays	0.850* (0.493)	0.813 (0.501)	0.233** (0.104)	0.107	0.365 (0.224)	0.160 (0.182)
Inpatient cost	2,817.574 (2,481.019)	2,819.489 (2,530.206)	552.870 (445.942)	0.859	1,365.368 (1,000.655)	91.622 (737.963)
Self-treatment incidence	-0.017 (0.396)	-0.038 (0.406)	0.254** (0.110)	0.980	0.162 (0.175)	0.336 (0.292)
Self-treatment cost	51.118 (280.722)	37.929 (282.275)	179.533*** (62.142)	0.675	182.578** (89.436)	362.687** (167.998)
Health check incidence	-0.053 (0.358)	-0.072 (0.369)	-0.029 (0.105)	0.513	0.128 (0.167)	-0.424 (0.277)
Forgone outpatient incidence	0.121 (0.241)	0.155 (0.250)	-0.006 (0.056)	0.523	0.000 (0.085)	0.018 (0.157)
Forgone inpatient incidence	0.301 (0.191)	0.307 (0.198)	0.084** (0.044)	0.684	0.018 (0.069)	0.142 (0.131)
Self-reported health	0.912 (0.830)	1.014 (0.873)	0.380* (0.205)	0.984	0.165 (0.324)	0.522 (0.464)
First-stage estimates of IV:						
≥ full retirement ages	0.151*** (0.045)	0.148*** (0.045)	0.301*** (0.031)		0.329*** (0.053)	0.219*** (0.055)
≥ early retirement ages			-0.011 (0.025)		0.032 (0.044)	-0.018 (0.051)
Covariates	Yes	Yes	Yes		Yes	Yes
Age polynomial	2	3	3		3	3
Observations	3,334	3,334	3,542		2,050	1,492

Notes:*Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses show robust standard errors clustered at the person level. Column (1) control for age polynomials up to the second degree (age, age²), and column (2) up to the third degree (age, age² and age³). Similar to column (1) of Table 3.7, we use the indicator of being at or above the SRA as the instrumental variable for retirement. Column (3) to (5) have the same specification as column (1) of Table 3.7 (±10 years) except for the instrumental variables. Besides the dummy variable being above full retirement ages, we include an extra instrument variable being above early retirement ages. The column next to column (3) show the P-value for the test of over-identifying restrictions. Column (4) and (5) are separate estimates for men and women. Covariates refer to age polynomials (age, age², and age³), binary variables for male, having a partner, having mid-education, and having high education.

Appendix 3.B Figures

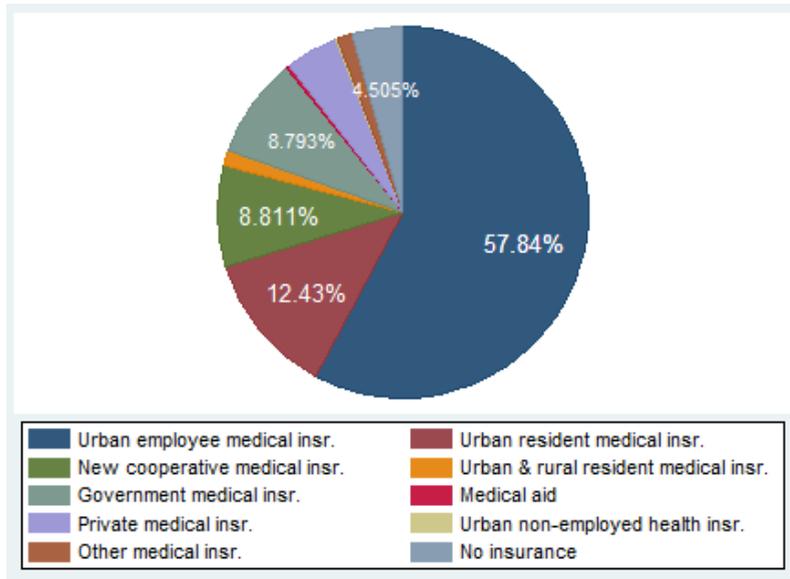


Figure 3.B.1: Health Insurance Coverage by Type of Insurance

(Source: own calculations from our sample)

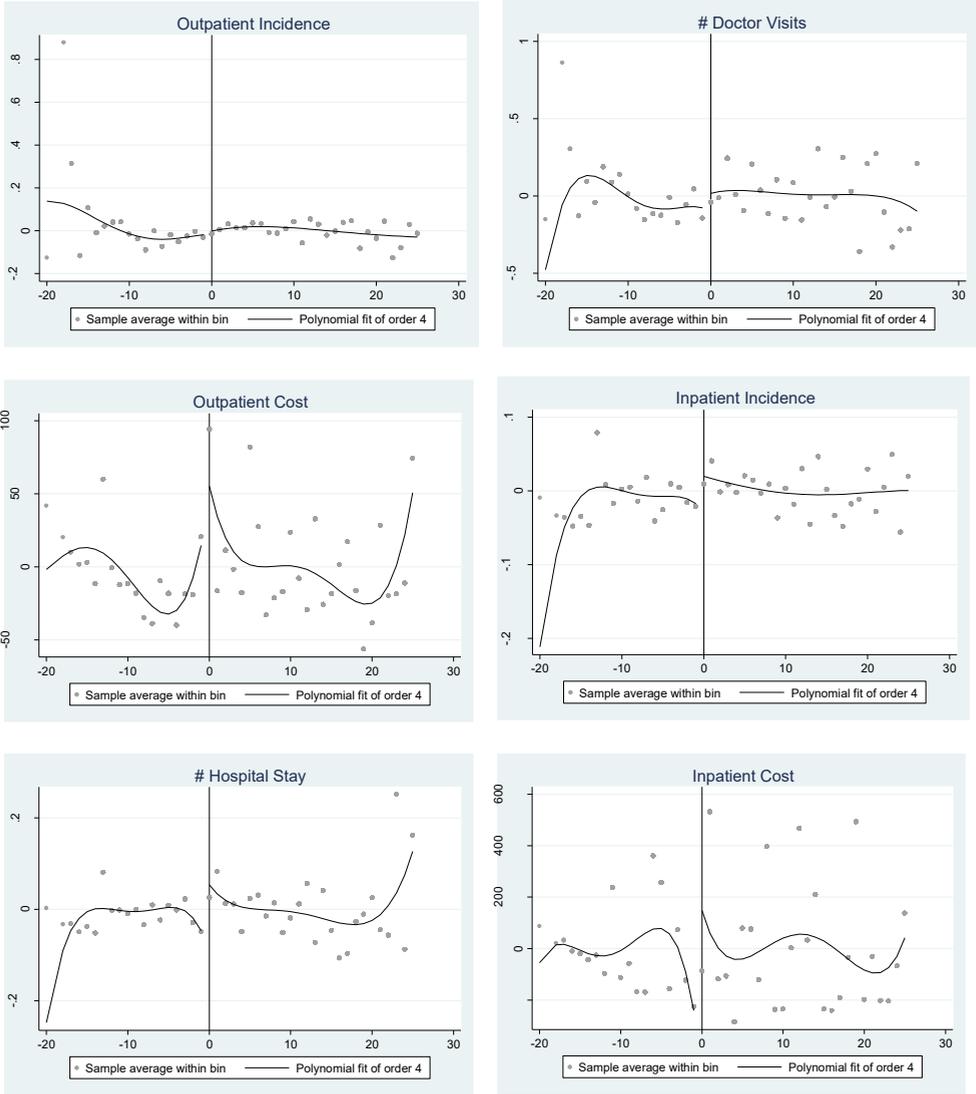


Figure 3.B.2: RD Plots of Residualized Outcome Variables by Normalized

Age⁶⁸

⁶⁸ We use a uniform kernel to construct local-polynomial estimators, assuming equal weights for all observations.

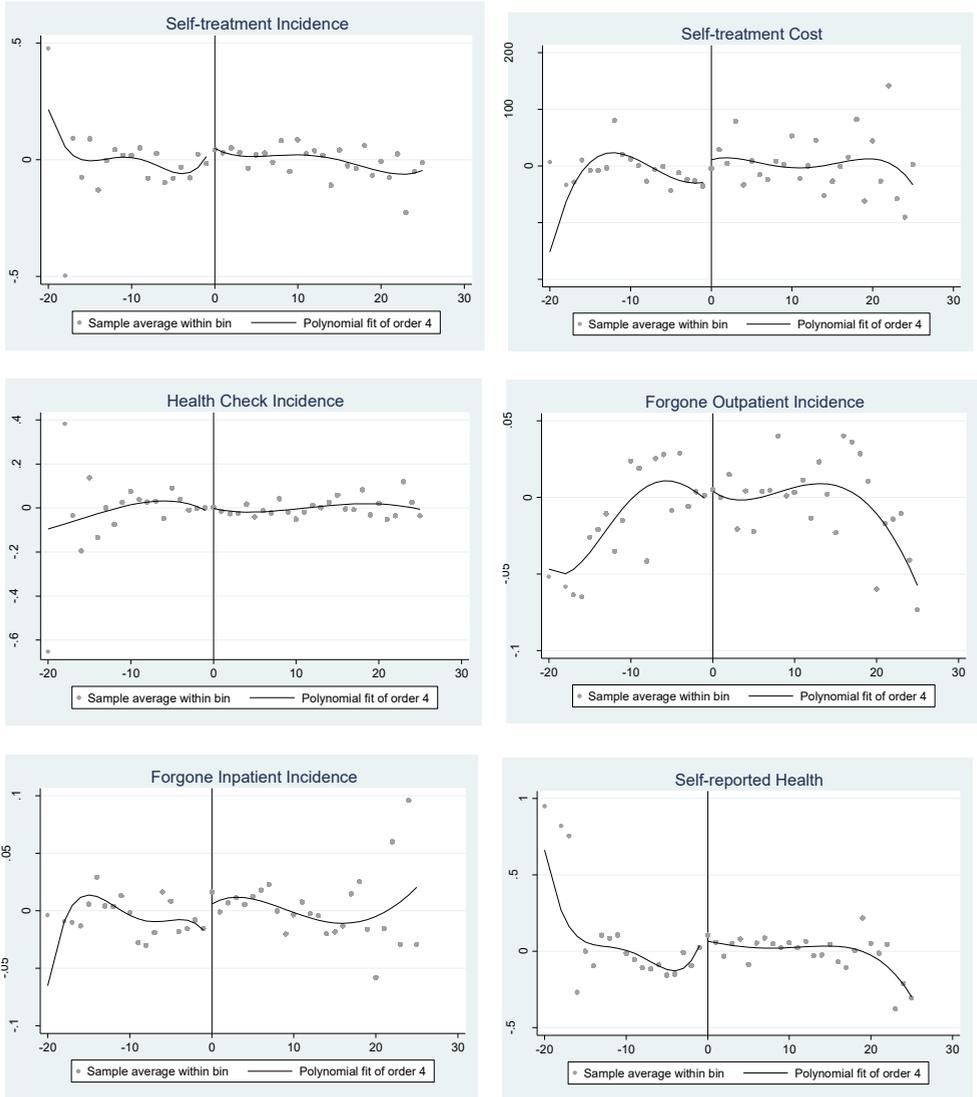


Figure 3.B.2: (Continued)

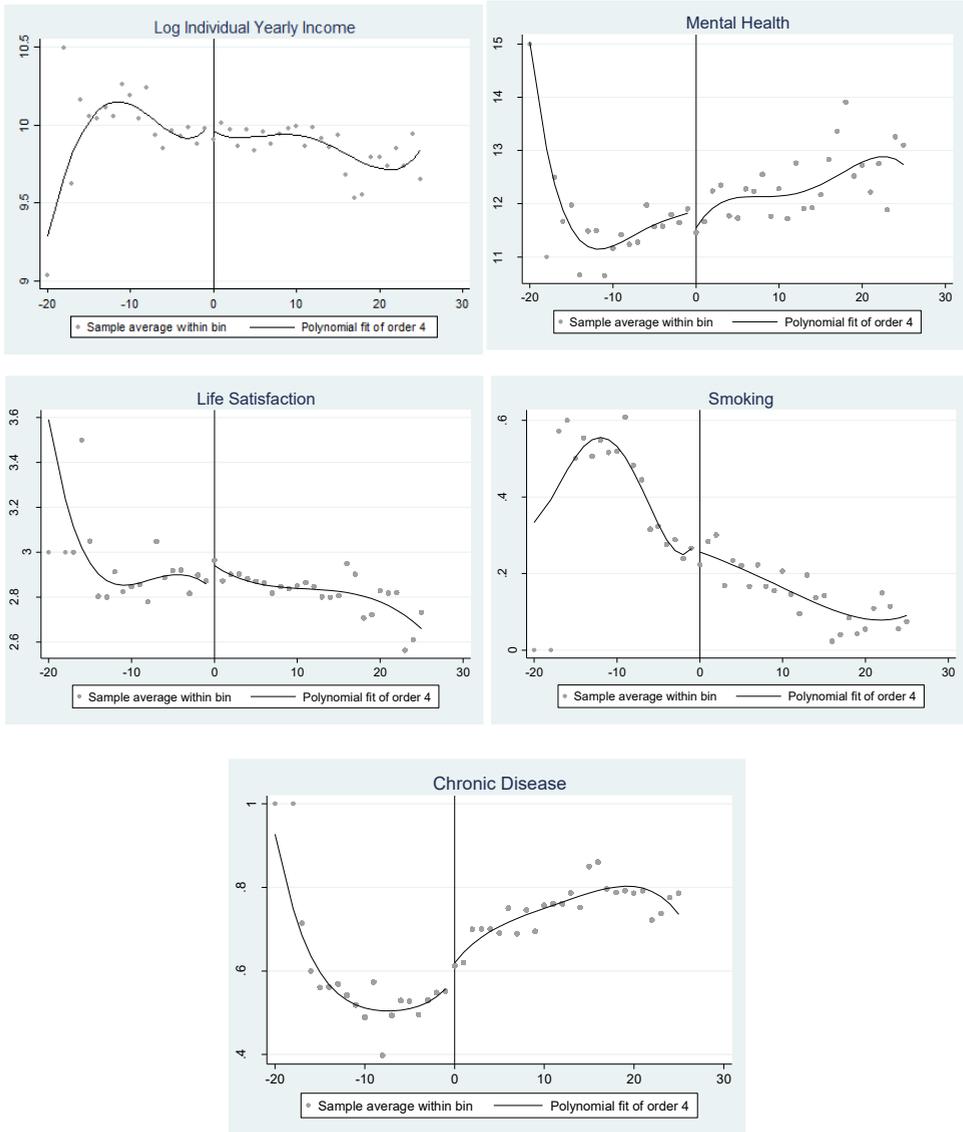


Figure 3.B.3: RD Plots of Mechanism Variables by Normalized Age

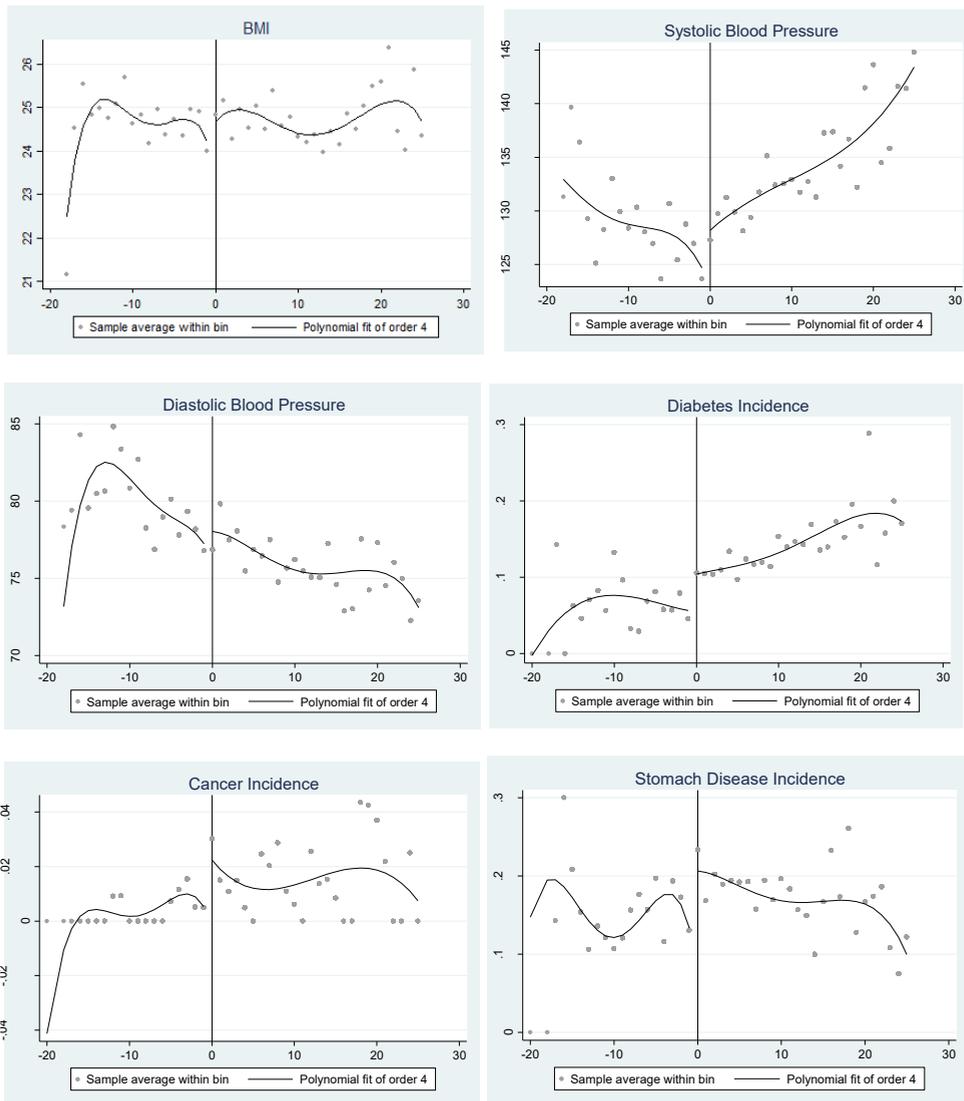


Figure 3.B.3: (Continued)

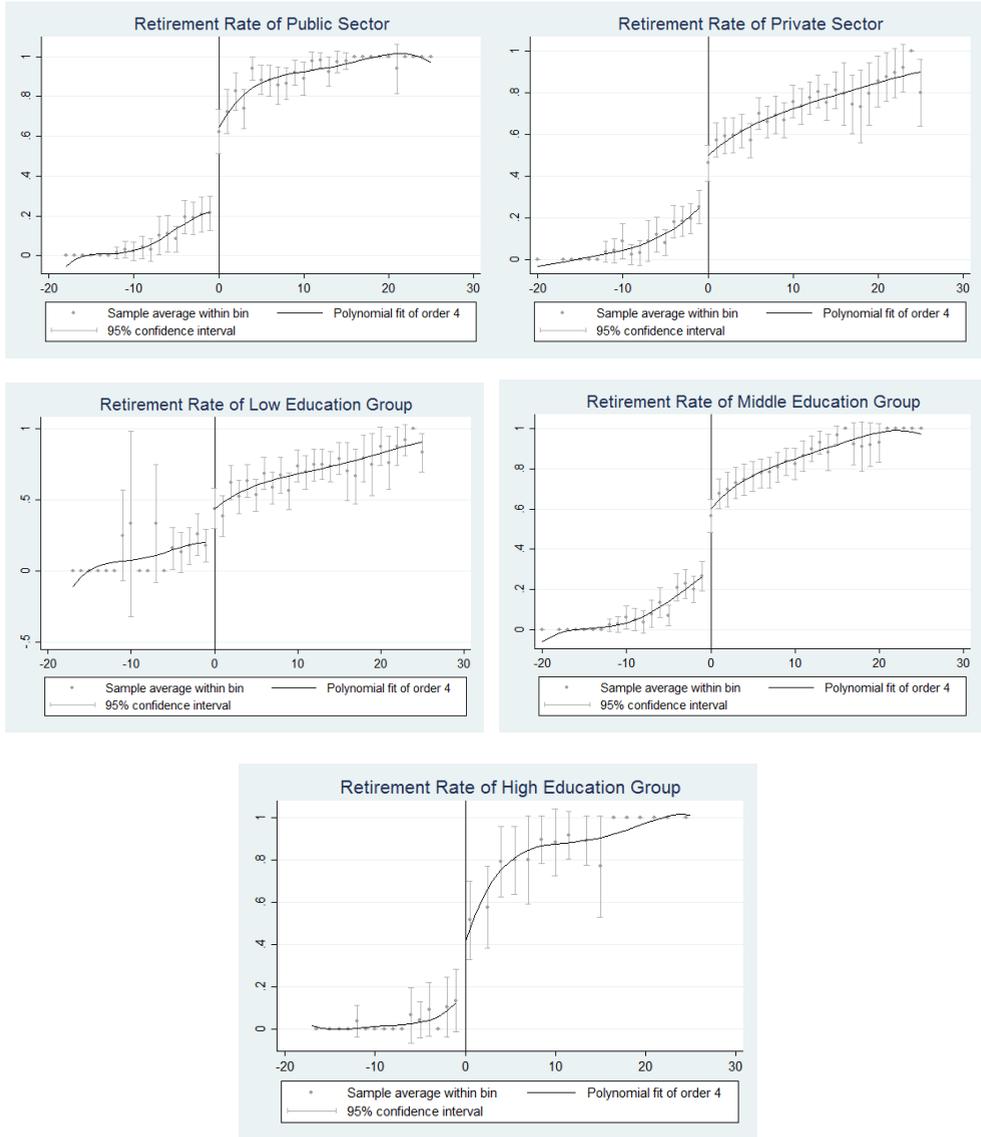
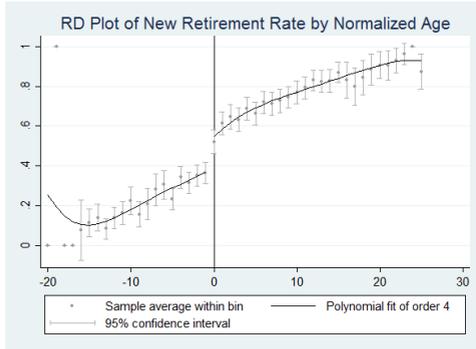
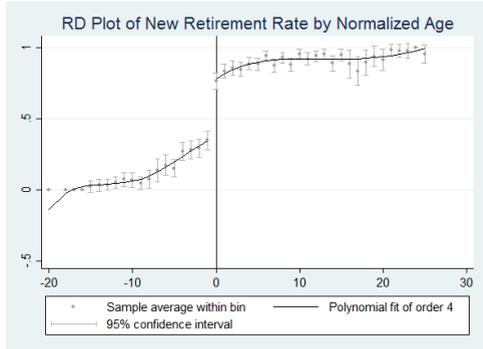


Figure 3.B.4: RD Plot of the First Stage by Sectors and Education Groups



(A) Stop Working



(B) Processed Retirement

Figure 3.B.5: Retirement Rate for Alternative Definition of Retirement by Normalized Age

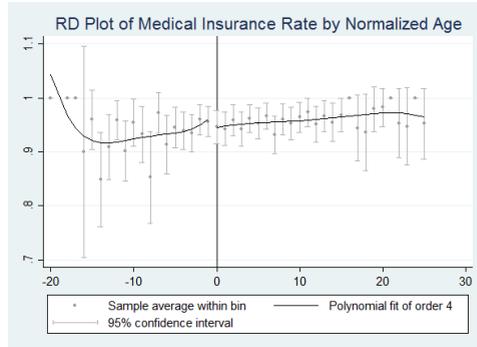
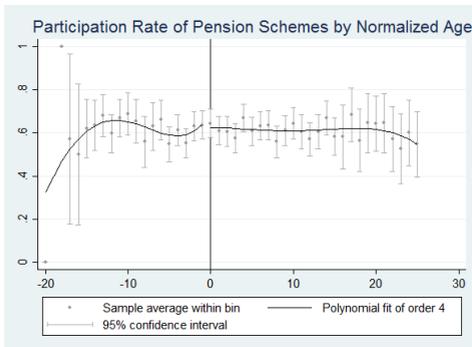


Figure 3.B.6: RD Plots of Pension and Health Insurance Coverage

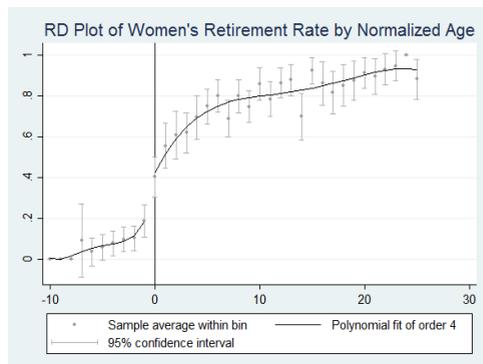
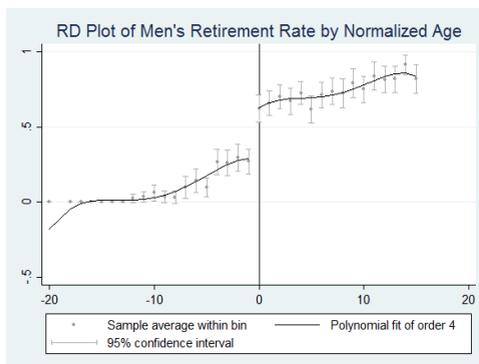


Figure 3.B.7: RD Plots of Retirement Rate by Gender

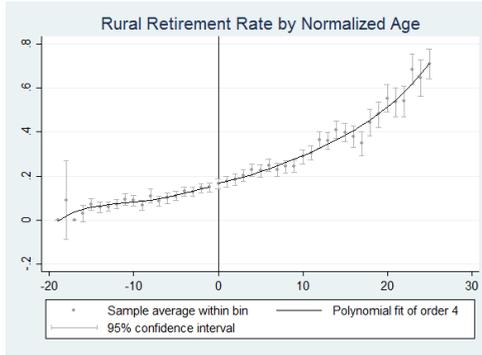


Figure 3.B.8 Retirement (Stop Working) Rate of Rural Population

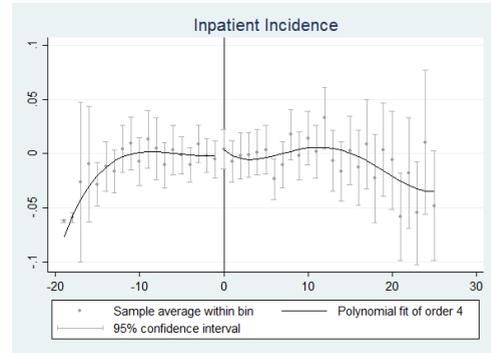
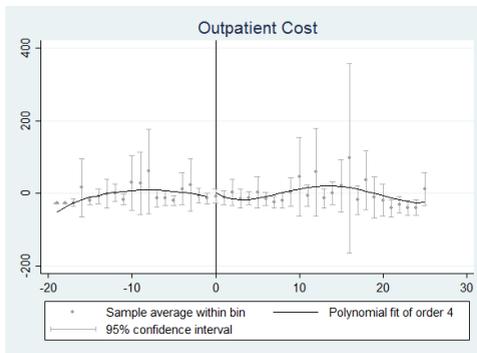
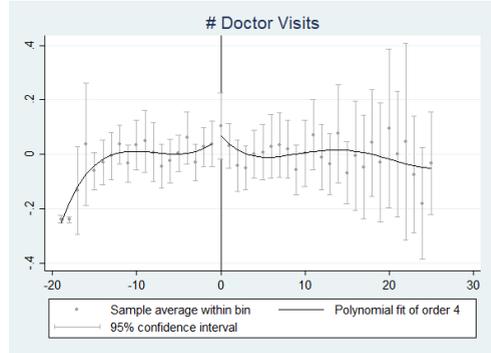
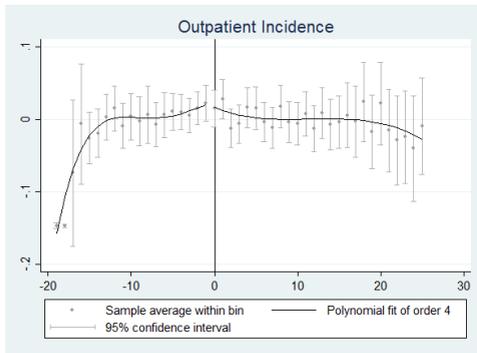


Figure 3.B.9: Health Utilization by Normalized Age of Rural Population

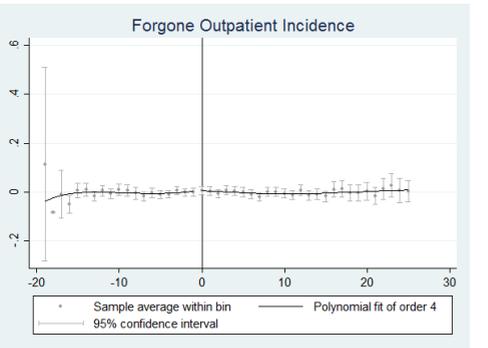
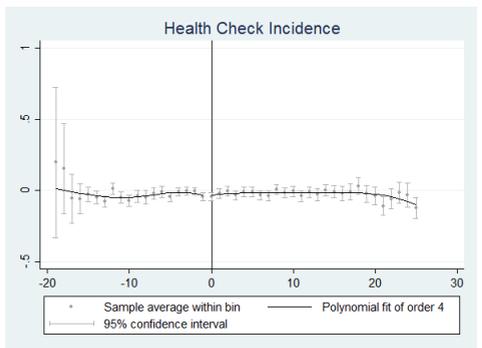
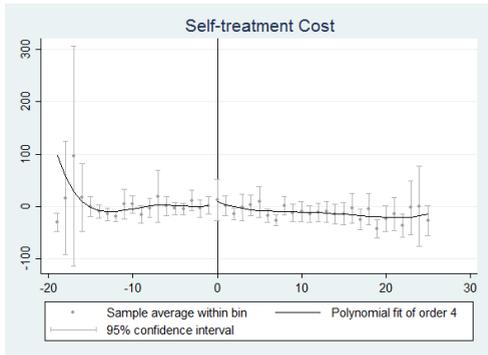
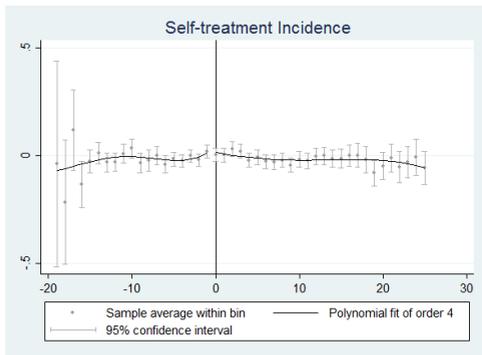
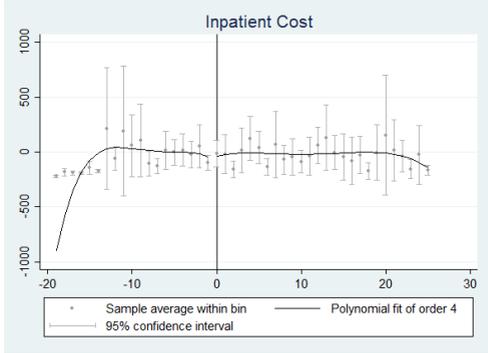
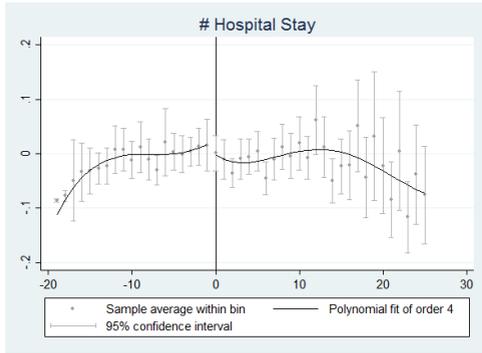


Figure 3.B.9: (Continued)

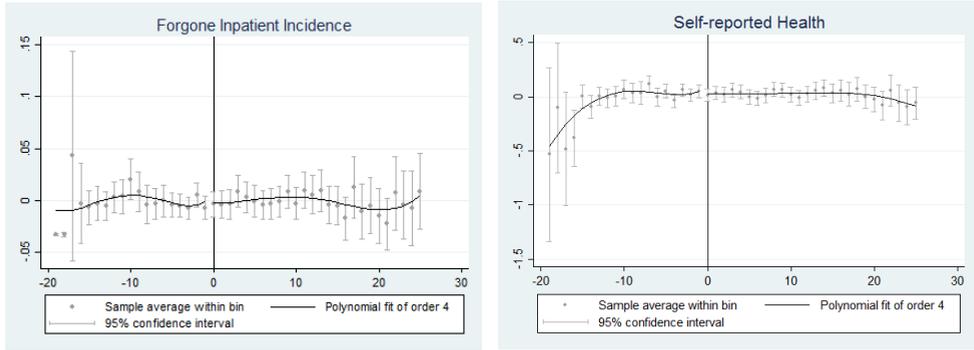


Figure 3.B.9: (Continued)

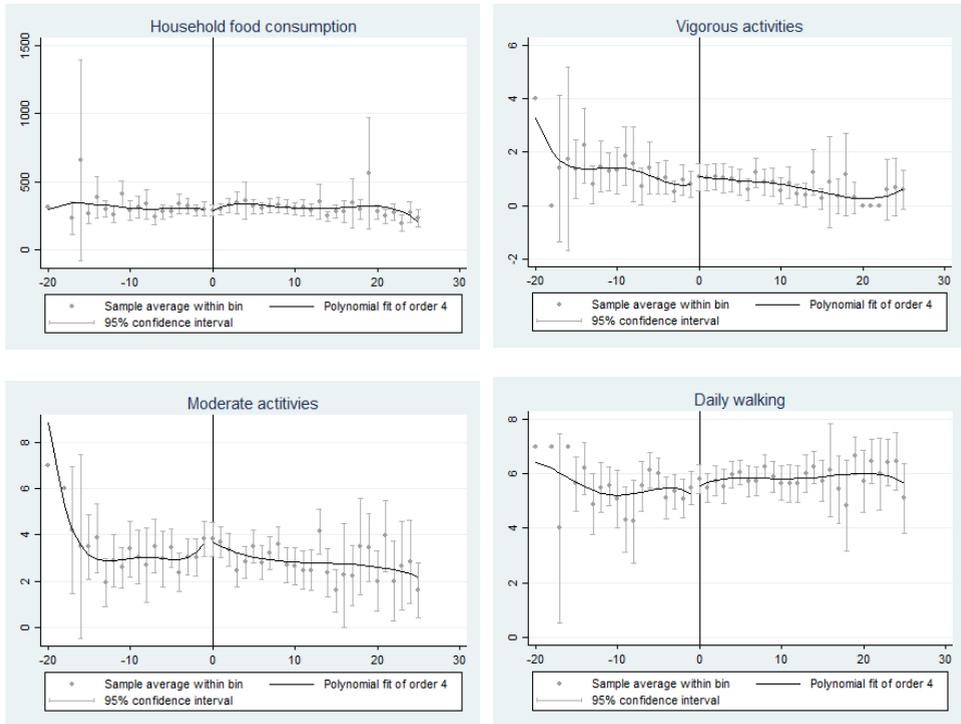


Figure 3.B.10: Food Consumption and Physical Activities by Normalized Age

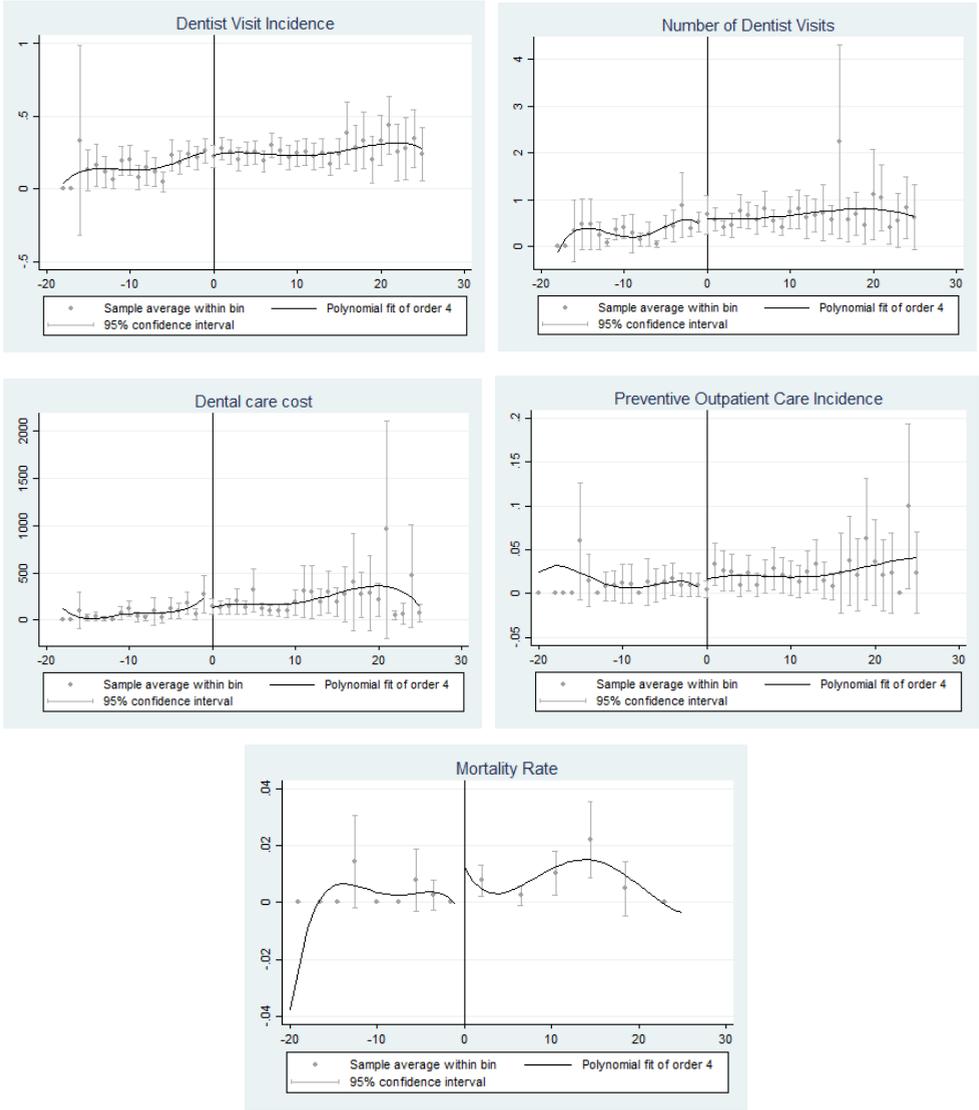


Figure 3.B.11: RD Plots of Additional Variables by Normalized Age

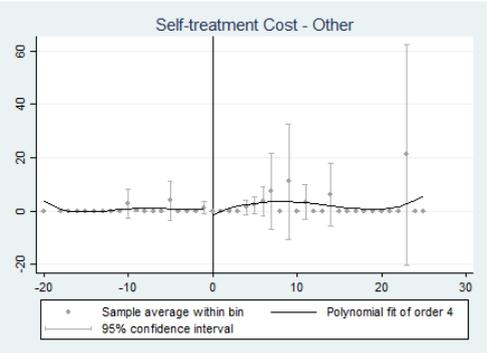
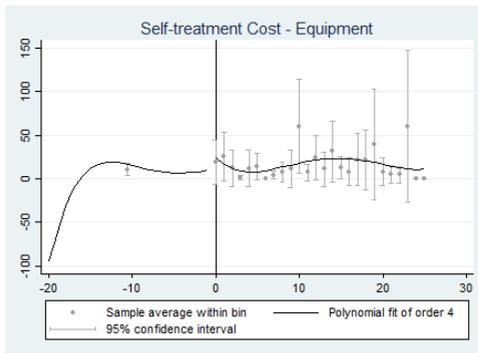
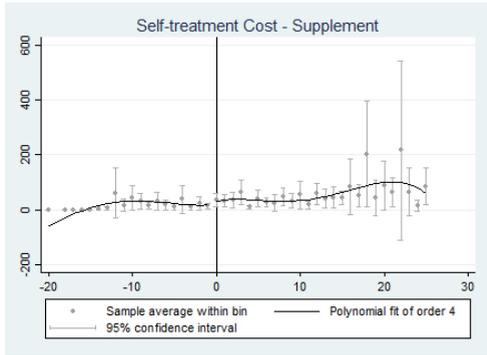
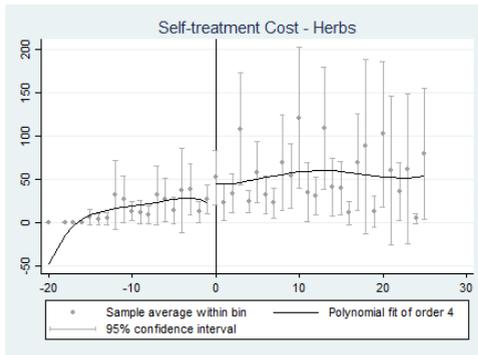
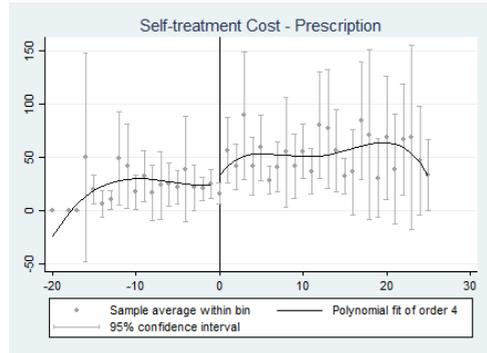
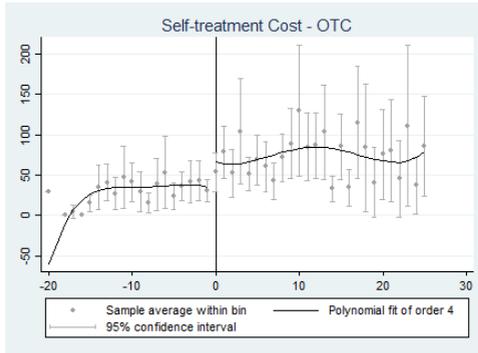


Figure 3.B.11: (Continued)

Appendix 3.C Variables Used in Further Analysis

In additional analysis, the variables that we used are defined as follows:

(1) Mental health: the total score on a scale from 8 to 32 is added up based upon 8 questions about negative feelings last week. The larger the mental health score, the worse the respondent's mental health.

(2) Life Satisfaction: self-reported life satisfaction on a scale from 1 to 5: 1: completely satisfied; 2: very satisfied; 3: somewhat satisfied; 4: not very satisfied; 5: not at all satisfied.

(3) Individual income (in RMB): individual yearly wage, pension and other income (e.g. transfer payments from government etc.).

(4) Chronic disease: whether the respondent has any of the following diseases: hypertension, dyslipidemia, diabetes, malignant tumor, chronic lung diseases, liver disease, heart disease, stroke, kidney disease, digestive disease, psychiatric problems, memory-related disease, arthritis or rheumatism, and asthma.

(5) Smoking: a dummy indicating whether the respondent smokes or not.

(6) BMI: Body Mass Index calculated with the formula: $BMI = \text{Weight in Kg} / (\text{Height in cm})^2$.

(7) Systolic blood pressure, and (8) Diastolic blood pressure measured in mmHg. Weight, height, and blood pressure are biomarkers which are measured during the interview.

(9) Diabetes, (10) Cancer, and (11) Stomach diseases: self-reported incidence of having diabetes, cancer, or stomach diseases.

(12) Pension: whether the respondent is accumulating or claiming a pension.

(13) Health insurance: whether the respondent is the policyholder/primary beneficiary of any type of health insurance.

(14) Mortality: a binary indicator for death between the two waves.

Variables (15) – (17) are only available in the second wave.

(15) Incidence of dentist visits: a binary indicator for visiting dentists in the past year.

(16) Number of dentist visits: the number of dentists visits in the past year.

(17) Dental cost: the total out-of-pocket cost of the dental care in the past year.

(18) Incidence of preventive outpatient care: whether the last doctor visit in the past month is for immunization, consultation or medical check-up.

(19) Incidence of working in public sector: whether currently work for or have retired from a government organization, institution, or 100% State owned firm.

References for Chapter 3

- Atalay, Kadir, and Garry F. Barrett. "The Causal Effect of Retirement on Health: New Evidence from Australian Pension Reform." *Economics Letters* 125.3 (2014): 392-395.
- Behncke, Stefanie. "Does Retirement Trigger Ill Health?" *Health Economics* 21.3 (2012): 282-300.
- Bejarano, Hernán, Hillard Kaplan, and Stephen Rassenti. "Effects of Retirement and Lifetime Earnings Profile on Health Investment." *ESI Working Paper* No. 14-21 (2014).
- Bertoni, Marco, Giorgio Brunello, and Gianluca Mazzarella. "Does Postponing Minimum Retirement Age Improve Healthy Behaviours Before Retirement? Evidence from Middle-Aged Italian Workers." *IZA Discussion Paper* No. 9834 (2016).
- Bonsang, Eric, Stéphane Adam, and Sergio Perelman. "Does Retirement Affect Cognitive Functioning?" *Journal of Health Economics* 31.3 (2012): 490-501.
- Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82.6 (2014): 2295-2326.
- Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocio Titiunik. "rdrobust: Software for Regression Discontinuity Designs." Working paper, University of Michigan, 2016.
- Che, Yi, and Xin Li. "Retirement and Health: Evidence from China." *China Economic Review* 49 (2018): 84-95.
- Coe, Norma B., and Gema Zamarro. "Retirement Effects on Health in Europe." *Journal of Health Economics* 30.1 (2011): 77-86.
- Coe, Norma B., Hans-Martin von Gaudecker, Maarten Lindeboom, and Jürgen Maurer. "The Effect of Retirement on Cognitive Functioning." *Health Economics* 21.8 (2012): 913-927.
- Coe, Norma B., and Gema Zamarro. "Does Retirement Impact Healthcare Utilization?." *CESR-Schaeffer Working Paper* 2015-032 (2015a).
- Coe, Norma B., and Gema Zamarro. "How Does Retirement Impact Health Behaviors? An International Comparison." *CESR-Schaeffer Working Paper* 2015-033 (2015b).
- Currie, Janet, and Brigitte C. Madrian. "Health, Health Insurance and the Labor Market." *Handbook of Labor Economics* 3 (1999): 3309-3416.
- Eibich, Peter. "Understanding the Effect of Retirement on Health: Mechanisms and Heterogeneity." *Journal of Health Economics* 43 (2015): 1-12.
- Fan, Jianqing. "Design-adaptive Non-parametric Regression." *Journal of the American Statistical Association* 87.420 (1992): 998-1004.

- Fé, Eduardo, and Bruce Hollingsworth. "Estimating the Effect of Retirement on Health via Panel Discontinuity Designs." Unpublished manuscript, University of Manchester / University of Lancaster, 2011.
- Galama, Titus, Arie Kapteyn, Raquel Fonseca, and Pierre-Carl Michaud. "A Health Production Model with Endogenous Retirement." *Health Economics* 22.8 (2013): 883-902.
- Godard, Mathilde. "Gaining Weight through Retirement? Results from the SHARE Survey." *Journal of Health Economics* 45 (2016): 27-46.
- Grøtting, Maja Weemes, and Otto Sevaldson Lillebø. "Health Effects of Retirement. Evidence from Norwegian Survey and Register Data." *UiB Working Papers in Economics* No. 02/17 (2017).
- Hagen, Johannes. "The Effects of Increasing the Normal Retirement Age on Health Care Utilization and Mortality." *Journal of Population Economics* 31.1 (2018): 193-234
- Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69.1 (2001): 201-209.
- Hernaes, Erik, Simen Markussen, John Piggott, and Ola L. Vestad. "Does Retirement Age Impact Mortality?" *Journal of Health Economics* 32.3 (2013): 586-598.
- Imbens, Guido W., and Thomas Lemieux. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142.2 (2008): 615-635.
- Insler, Michael. "The Health Consequences of Retirement." *Journal of Human Resources* 49.1 (2014): 195-233.
- International Labour Organization. *Key Indicators of the Labour Market*. International Labour Organization, Genève, 2018. <http://www.ilo.org/ilostat>. 1 June. 2018.
- Iparraguirre, José. "Physical Functioning in Work and Retirement: Commentary on Age-Related Trajectories of Physical Functioning in Work and Retirement—the Role of Sociodemographic Factors, Lifestyle and Disease by Stenholm et al." *Journal of Epidemiology and Community Health* 68 (2014): 493-499.
- Kim, Jinhee, Seung-Eun Cha, Ichiro Kawachi, and Sunmin Lee. "Does Retirement Promote Healthy Behaviors in Young Elderly Korean People?" *Journal of Behavioral Health* 5.2 (2016): 45-54.
- Kuhn, Michael, Stefan Wrzaczek, Alexia Prskawetz, and Gustav Feichtinger. "Optimal Choice of Health and Retirement in a Life-Cycle Model." *Journal of Economic Theory* 158 (2015): 186-212.
- Laaksonen, Mikko, Niina Metsä-Simola, Pekka Martikainen, Olli Pietiläinen, Ossi Rahkonen, Raija Gould, Timo Partonen, and Eero Lahelma. "Trajectories of Mental Health before and after Old-Age and Disability Retirement: a Register-Based Study on Purchases of Psychotropic Drugs." *Scandinavian Journal of Work, Environment & Health* (2012): 409-417.

- Lang, Iain A., Neil E. Rice, Robert B. Wallace, Jack M. Guralnik, and David Melzer. "Smoking Cessation and Transition into Retirement: Analyses from the English Longitudinal Study of Ageing." *Age and Ageing* 36.6 (2007): 638-643.
- Lee, David S., and Thomas Lemieux. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48.2 (2010): 281-355.
- Lei, Xiaoyan, Li Tan, and Yaohui Zhao. "The Impact of Retirement on Health: Evidence from China." Unpublished Manuscript, China Center for Economic Research, Peking University, Peking (2011).
- Mayer Brown JSM. *Guide to Employment Laws in the PRC*. Mayer Brown JSM, Hong Kong, 2008. https://www.mayerbrown.com/files/Publication/8cc25b33-7ada-49da-b444-cf697e1bc63f/Presentation/PublicationAttachment/79dcf986-2913-492b-97b3-ecb41a5f91b1/JSM_PRC_Employment_May2008.PDF. 17 Feb. 2017.
- McGarry, Kathleen. "Health and Retirement Do Changes in Health Affect Retirement Expectations?" *Journal of Human Resources* 39.3 (2004): 624-648.
- Mein, Gill, P. Martikainen, Harry Hemingway, and Michael G. Marmot. "Is Retirement Good or Bad for Mental and Physical Health Functioning? Whitehall II Longitudinal Study of Civil Servants." *Journal of Epidemiology and Community Health* 57.1 (2003): 46-49.
- National Bureau of Statistics of China. *National Data*. National Bureau of Statistics of China, Beijing, 2016. <http://data.stats.gov.cn/english/easyquery.htm?cn=C01>. 25 May. 2018.
- Neuman, Kevin. "Quit your Job and Get Healthier? The Effect of Retirement on Health." *Journal of Labor Research* 29.2 (2008): 177-201.
- Shai, Ori. "Is Retirement Good for Men's Health? Evidence Using a Change in the Retirement Age in Israel." *Journal of Health Economics* 57 (2018): 15-30.
- Shigeoka, Hitoshi. "The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection." *American Economic Review* 104.7 (2014): 2152-84.
- van der Heide, Iris, Rogier van Rijn, Suzan Robroek, Alex Burdorf, and Karin Proper. "Is Retirement Good for Your Health? A Systematic Review of Longitudinal Studies." *BMC Public Health* 13.1 (2013): 1180.
- World Health Organization. *Global Health Observatory Data Repository*. World Health Organization, Genève, 2018. <http://apps.who.int/gho/data>. 21 August. 2018
- Zantinge, Else M., Matthijs van den Berg, Henriëtte A. Smit, and H. Susan J. Picavet. "Retirement and a Healthy Lifestyle: Opportunity or Pitfall? A Narrative Review of the Literature." *The European Journal of Public Health* 24.3 (2014): 433-439.

Chapter 4

The Impact of a Disability Insurance Reform on Work Resumption and Benefit Substitution in the Netherlands⁶⁹

4.1 Introduction

In many western countries the number of disability benefit recipients and the share of the disability insurance (DI) program in the total public expenditure have grown in the past decades. During the past 50 years, benefit receipt in the Social Security Disability Insurance (SSDI) rose from less than 1 to 5 percent in the United States, and that in various DI programs rose from 1 percent to 7 percent in the United Kingdom (Autor et al., 2019). The Netherlands, in the early 2000s, became one of the countries with the highest fraction of disabled workers in the working population, as the total number of disability benefit recipients reached almost one million whereas the working population was around 7 million. During the period from 1990 to 2005, the total expenditure on disability benefits accounted for approximately 2.5 percent of the gross domestic product in the OECD countries, on average (OECD, 2010). The rapid expansion of DI programs raises concerns on the sustainability of public finance and labor market participation.

Governments seek to DI reforms to reduce disability benefit claiming and increase labor participation among the sick individuals. A large body of literature analysing different designs of DI programs and reforms (e.g., Autor and Duggan, 2003; Karlström et al., 2008; De Jong et al., 2011; Staubli, 2011; Campolieti and Riddell, 2012; Borghans et al., 2014; Burkhauser et al., 2014; Moore, 2015; Autor et al., 2016; Gruber, 2000; Campolieti, 2004; Maestas and Song, 2011; Kostøl and Mogstad, 2014; Deshpande, 2016; Mullen and Staubli, 2016; Fevang et al., 2017; Koning and van Sonsbeek, 2017; Ruh and Staubli, 2018) generally finds that restricting entitlement

⁶⁹ This chapter is coauthored with Tunga Kantarcı and Jan-Maarten van Sonsbeek. This research is supported by the Network for studies on Pensions, Aging and Retirement (Netspar) under grant number LMVP 2014.03. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of Netspar. We thank the UWV, and in particular Lucien Rondagh and Roel Ydema, for providing the sickness data. We thank Jennifer Alonso-García, Jochem de Bresser, Meltem Daysal, Pilar García-Gómez, Marike Knoef, Pierre Koning, Martin Salm, Jan van Ours, Arthur van Soest, seminar participants at Tilburg University, and conference participants at the Netspar Pension Day 2018, at the Netspar International Pension Workshop 2019, and at the 6th Workshop of the DGGÖ Health Econometrics for their helpful comments.

and reducing the benefit level increase labor participation and inflow into alternative benefit programs but that these effects are heterogeneous across subgroups of sick individuals.⁷⁰

We add to this strand of the literature by providing new evidence from a DI reform introduced in the Netherlands. In 2006, a new DI system, the Work and Income According to Labor Capacity Act (WIA) came into effect, as a successor to the Disability Insurance Act (WAO). The WIA reform introduces a basket of changes, among other things, extending the sickness benefit scheme that precedes the disability benefit scheme (i.e., the waiting period before sick individuals can apply for the benefit) from one to two years, tightening the criteria to enter the disability benefit scheme, and introducing financial incentives for work resumption at the more advanced stage of the disability benefit scheme, etc. The measures of WIA reform aims at providing strong incentives to facilitate the work resumption for sick individuals. For example, employers are obliged to compensate employees for wage loss during the waiting period for a maximum of two years instead of one year. To contain financial cost, employers may try harder to urge their employees back to work. Another example is that sick individuals with partial DI benefits in the new system will have to utilise their remaining working capacity to keep the benefit level from dropping.

However, concerns have been raised that the reform may also bring counter-incentives. For example, employers may become reluctant to hire workers with health problems as they have to bear a higher financial burden (Koning and Lindeboom, 2015). Sick individuals may face larger human capital depreciation during a longer waiting period, which may negatively influence their employment

⁷⁰ Instead of exploiting exogenous variation in benefit rules due to policy reforms, Chen and van der Klaauw (2008), Maestas et al. (2013), French and Song (2014), and Gelber et al. (2017) exploit variation in benefit eligibility rules or benefit levels imbed in disability benefit programs to analyze the causal effect of benefit incentives on benefit receipt and labor supply. A smaller literature investigates how incentive changes on the employer side influence disability benefit receipt. For example, De Jong and Lindeboom (2004) show that mandating firms to use preventive and reintegration measures to reduce sickness absenteeism does not decrease absence rates. Koning (2009) and Groot and Koning (2016) analyze the introduction and abolishment of experience rating for firms' disability insurance premium, and find that experience rating effectively decreases disability benefit receipt. Another strand of the literature pays attention to the non-economic outcomes of disability reforms (Dahl and Gielen, 2018; García-Gómez and Gielen, 2018). García-Gómez and Gielen (2018), for example, find that despite the gains in public finances, stricter eligibility criteria reduce life expectancy among women.

potential.⁷¹ With mixed economic predictions, it is a priori unclear how WIA reform would perform in reducing disability benefit claiming and encouraging work resumption.

This paper examines the overall effect of the WIA reform, considering all the measures of the reform as a whole, on sick individuals' labor market participation and social benefit claims, compared with the old DI system, WAO. Previous research analysing the impact of the stricter rules of the WIA regime in comparison to the rules of the older WAO regime is limited to Van Sonsbeek and Gradus (2013). Based on data on disability applications, they find that the stricter eligibility criteria of the WIA have led to a sharp fall in the number of new benefit claims, in addition to what has already been achieved through previous reforms, generating large budgetary savings for the government. The lack of research on the impact of the WIA reform is due to lack of data on sickness absence. The sickness benefit was reformed in 1994, 1996 and 2004 to mandate the employer to pay during the sickness period 70 percent of the earnings before sickness. Since no sickness benefit is paid by the government, but wage is paid by the employer, there is no registration of sickness absence by the government since these reforms. New reintegration regulations for employers were introduced in the sickness scheme in 2002 ("Gatekeeper protocol"). Only after this year the government started to register sickness cases to monitor whether employers comply with the new regulations. This has so far hampered evaluations of the disability reform directed at the sickness absence period. The existing research on the impact of the WIA reform is therefore limited in several respects. First, it is not analyzed to which extent the decrease in disability benefit use led to an increase in labor participation or use of benefits from alternative benefit programs such as the unemployment benefit. Second, it is not known whether the effects of the WIA reform are structural or fade in the long run, for example because people who do not first enter the WIA later become more sick and still become incapacitated for work. Third, little is known about how the effects of the WIA reform vary across subgroups of sick individuals.

In this study we exploit unique administrative data from the Employee Insurance Agency (UWV) on individuals who fell sick in the third and fourth quarters of 2003 and the first quarter of 2004. The two groups of individuals who fell sick in the last

⁷¹ A related study from the US, Maestas et al. (2015), finds that longer processing times reduce the employment and earnings of SSDI applicants for multiple years following the application.

two quarters of 2003 are insured under the old WAO scheme but are subject to different eligibility criteria, while the third group of individuals who fell sick in the beginning of 2004 is insured under the WIA scheme and is subject to additional and new eligibility criteria. To investigate the overall effect of the reform, we employ a difference-in-difference approach, comparing the labor market and benefit claiming behavior of the three groups of individuals before they fall sick and after they become eligible for disability benefits. The three groups of sick individuals are comparable in background characteristics, and economic shocks are likely to affect the behavior of the three groups in similar ways since eligibility for different disability schemes are determined by falling sick within very close proximity in time. This allows to attribute the differences in the labor market and benefit claiming behavior across the three groups to the differences in the rules of the disability benefit schemes that apply to these groups.

The regression results show that the WIA reform substantially reduced disability benefit receipt and the amount of disability benefits received. Individuals respond by increasing their labor participation and earnings. They also increase their use of unemployment benefit, and the amount they receive from unemployment benefit. Use of general assistance and small benefits decreases but the effects are small. The impact of the reform on disability benefit receipt and work resumption is persistent over time, while the spillover effect on alternative social security programs dies out in about seven years after individuals become eligible for disability benefits.

The reform is most effective among prime-aged workers and workers with regular contracts. Older individuals are less able to compensate the decrease in disability benefit receipt or income with higher labor participation or wages, whereas they more often rely on unemployment benefit. However, individuals who are unemployed at the time of falling sick are the worst affected by the WIA regime. Their disability benefits drop substantially, and their probability of working and their salary even decrease, compared with their counterparts insured under the WAO regime. Their worsening prospect of work resumption is possibly due to a larger scarring effect and more human capital loss as a result of staying in the waiting period for an extra year in the new DI system. Their loss of disability benefits and earnings is not compensated by increases in unemployment benefit or alternative benefits, which raises equality concerns for this vulnerable labor market group.

The paper proceeds as follows. Section 4.2 describes the institutional context. Section 4.3 describes the data. Section 4.4 presents descriptive statistics. Section 4.5 describes the empirical strategy. Section 4.6 presents the baseline results. Section 4.7 analyzes effects over time. Section 4.8 studies heterogeneous effects of the reforms. Section 4.9 presents sensitivity checks. Section 4.10 discusses policy implications and concludes.

4.2 Institutional setting

The general procedures for a DI applicant in the WAO system

Any individual insured with Dutch public DI enters the DI system as of the day of falling sick. Individuals falling sick before **1 October 2003** are subject to the old DI system, the Disability Insurance Act (WAO).⁷² The system comprises two stages (schemes), the sickness benefit stage (a.k.a. the sickness scheme, or the waiting period) and the DI benefit stage (disability scheme).

(1) Sickness benefit stage: An individual who earns wage or receives unemployment benefit is first admitted to the sickness scheme if he is unable to perform his work because of illness or injury irrespective of its cause. The maximum duration of the sickness scheme is **one year** as long as the individual remains sick. The employer is responsible to pay 70 percent of the former wage during the one year duration of the scheme. Most employers, however, pay the full amount of the former wage. For the unemployed sick individual in this stage, his unemployment benefit will be replaced by the sickness benefit that amounts to 70 percent of his former wage granted by the government.

The sick individual is invited to apply for the DI benefit at the end of the waiting period.⁷³ Case workers and experts will assess the applicant's health status and potential earnings, and calculate the "disability grade" which reflects the individual's lost earning ability. The disability grade is determined by dividing the estimated wage loss due to disability by the former wage, where estimated wage loss is given by the difference between the former wage and the potential wage that the sick

⁷² WAO came into effect in 1967 to insure against loss of earnings due to long-term disability. The act was amended several times but since the main amendments in 1993 it preserved its main features until the WIA reform.

⁷³ Sick individuals can file for DI application since 39 weeks as of falling sick. For WIA it is 87 weeks.

individual can still earn. An ergonomist determines the potential wage by taking the average of the highest wages the sick individual could still earn in three suitable occupations. The individual will be admitted to the DI benefit scheme only if his disability grade is at least **15 percent**.

(2) DI benefit stage:

The DI benefit stage has two sub-stages. The individual is first entitled to the “Wage-loss benefit” that replaces 70 percent of the former wage multiplied by the disability grade. The duration of the benefit depends on the age of the individual and is limited to a maximum of 6 years. When the Wage-loss benefit expires, the disabled individual is entitled to the “Follow-up benefit” that is lower than the Wage-loss benefit and pays the minimum wage and an additional amount that depends on the former wage and the age at which the individual has become entitled to the benefit. The benefit is paid as long as the individual is disabled but expires when the individual becomes entitled to the state pension.

The transitional WAO system

Before the WAO was abolished entirely, however, a transitional DI system was introduced on 1 October 2004 for people who have fallen sick during the period **from 1 October 2003 until 31 December 2003**.

The transitional DI system only changed one feature, compared to the old WAO system: the criteria to enter the DI benefit scheme have been made stricter. In particular, the transitional WAO has adopted a broader definition of what work can still be done by the applicant. Under the new definition, it is easier to find potential jobs that the individual can still perform. The wage loss due to disability can therefore be smaller, and the “disability grade” can be systematically lower. As a result, it is more difficult to reach the minimum disability grade to become eligible for the disability benefit, or to reach a higher disability grade that leads to a higher Wage-loss benefit.

The WIA system

The Work and Income Act (WIA) came into effect on 1 January 2006 for people who have fallen sick from **1 January 2004 onwards**. Besides inheriting the change made in the transitional WAO system, the WIA system further introduced major changes in both the sickness and disability schemes to facilitate work resumption.

(1) Sickness benefit stage: The maximum duration of this stage extends from one year to **two years**.⁷⁴ The strong incentive for the employer to facilitate work resumption is that the employer is obliged to compensate the employee for wage loss during the two years period of the scheme. The compensation must amount to 70 percent of the former wage. Most employers, however, pay the full amount of the former wage during the first year of sickness, and many pay more than 70 percent of the former wage during the second year of sickness.

WIA system inherited the stricter eligibility criteria of the disability scheme of the transitional WAO system that uses the broader definition of what work can still be done by the applicant. In addition, it increased the minimum grade of disability required to enter the DI scheme from 15 to **35 percent**. Therefore, workers with limited disability are expected to resume working with adaptations, or to apply for unemployment benefit.

(2) DI benefit stage:

The WIA system introduced a distinction between full and partial disability, and accordingly two specialised disability schemes. If the wage loss is more than 80 percent and there is no potential for any degree of recovery, the worker is admitted to the Full Invalidation Benefit Regulation (IVA), and is entitled to a benefit that replaces 75 percent of the former wage. Admission to the scheme is limited to a selective group of impairments that are expected to be permanent so that moral hazard is unlikely at least among this small group of workers.

If the wage loss is more than 35 percent and less than 80 percent, or if the wage loss is more than 80 percent but there is still a potential for recovery, the worker is insured under the Return to Work Regulation (WGA). The eligible worker is first entitled to the “Wage-related benefit”. Like the Wage-loss benefit of the WAO, the Wage-related benefit is related to the former wage. It replaces 70 percent of the former wage multiplied by the disability grade if the individual utilises his remaining work capacity to its full potential. The benefit has an **unemployment benefit component** that compensates the individual if he is not able to utilise his remaining work

⁷⁴ Strictly speaking, the extension of the sickness benefit from one to two years is not part of the WIA, but is part of a separate law, called Verlenging Loondoorbetalingsverplichting bij Ziekte (VLZ). However, both laws are part of the same package of reforms and their starting dates were synchronised so that both laws are effective from 1 January 2004 onwards.

capacity.⁷⁵ The duration of the benefit depends on the employment history, and is limited to a maximum of 38 months.

When the Wage-related benefit expires, the disabled individual is entitled to one of two types of benefits depending on whether he utilises more than 50 percent of his remaining earning capacity. If the individual utilises at least 50 percent of his remaining earning capacity, he is entitled to the “Wage-supplement benefit” which replaces 70 percent of the former wage multiplied by the disability grade. If the individual utilises less than 50 percent of his remaining earning capacity, he is entitled to the less generous “Follow-up benefit” which replaces 70 percent of the minimum wage multiplied by the disability grade. These mean that both the Wage-supplement and the Follow-up benefits make flat rate payments and hence disregard how much the individual is working below or above the threshold utilisation rate of remaining work capacity.

Both benefits are paid as long as the individual is disabled but expire when the individual becomes entitled to the state pension. At a given disability grade, the difference between the Wage-related benefit and the Follow-up benefit is as large as 70 percent of the difference between the former wage and the minimum wage, giving the partially disabled workers with higher former wages a stronger incentive to utilise at least 50 percent of their remaining work capacity when the wage-related benefit expires.

In addition, the WIA system extended the “experience rating” period. The experience rating refers to the differentiation in the premium firms pay to the disability insurance program. The premium amount is based on the costs of the disability benefits of the employees from the past. Hence, firms with high disability costs are punished with a higher premium. In the WAO, experience rating applied to employer contributions to disability insurance for a period of 5 years for all disabled workers. In the WIA, however, the experience rating period is extended to 10 years and applied to

⁷⁵ During participation in the disability scheme, the individual is eligible for the unemployment benefit (UB). The amount of the UB is a certain fraction of the remaining earning capacity. In the WAO, the individual is required to file an application to claim the UB. Therefore the DI and UB are always separate in WAO. In the WIA, however, the UB is integrated into the disability benefit, and therefore no application for UB is required. In fact, the duration of the Wage-related benefit is determined by the duration of the UB.

employer contributions for disabled workers participating in the WGA, but not in the IVA scheme.

The differences of the three DI systems are summarized in Table 4.1.

Table 4.1: Comparisons for different disability insurance systems

Disability insurance (DI) systems		WAO	Transitional WAO	WIA		
Applicable to individuals falling sick...		before 1 Oct. 2003	from 1 Oct. to 31 Dec. 2003	as of 1 Jan. 2004		
Sickness benefit stage	Max. length of the waiting period	1 year		2 years		
	The way to calculate the disability grade (DG)	Easy to have a high DG	Systematically harder to get a high DG			
	Min. DG required for DI	15%		35%		
DI benefit stage	First-stage DI benefit	Wage-loss benefit = $70\% \times \text{previous wage} \times \text{DG}$		WGA: partially disabled	IVA: permanently disabled	
		Max. 6 years		Wage-related benefit = $70\% \times \text{previous wage} \times \text{DG} + \text{Unemployment benefit component}$	75%*previous wage; Max. until state pension age	
	Second-stage DI benefit	Follow-up benefit < wage-loss benefit		Max. 38 months		
				Wage supplement		
				If use \geq 50% of the remaining working capacity:		If use < 50% of the remaining working capacity:
		Same as Wage-related benefit in the 1st stage	Minimum wage* DG			
	Max. until state pension age		Max. until state pension age			
	Experience rating period for the employer	5 years		10 years		

4.3 Data

Data source

We use unique administrative data from the Employee Insurance Agency (UWV) on three cohorts of sick people who face different criteria to enter the disability scheme and different incentives to resume working if participating in the disability scheme. In particular, the data contains information on individuals who fell sick in the third quarter of 2003, fourth quarter of 2003, and the first quarter of 2004 and therefore became eligible to participate in the WAO, transitional WAO, and the WIA schemes, respectively.⁷⁶ For these people we observe the beginning and ending dates of sickness, gender, date of birth, etc. These people either earn wage or receive unemployment benefit at the time they fall sick since people of other labor market groups are not eligible to enter the sickness scheme. For people in employment, we observe whether they hold a regular contract, temporary contract, or a contract through a temporary work agency.

We merge the administrative data on sickness with administrative data on labor participation, salary, and benefits, all available on a monthly basis from Statistics Netherlands. The benefits are from various benefit schemes which include the disability insurance (DI), unemployment benefit (UB), general assistance (GA) for low-wage earners, and other benefits (OB) from a large number of smaller benefit programs.⁷⁷ The data from Statistics Netherlands extend from January 1999 to December 2015 and allow to study the differences in benefit claiming and labor market behavior of the three cohorts of sick individuals over a long period of time.

⁷⁶ According to the “Gatekeeper protocol”, it is not compulsory to report sickness cases that last shorter than 13 weeks. So the administration data of sickness registration do not necessarily capture all the short-term sickness cases.

⁷⁷ We also made two checks using available yearly data. First, we checked if the probability of being self-employed and the self-employment earnings differ across cohort over calendar years. We find that all the three cohorts has increasing possibility to be self-employed, and their self-employment earnings are also increasing. But the increasing trend is no significantly different across cohorts. Second, we check the receipt and yearly amount of other benefits lumped with the sickness benefit which was paid by the government to sick individuals without an employer during the waiting period. We find that in the second year of the waiting period (year 2005 to 2006) for the WIA cohort (1st quarter of 2004), the receipt and amount of this broader definition of other benefits have a large increase due to the extension of the waiting period and the mechanic increase of the sickness benefit. The analysis on these yearly data are not included in the current analysis. But the descriptive plots of them can be found in Appendix 4.B.

Sample restrictions

The initial sample of sick people consists of 51,319,668 observations for 251,567 individuals. We impose a number of restrictions on the initial sample.

First, we drop sickness cases last shorter than 180 days in the data to ensure that the three cohorts of sick individuals are comparable in the number of days spent in sickness. We drop sickness cases that are shorter than 90 days because only sickness cases longer than that duration are mandatory to be reported, according to the Gatekeeper Protocol. By checking the distribution of the sickness duration across individuals falling sick in different months, we also detected that sickness cases between 90 to 180 days are slightly under-reported among individuals who fell sick in the third quarter of 2003, in particular, in July and August.⁷⁸ To ensure a comparable distribution of sickness duration across three cohorts of sick individuals, we further drop the sickness cases that last shorter than 180 days. This restriction leads to a sample of 19,690,488 observations for 96,522 individuals. Although this restriction reduces more than half of the observations, the long sickness spells are of the main interest as they account for a major share of benefits paid. On the other hand, focusing on relatively long sickness cases may raise concerns on overlooking behavioral responses early in the waiting period. In sensitivity analysis, we restrict sample to alternative duration of sickness cases to see how this concern together with the under-reporting issue would affect the results.

Second, we only keep individuals who do not have a record of DI benefit before. Existing DI recipients can blur the definition of groups. For example, a partially disabled individual with a WAO benefit can still fall sick when working part time under the WIA regime (e.g. fall sick in the 1st quarter of 2004). Then it is less clear if we should assign him to the WAO group or the WIA group. To abstract away from the complications by the existing DI recipients, we only focus on newcomers to the DI system. This restriction leads to a sample of 16,028,076 observations for 78,569 individuals.

Third, we drop individuals if they are participants of the disability schemes for the self-employed (WAZ) or young people (WAJONG) since the institutional rules and incentives for work resumption are very different for them. This restriction leads to

⁷⁸ See Appendix 4.A for checks on the distribution of sickness duration across cohorts.

a sample of 15,864,876 observations for 77,769 individuals and constitute the study sample.

Treatment and control groups

As described above, sick individuals become eligible to participate in one of three disability schemes depending on the date they fall sick. This allows to construct control and treatment groups and compare their responses to the disability reform in a quasi-experimental research design. In particular, we categorise the sick individuals into three groups: the WAO group, the transitional WAO group, and the WIA group, which consist of individuals who fell sick in the third quarter of 2003, fourth quarter of 2003, and first quarter of 2004, respectively. We consider the WAO group as the control group, and the transitional WAO and the WIA group as the treatment group 1 and 2, respectively.

Note that the assigned group for a given individual is not changing over time. For an individual who falls sick in the third quarter of 2003, he is forever in the WAO group, even if he does not get a WAO benefit in the end, or even if he falls sick again years later and enters the WIA scheme. This definition keeps groups comparable, but influences the interpretation of comparison between control and treatment group. Our WAO group has 26111 individuals. 9332 individuals have ever claimed DI. Among them, 2208 individuals have claimed WIA benefit for at least one month. The majority of DI recipients in WAO group claim WAO benefit in most of the time.

Another complication to the definition of the control group is the “re-examinations” of the WAO participants younger than age 50 on 1 July 2004 that have taken place from 2004 until 2008 based on the eligibility criteria of the transitional WAO scheme (Mandicó et al., 2018). As a result of this re-examination, the majority of the control group gradually subjects to the same rules as transitional WAO group.

Therefore, when we compare WIA group to the control group, what we capture is not the “pure” difference between WIA and WAO system, but rather the difference between WIA and a mixture of WAO, transitional WAO, and a small proportion of WIA. The “pure” difference between WAO system and WIA system throughout the sample periods would be larger in absence of these complications.

Outcome variables

Based on the available data on labor and benefits, two sets of five outcome variables are defined and used to compare behavioral responses of control and treatment groups to the disability reform. The first set considers dummy variables indicating the labor participation and the benefit receipt of disability benefit (DI), unemployment benefit (UB), general assistance (GA), and other benefits (OB). The second set considers monthly income from labor and the four benefit programs. The income variables are **unconditional** on participation or receipt. That is, the income equals 0 if there is no labor participation or benefit receipt.

However, there is a complication for the definitions of three variables: “Amount of DI received”, “Receive UB or not”, and “Amount of UB received”. As described in Section 4.2, for DI recipients, UB is integrated into the DI in the WIA, while this is not the case in the WAO. As a result, the amount of DI, as well as the receipt and the amount of UB are no longer comparable across groups. We therefore use the following adjusted definitions:

The amount of DI is defined as: **the amount of DI and possibly also UB at the same time**. This means that for WAO recipients, this variable is the total amount of DI plus UB, while for WIA recipients, this variable is just the amount of DI (the UB component is already in the amount).⁷⁹ For an individual who does not have a DI, the variable equals 0 no matter he has a UB or not.

The receipt of UB indicates **the receipt of UB but not DI** at the same time. The amount of UB refers to the amount of UB without DI, i.e. it equals the amount of UB when the “receipt of UB but not DI” equals 1.

4.4 Descriptive statistics

Time trends in outcome variables

Figures 4.1a and 4.1b show the time profiles of labor participation and benefit receipt and income for control and treatment groups over a period of 17 years from January 1999 to December 2015. A time profile of a given group is generated as follows. First, within a group and in a given calendar month, the mean of the outcome variable (dummy variable that indicates labor participation or receipt of a benefit, or income

⁷⁹ This complication does not influence the receipt of DI. But strictly speaking, the receipt of DI can also be interpreted as the receipt of DI and possibly UB.

from work or benefits) is calculated. The set of means calculated for each month of the 17 year period are then used to plot the time profile. In the plots vertical lines are added at the first instance individuals could become entitled to the sickness and disability benefits in the WAO, transitional WAO or WIA scheme.

The top left plot in Figure 4.1a shows the time profile of probability of receiving DI. All groups do not receive a DI until waiting period expires.⁸⁰ A large fraction of about 20 percent of the sick individuals in the WAO and the transitional WAO cohorts claim DI when they become eligible for DI. This fraction shows no notable change until the end of the study period. The time trends of the WIA and WAO groups are similar, except in two respects. First, the inflow into the disability scheme for the WIA group is not as immediate as it is for the WAO groups when individuals become eligible for DI. It might be that stricter rules of the WIA make it difficult to claim the DI at the first attempt. Second, the time trends for the WAO groups show a decrease during the first year after they show a peak when these groups become eligible for DI. It might be that the WAO recipients have better chances of recovery shortly after entering the disability scheme. The WAO recipients are relatively healthier than the WIA recipients since in the WIA the minimum disability grade to enter the disability scheme is higher but also disability assessment is stricter.

Note that from year 2008 onwards, the lines for WAO and transitional WAO gradually converge. This can be explained by the re-examinations of the WAO participants from 2004 until 2008 based on the eligibility criteria of the transitional WAO scheme, as mentioned in Section 4.3. Due to the re-examination, the majority of the control group gradually became subject to the same rules as transitional WAO group. And we indeed observe this in the plot.

The probability of working shows a strong time trend that is common to both the control and treatment groups. It increases until the date individuals fall sick. This pattern does not reflect behavioral responses but it reflects the fact that individuals can enter the sickness scheme, and get reported as sick in the administrative data, only if they are working or receiving the UB at the time they fall sick (Section 4.3).

⁸⁰ A very small share receives DI slightly before the waiting period expires (the blue lines). In principal, individuals can apply for DI since 39 weeks as of falling sick for WAO and transitional WAO, 87 weeks for WIA. We allow for early recipients up to the earliest application time. Results are robust to dropping these early recipients.

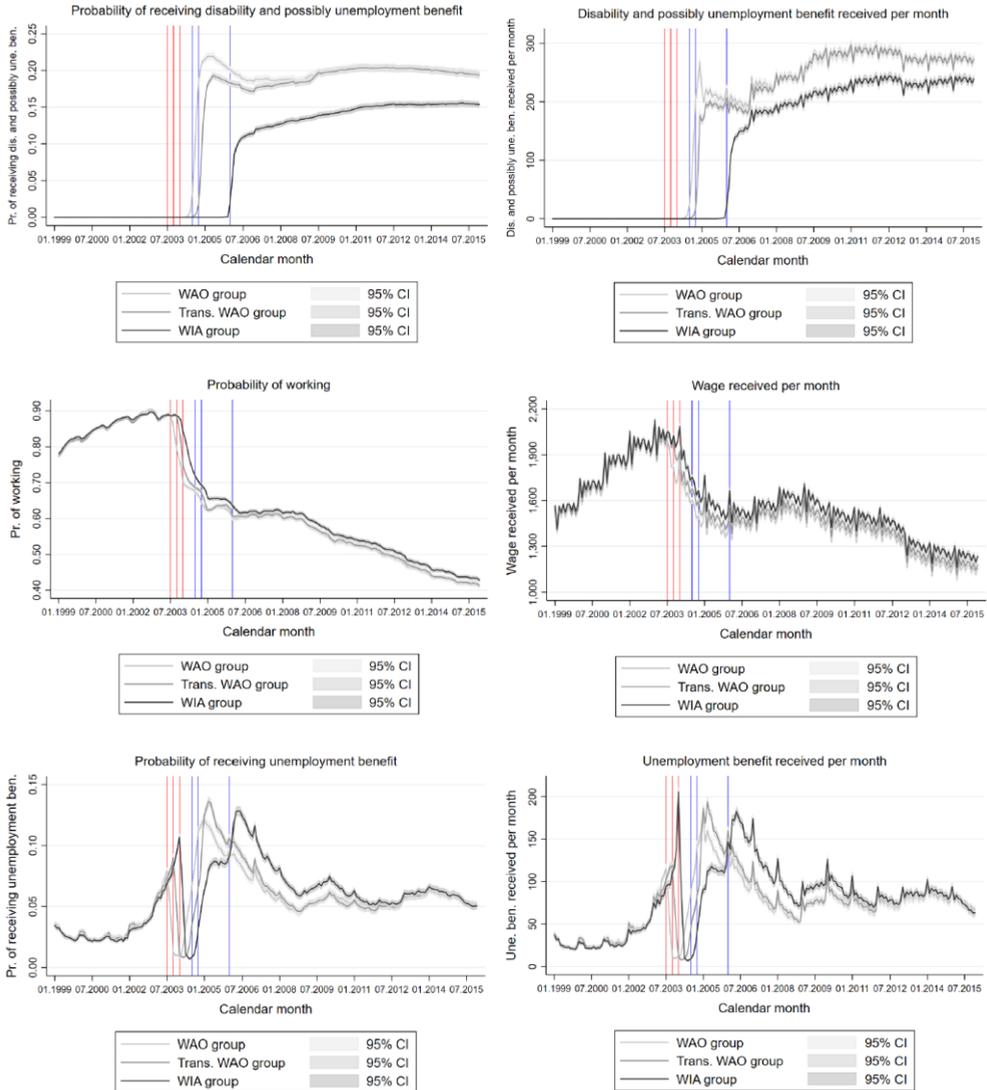


Figure 4.1a: Labor participation, benefit receipt (disability and unemployment), labor income and benefit income for control and treatment groups over calendar months

Note: In a given plot, a point on a given time profile represents the mean of the outcome variable (dummy variable that indicates labor participation or benefit receipt, or income from work or benefits) within a group (control or treatment) in a given calendar month. Around the mean is a 95 percent confidence interval. Disability benefit might be supplemented with unemployment benefit. Unemployment benefit is defined so that receiving disability benefit at the same time is not allowed. Each plot is based on the study sample of 15,864,876 observations for 77,769 individuals who fell sick during the period from July 2003 until March 2004. Vertical lines indicate the first instance individuals could become entitled to the sickness (red) and disability (blue) benefits in the WAO, transitional WAO and WIA schemes.

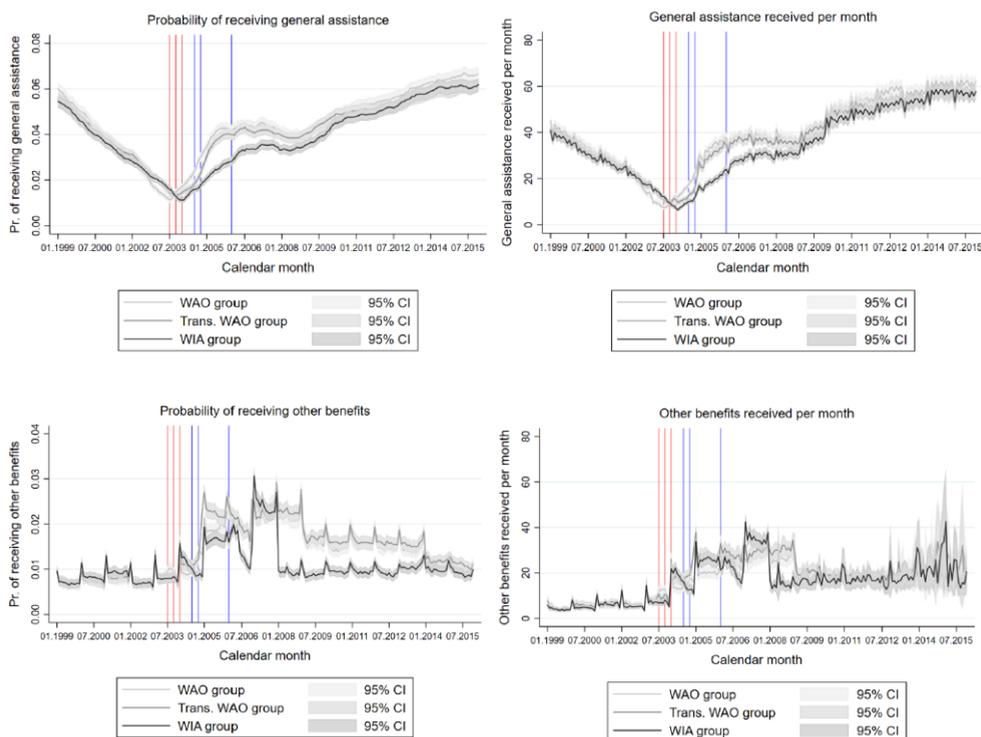


Figure 4.1b: Benefit (general assistance and other benefits) receipt and income for control and treatment groups over calendar months

Note: In a given plot, a point on a given time profile represents the mean of the outcome variable (dummy variable that indicates receipt of a benefit, or income from benefits) within a group (control or treatment) in a given calendar month. Around the mean is a 95 percent confidence interval. Each plot is based on the study sample of 15,864,876 observations for 77,769 individuals who fell sick during the period from July 2003 until March 2004. Vertical lines indicate the first instance individuals could become entitled to the sickness (red) and disability (blue) benefits in the WAO, transitional WAO and WIA schemes.

Before this time, these individuals can have another status outside the labor force (e.g. study in school). The probability of working decreases dramatically during the first year of sickness, remains fairly stable for about three years, but decreases further throughout the remaining months of the study period (e.g. people can gradually retire and exit labor market).

Unemployment benefit use is strongly related to the use of sickness and disability benefits. For the WAO groups, during the sickness period UB use decreases sharply because unemployed people who fall sick change their UB for sickness benefit. UB use rebounds thereafter because many of these people recover during the sickness

period. It peaks as individuals become eligible for DI because when the sickness period ends, those who apply but get rejected to enter the disability scheme turn to the UB. UB use decreases during the disability period. This is because UB is temporary and ends after a maximum of 38 months. For the WIA group a similar time trend is observed except that UB use increases further during the second year of sickness before it peaks when individuals become eligible for DI since members of this group have more time for recovery during the longer sickness period, and change their sickness benefit for UB.

The probability of receiving general assistance increases steadily from the time individuals fall sick. A similar but less pronounced increase is observed for the probability of receiving other benefits from various small benefit programs from the time individuals fall sick. These time patterns appear to be related to that of working. As individuals work less due to sickness, their earnings decrease enough to become entitled to these benefits.

In the right panels of Figures 4.1a and 4.1b, we show the time profiles of the means of income from work and benefit programs. In a mean calculation, we allow for zero income, therefore the effect on these income variables is a combination of effects on extensive margin (participation/receipt) and intensive margin (income conditional on participation/receipt). The time trends of the means of different types of income resemble those of participation in the labor market and benefit receipt in the left panels of the figures.

In most of the plots, the groups show similar time trend before falling sick. A closer look into the pre-sickness period show that, the lines of WAO group is slightly lower than other groups for the receipt and amount of GA between January 2002 and June 2003. And the lines of WAO group's labor participation and the amount of UB are slightly higher than other groups between January 2002 and January 2003, and between January and June 2003, respectively. Though the magnitude of the difference is very small, in section 4.9, we do robustness checks on this slightly deviating time trend.

Descriptive statistics before, during, and after the waiting period

Table 4.2 summarizes the sample mean of background characteristics and outcomes in control and treatment groups.

Table 4.2: Sample means of background characteristics and outcomes

Groups	Before			Waiting period			After		
	WAO (1)	Trans. WAO (2)	WIA (3)	WAO (4)	Trans. WAO (5)	WIA (6)	WAO (7)	Trans. WAO (8)	WIA (9)
Background characteristics									
Age of falling sick	39.690	40.137	40.628						
Female	0.441	0.437	0.440						
Foreign-born	0.188	0.177	0.170						
Work characteristics									
Regular contract	0.581	0.574	0.593						
Temporary contract	0.138	0.130	0.119						
Temporary contract via agency	0.048	0.049	0.034						
Unemployed	0.188	0.204	0.213						
Other	0.044	0.043	0.041						
Labor participation and benefit receipt									
DI	0	0	0	0.002	0.001	0.000	0.198	0.191	0.141
Work	0.858	0.856	0.860	0.733	0.725	0.689	0.534	0.536	0.539
UB	0.033	0.036	0.037	0.030	0.030	0.058	0.064	0.065	0.069
GA	0.036	0.035	0.032	0.018	0.017	0.021	0.050	0.048	0.047
OB	0.008	0.008	0.008	0.009	0.011	0.014	0.017	0.018	0.012
Income from labor and benefit programs									
DI	0	0	0	1.528	1.241	0.214	253.895	250.790	212.504
Salary	1768.763	1792.750	1820.622	1731.049	1731.576	1650.255	1376.214	1407.520	1453.732
UB	40.620	43.528	45.717	36.775	35.457	78.210	86.192	90.183	96.359
GA	27.657	26.730	25.061	12.260	11.763	15.194	47.287	44.936	44.009
OB	5.650	6.486	5.922	12.884	16.878	21.613	20.346	23.393	20.335
Individuals	26,111	26,217	25,441	26,111	26,217	25,441	26,111	26,217	25,441
Observations	1,439,381	1,518,527	1,551,421	313,332	314,604	610,584	3,573,931	3,515,137	3,027,959

Note: “Before” refers to the all the months from January 1999 to the month before falling sick. Each individual falls sick in different month, therefore the lengths of the “Before” period are different for each individual. “Waiting period” refers to the 12 months as of falling sick for WAO group and transitional WAO group, while 24 months as of falling sick for WIA group. For WAO and transitional WAO, “After” means the 13th month as of falling sick and onwards until December 2015. For WIA, it means the 25th month as of falling sick and onwards. The background characteristics are collected when individuals fall sick. They are time invariant, and averaged across individuals, rather than individual-calendar month observations.

The top panel of the table presents the sample means of a number of background characteristics in control and treatment groups. A sample mean is calculated as the average of a given characteristic of all individuals in a given group at the time these individuals fall sick. In all groups, the average age of falling sick is almost 40, the fraction of men is slightly higher than that of women, and the majority of the sample is native-born. About 60 percent hold a regular work contract, less than 20 percent

hold a temporary contract or a contract through a temporary work agency, and about 20 percent is unemployed. Sample means of background characteristics slightly differ across groups. In our formal empirical analysis, we control for time-invariant characteristics, through individual fixed effects.

The bottom panel of Table 4.2 presents the sample means of labor participation and benefit receipt and income from labor and benefit programs for control and treatment groups before falling sick, during the waiting period, and after the waiting period expires.

“Before” refers to months from January 1999 to the month before falling sick. Each individual fell sick in different months, therefore the lengths of the “Before” period are different for each individual. “Waiting period” refers to the 12 months as of falling sick for WAO group and transitional WAO group, while 24 months as of falling sick for WIA group. “After” refers to months after the waiting period expires, which is the institutional time when an individual becomes (potentially) eligible for DI.⁸¹ For WAO and transitional WAO, “After” means the 13th month as of falling sick and onwards until December 2015. For WIA, it means the 25th month as of falling sick and onwards.

During the “Before” period, the mean differences are small across groups. In the formal empirical analysis, these mean differences in the “before” period can be controlled for by using a difference-in-difference approach. In “After” periods where people are eligible for DI, the probability of receiving DI for WIA is lower than WAO and transitional WAO by 0.05. Salary and the amount of UB in WIA are slightly larger than in WAO and transitional WAO, netting out the difference in the “Before” period. For other variables, the differences are less noticeable.

Sample means differ across groups already in the waiting period, suggesting that measures of the reform (e.g. extending the waiting period) may already play a role in the waiting period. However, these group differences cannot be interpreted as the effect of the reform during the waiting period, because the calendar month shocks in WIA’s second year of waiting period does not happen in WAO and transitional WAO.

⁸¹ We use “potentially” because some individuals may recover before they can apply for DI, so that eventually they are not really “eligible” for a DI. But the institutional time that they can be potentially eligible for DI still applies to them.

We need an empirical strategy that can account for effects of measures both within and after the waiting period.

4.5 Empirical strategy

We use a difference-in-differences (DID) approach to identify the causal effects of the transitional WAO and the WIA reforms on labor participation and benefit receipt and on income from labor and benefit programs. The first difference is between treatment and control groups. Treatment groups are those who fall sick in the last quarter of 2003 or the first quarter of 2004, subject to WAO and transitional WIA schemes. We compare them to individuals in the control group, who fall sick in the third quarter of 2003, and are subject to a less restrictive benefit regime WAO.

The second difference is between “before” and “after”. Since individuals fall sick at different calendar months from July 2003 to March 2004, we do not compare individuals over calendar time. Instead, we compare people over “event time”. The “before” period in “event time” refers to all months before an individual falls sick, where all the DI systems are not yet available.

An intuitive definition of “after” period is “since the onset of sickness”, so that we compare individuals at the same time elapsed since falling sick. However, this method is problematic for WIA reform. Because WIA extends the waiting period from one year to two years, the main incentive of receiving DI benefit kicks in one year later than in (trans.) WAO group. Comparing people since falling sick will end up comparing WIA people in the second year still waiting for DI application with WAO people already claiming DI for half a year. This creates difficulties in interpreting the effect of the reform.

To make sensible comparisons, we compare people at the end of the waiting period, when DI incentives kick in for all groups. Figure 4.2 shows the conceptual plot for this idea.

Had it not been the WIA reform, the potential outcome of the Treatment group 2 should have been parallel to that of the control group. And treatment group 2 would have entered DI one year since falling sick just as WAO people do. Now under WIA reform, WIA group has to stay in the waiting period for two years. The difference between WIA people’s outcome at 2 years since sick (the black dot) and the WAO

people’s outcome at 1 year since sick (the grey dot), is the joint effort of both extending the waiting period, and other reform measures like tightening the criteria. So comparing WIA people’s outcome since their two-year waiting period expires with WAO people’s outcome since their one-year waiting period expires, gives the total effect of reform measures both within and after the waiting period.

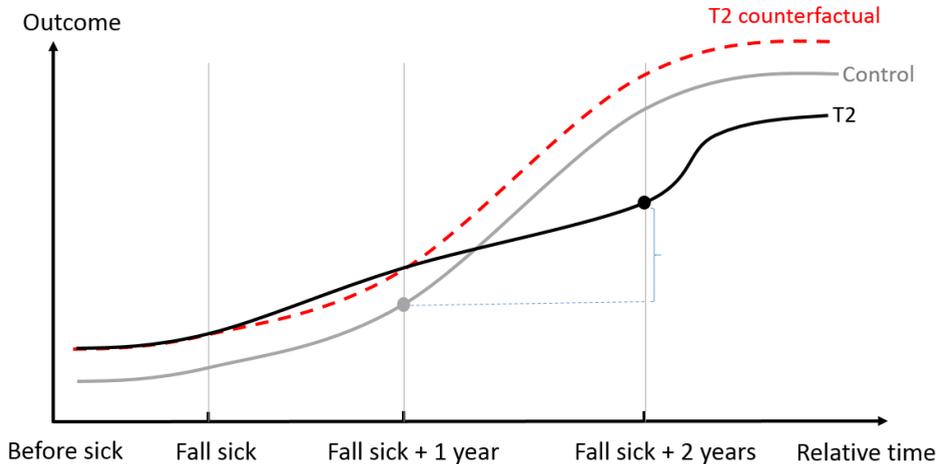


Figure 4.2: Conceptual plot for the empirical design

Therefore we define “after” period as: one year since falling sick and onwards for WAO and transitional WAO; Two years since falling sick and onwards for WIA. In other words, this “after” period starts when the waiting period expires, which is also the institutional time of (potentially) being eligible for DI.

In order to compare individuals along event time, we drop the observations in the “waiting period”, only keeping data in “before” and “after” periods.

We implement the DID comparison using the following regression:

$$y_{it} = \gamma_1(Treat_i^{Trans. WAO} \times Post_t) + \gamma_2(Treat_i^{WIA} \times Post_t) + \delta Post_t + \lambda_{it} + \alpha_i + \varepsilon_{it} \quad (4.1)$$

i indexes individuals.

t indexes the months of event time. t indexes the months of the “before” period with values from -61 to 0 . E.g. $t = 0$ means the last month before falling sick. 1 to 138 refer to the “after” period. E.g. $t = 1$ means the first month as of being (potentially) eligible for DI. Due different lengths of waiting period, $t = 1$

corresponds to the 25th month since falling sick for WIA, the 13th month since sick for WAO and transitional WAO.

y_{it} is the outcome variable of interest.

λ_{it} is a vector of calendar month dummies from January 1999 to December 2015.⁸² It is individual and event-month specific because individuals enter the same calendar month at different event months.⁸³

α_i is an individual-specific, time-invariant intercept term.

ε_{it} represents the individual-specific, time-varying shocks that are not observed.

$Treat_i^{Trans. WAO}$ and $Treat_i^{WIA}$ are dummy variables that indicate the treatment groups, i.e. the transitional WAO and the WIA groups, respectively. We do not control for two treatment group dummies because they are time-invariant and therefore absorbed individual fixed effects α_i .

$Post_t$ is a dummy variable that indicates the “after” period.

We interact $Treat_i^{Trans. WAO}$ and $Treat_i^{WIA}$ with $Post_t$ to capture the mean difference in the outcome variable between the treatment and control groups during the “after” period compared to the mean difference between the two groups during the “before” period. In this comparison, the latter difference aims to account for differences between the groups due to factors other than the policy reform. γ_1 and γ_2 are the coefficients of main interest and reflect the effects of the transitional WAO and WIA reforms. Standard errors are adjusted for clustering at the individual level.

The assumptions needed to obtain unbiased estimates of these coefficients are (1) “random” timing of falling sick: the time of falling sick are not correlated with unobserved factors that both influence the assignment of the group and the outcome. (2) Common trend assumption: the potential outcomes of treatment (in absence of reforms) and control groups are parallel to each other. We provide evidence and robustness checks for these assumptions in Section 4.9.

⁸² January 1999 is set as the base month. Due to dropping the data in waiting period, no observation is in March to June 2004.

⁸³ Strictly speaking, individuals that fall sick in the same month share the same λ_{it} .

4.6 Main results

Here we present the baseline DID estimates of the effects of the transitional WAO and WIA reforms based on Equation (4.1). Panel A of Table 4.3 presents the results for labor participation and benefit receipt, and Panel B of Table 4.3 presents the results for income from labor and benefit programs. The estimated effects of the transitional WAO and the WIA reforms are always interpreted as the effects of the new rules of the transitional WAO and WIA regimes compared to the old rules of the WAO regime.

The effects of transitional WAO and WIA show similar pattern: the receipt and the amount of DI claim reduce, while labor participation and salary increase, as intended by the policy design. But not everyone successfully resume working. The receipt and amount of unemployment benefit (UB) also increase. The effect on general assistance (GA) and other benefits (OB) are very limited.

Table 4.3: Main results

Panel A:	DI receipt	Labor Participation	UB receipt	GA receipt	OB receipt
Trans. WAO × Post	-0.008** (0.003)	0.007** (0.003)	0.003** (0.001)	-0.003* (0.001)	0.001*** (0.001)
WIA × Post	-0.058*** (0.003)	0.018*** (0.004)	0.014*** (0.001)	-0.004*** (0.002)	-0.005*** (0.001)
Post	0.203*** (0.002)	-0.368*** (0.003)	0.010*** (0.001)	0.009*** (0.001)	0.003*** (0.001)
Panel B:	DI amount	Salary	UB amount	GA amount	OB amount
Trans. WAO × Post	-5.528 (4.300)	19.284 (12.802)	5.319*** (1.882)	-2.841* (1.487)	2.262* (1.225)
WIA × Post	-50.813*** (4.413)	55.659*** (13.332)	22.824*** (1.952)	-4.898*** (1.550)	0.248 (1.230)
Post	282.124*** (3.411)	-389.573*** (11.262)	17.368*** (1.838)	19.042*** (1.371)	11.393*** (2.721)
Calendar month dummies	Yes				
Individual fixed effects	Yes				
Individuals	77,769				
Observations	14,626,356				

Notes: DI stands for disability insurance benefit with possibly unemployment benefit at the same time. UB stands for unemployment benefit without disability benefit. GA and OB are general assistance and other benefits, respectively. All regressions employ the linear regression model with fixed effects given by Eq. (4.1) and include calendar month dummies with January 1999 as the base month. The “before” and “after” are periods of event time and correspond to the months before individuals fall sick, and subsequently the months after individuals become eligible for disability benefits. Standard errors, in parentheses, are adjusted for clustering at the individual level. ***, **, * indicate statistical significance at the 0.01, 0.05, 0.10 levels, respectively.

The effects of transitional WAO reforms are much smaller than WIA reform, not only because transitional WAO reform is only one reform measure nested in WIA reform, but also because the re-examination during 2004 to 2008 assimilates the WAO and transitional WAO group. Had there not been the re-examination, we would see a larger effect of transitional WAO.

The transitional WAO reform reduces the probability of receiving DI by 0.8 percentage points, The WIA reform, as our main focus, reduces the probability of receiving DI by 5.8 percentage points, on average, during the years after the reform has come into effect. The reduction of 5.8 percentage points corresponds to a 28.6% drop (0.058/0.203) in DI receipt for the control group. Van Sonsbeek and Gradus (2013) also find large effect of WIA. They showed that due to the WIA reform the number of disability benefit awards in the working population decreased by 40 percent at the onset of the reform, although the impact of the reform slightly decreased over time.

A back-of-envelope calculation in Appendix 4.C show that half of the reduction of DI receipt in WIA comes from the reduction of observations who claim partial DI benefit while work. This is consistent with the policy design of limiting the access to partial DI benefit and encouraging people with mild disability to work. The rest comes from the reduction of people who claim DI benefit but do not work. They can be permanent disabled people that no longer work, or people who claim partial DI but cannot find a job for the remaining working capacity. The reduction of the latter group can result in a larger inflow into unemployment benefit.

Both reforms increase labor participation. The probability of working is 0.7 and 1.8 percentage points higher for the transitional WAO and WIA group, respectively, compared to the WAO group. Borghans et al. (2014) find that a 1993 Dutch DI reform with more stringent re-examination rules increased the fraction employed by 2.9 percentage points. Though their findings were under a more lenient DI regime where there was larger room for changes and they focus on individuals younger than age 45, still the effect of WIA on labor participation is not particularly large compared with the past Dutch DI reform. According to the back-of-envelope calculation in the Appendix 4.C, the increase in probability of working of WIA reform comes from a larger fraction of people working without having DI. But this

increase is partially offset by the fact that fewer people can work while claiming partial DI benefit.

Individuals who could not access the disability benefit might have turned to benefits from other benefit programs. We find evidence that the reforms induced sick individuals to turn to the unemployment benefit. Compared to the control group, both treatment groups increase their UB receipt, and the magnitude of the effects are slightly smaller than those of labor participation. Both treatment groups, however, become less likely to claim general assistance. It might be that the income earned above the subsistence minimum due to work resumption or benefit substitution limits the access to GA. Another possibility is that there is less need for claiming GA on top of the DI because the DI itself can more easily be topped up to the social minimum income.⁸⁴ Similar reasons can explain the decrease in receiving other benefits for the WIA group.

Earlier studies in other countries also find evidence that tightening the eligibility criteria leads to more take-up of other benefits (Karlström et al., 2008; Staubli, 2011). Studies on earlier disability insurance reforms in the Netherlands, however, show mixed results. Borghans et al. (2014) find that the disability reform in 1993, which is similar to the transitional WAO reform, led to more benefit claims from other benefit programs, while for reforms on the employer side, Koning and van Vuuren (2010) and De Jong et al. (2011) find that the experience rating reform in 1998 and the “Gatekeeper protocol” reform in 2002 had no spillover effects on UB.

In terms of the amount of benefits and salary, compared to the WAO regime, under the WIA regime individuals received 50.8 euros less disability benefits, while they earned 55.7 euros more salary, and received 22.8 euros more unemployment benefits, on average. This shows that, due to the WIA reform, the decrease in disability benefits received is compensated by higher wages and higher unemployment

⁸⁴ If the benefit received from a benefit scheme (sickness, disability, or unemployment scheme), or the wage earned during the second year of sickness (in the WIA), is lower than the applicable social minimum, it is supplemented up to the social minimum according to the Supplementary Benefits Act (Toeslagenwet). The total of the benefit and the social minimum supplement cannot exceed the former wage. If the individual is living with a partner, the supplement is granted if the total income of the individual and the partner is below the social minimum. If the individual is living alone, the amount of the supplement depends on whether the individual has children.

benefits so that, on balance, individuals earn a higher income (22.8 euros), on average.

A back-of-envelope calculation shows that the main contributor to the effect of WIA reform on benefit and salary amount is at the extensive margin. According to Table 4.2, the amount of DI conditional on having a DI is $253.895/0.198 = 1282.3$ and $212.504/0.141 = 1507.1$ for WAO and WIA, respectively. Conditional on having a DI, the average amount of DI benefit is higher for WIA than WAO, which is consistent with the fact that people have to be more severely disabled to get WIA benefit. The reduction in DI benefit mainly comes from the reduction of receipt (extensive margin). For the salary increase of 55.7 euros, the majority, roughly $0.018 \times (1376.214/0.534) = 46.4$ euros, comes from the increase in WIA's labor participation.⁸⁵ For the increase of UB, $0.014 \times (86.192/0.064) = 18.9$ euros out of 22.8 euros can be attributed to the increase at the extensive margin. Similar calculations can be done for transitional WAO group and the effects are also mainly from the extensive margin.

WIA reform reduces 4.9 euros of GA, and has almost no effect on other benefits. While transitional WAO reform reduces 2.8 euros of GA and increases 2.3 euros of OB. But the magnitude of effects on GA and OB is very small compared to programs like UB and DI.

4.7 Effect over time

As found in the literature, effects of DI reform can rebound or decrease over time (e.g. Staubli, 2011; Borghans et al., 2014). To study how the effects of transitional WAO and WIA reforms change over time, we consider the following regression:

$$y_{it} = \sum_{l=1}^{10} \gamma_{1l} (\text{Treat}_i^{\text{Trans. WAO}} \times d_{lt}) + \sum_{l=1}^{10} \gamma_{2l} (\text{Treat}_i^{\text{WIA}} \times d_{lt}) + \sum_{l=1}^{10} \delta_l d_{lt} + \lambda_{it} + \alpha_i + \varepsilon_{it} \quad (4.2)$$

Compared to Equation (4.1), the Post_t dummy, which indicates the entire “after” period, is replaced by 10 event year dummies, $d_{1t}, d_{2t}, \dots, d_{10t}$, indicating the 1st, 2nd, ..., 10th year since the “after” period begins. For example, d_{2t} indicates the second year from the time the individual becomes eligible for disability benefit.

⁸⁵ $1376.214/0.534$ gives the average salary of WAO in the “after” period conditional on working. Numbers are from Table 4.2.

Similar to equation (4.1), the “before” period is chosen as the base period for comparison. The interaction terms of treatment and event year dummies capture the mean difference in the outcome variable between the treatment and control groups in a given event year compared to the mean difference between the two groups in the pre-sickness period. The coefficients of the interaction terms γ_{1l} and γ_{2l} are effects of transitional WAO and WIA reforms over 10 event years, respectively. In Figures 4.3a and 4.3b, we present these coefficient estimates of treatment and event year dummies interactions over the “after” period of ten event years based on the regression given by Equation (4.2).

As a result of the re-examination during 2004 to 2008 discussed in section 4.3, the impact of the transitional WAO reform on labor participation and benefit use is short-lived, with effects all disappearing in 4 years since eligible for DI.

The WIA reform shows a substantial and persistent reduction in the receipt and amount of DI.

The magnitude of the effect is the largest in the first year and shrinks to half in four years, also related to the re-examination between 2004 and 2008. The effect remains stable afterwards. The effects of WIA on labor participation and salary are significant and persistent over time, while the effects on UB, GA, and OB are larger at the beginning but become smaller and eventually statistically indistinguishable from zero in about seven years. This can be a consequence of the unemployment benefit being temporary. The effects on GA and OB, though statistically significant, have rather limited economic significance compared to the magnitude of effects on DI and UB.

4.8 Heterogeneous effects

Disability reforms can affect sick individuals with different background and labor market characteristics differently.

Gender

Previous literature finds gender difference in effects of DI reform. For example, women experienced a larger increase in the probability to start working and to participate in other social benefit programs than men in the 1993 Dutch reform that tightens the entrance criteria to DI scheme (Borghans et al., 2014). We check if

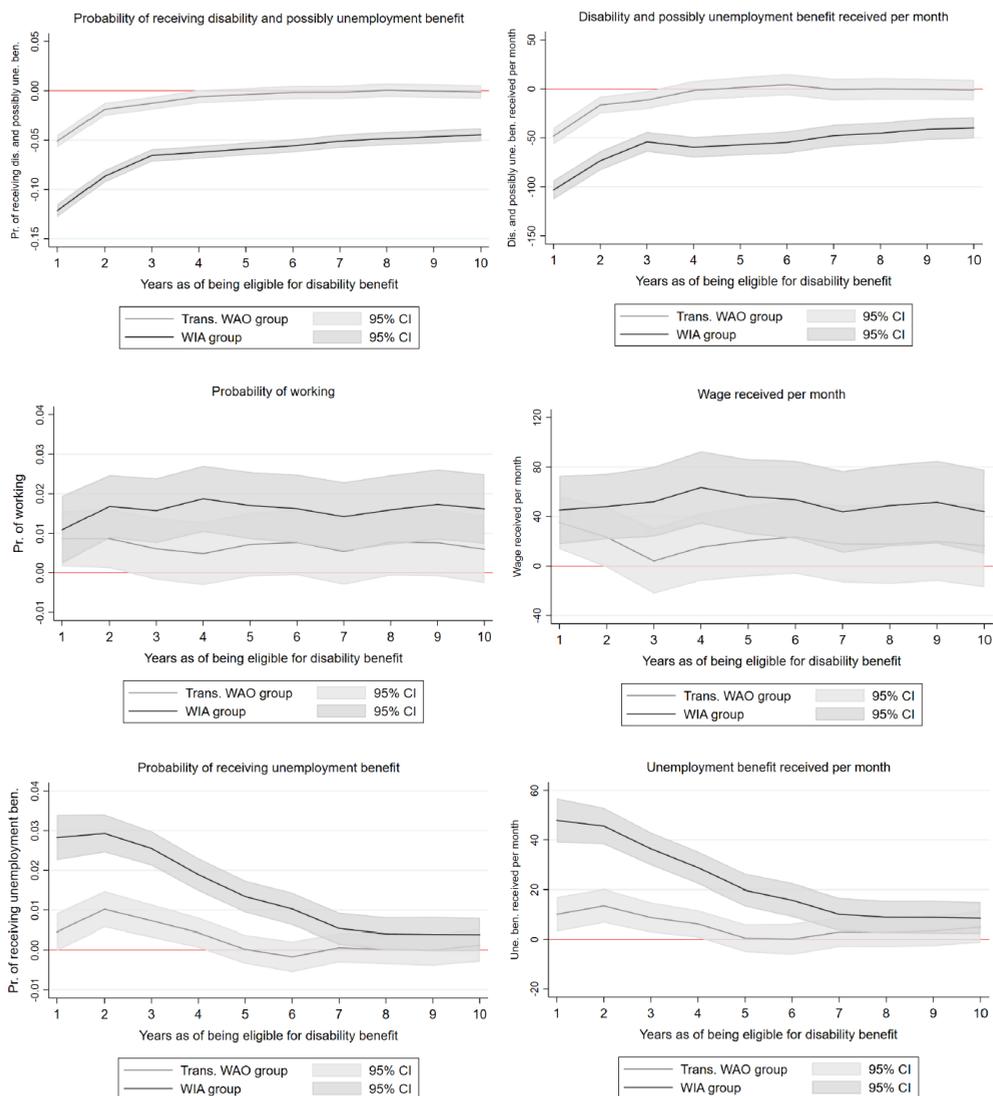


Figure 4.3a: Effects of reforms over time on labor participation and benefit receipt

Note: The plots present coefficient estimates of treatment and annual dummy interactions for control and treatment during the “after” period from regressions of equation (4.2) for labor participation, benefit receipt (disability and unemployment benefit), labor income and benefit income. Around each estimate is a 95 percent confidence interval. Regressions are based on the data available on a monthly basis for the reform period of ten years. The “before” period is the base for comparison for annual dummies. Labor participation, benefit receipt, labor income or benefit income is the outcome, and annual dummies, treatment and annual dummy interactions, calendar months dummies and time-invariant individual fixed effects are controls. Each regression uses 14,626,356 observations for 77,769 individuals who fell sick during the period from July 2003 until March 2004. Standard errors are adjusted for clustering at the individual level.

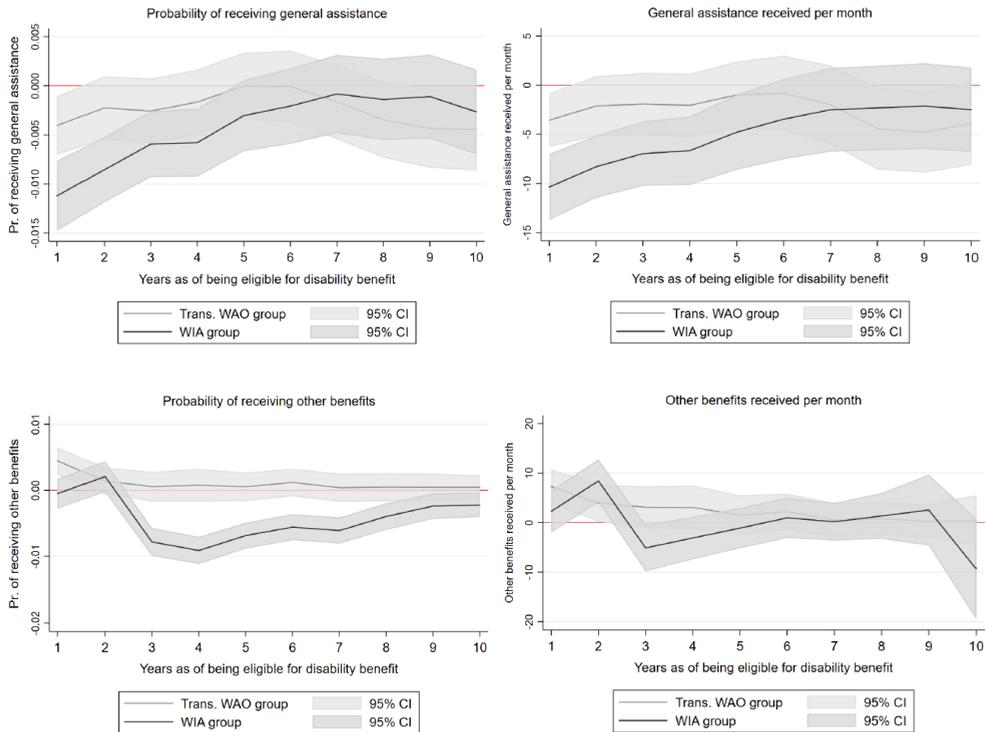


Figure 4.3b: Effects of reforms over time on salary and the amount of benefit

Note: The plots present coefficient estimates of treatment and annual dummy interactions for control and treatment during the “after” period from regressions of equation (4.2) for benefit receipt (general assistance and other benefits) and benefit income. Around each estimate is a 95 percent confidence interval. Regressions are based on the data available on a monthly basis for the reform period of ten years. The “before” period is the base for comparison for annual dummies. Benefit receipt or benefit income is the outcome, and annual dummies, treatment and annual dummy interactions, calendar months dummies, and time-invariant individual fixed effects are controls. Each regression uses 14,626,356 observations for 77,769 individuals who fell sick during the period from July 2003 until March 2004. Standard errors are adjusted for clustering at the individual level.

transitional WAO and WIA reforms also display such gender difference. Table 4.4 splits out the main results in Table 4.3 by gender.

For transitional WAO reform we do not see a significant difference between men and women. For WIA reform, the gender difference operates in the opposite way compared to the 1993 reform. DI reduction and UB increase are larger among men than women. This result is expected. Men earn, on average, higher wages than women. In the Dutch disability scheme, a higher pre-disability wage means a higher probability of getting a partial disability benefit. This is because the pre-disability wage is compared to a (lower) fictitious new wage that can be earned given one’s

disability. A higher pre-disability wage implies more alternative work opportunities with lower fictitious new wages (Section 4.2).

Table 4.4: Main results by gender

Age of falling sick Coefficient estimates	Female				Male			
	Trans. Post	WAO ×	WIA × Post		Trans. Post	WAO ×	WIA × Post	
DI receipt	-0.006 (0.004)		-0.049*** (0.004)		-0.009** (0.004)		-0.066*** (0.004)	
Labor Participation	0.009* (0.005)		0.019*** (0.005)		0.006 (0.005)		0.017*** (0.005)	
UB receipt	0.004* (0.002)		0.013*** (0.002)		0.002 (0.002)		0.015*** (0.002)	
GA receipt	-0.002 (0.003)		-0.005** (0.003)		-0.004** (0.002)		-0.003* (0.002)	
OB receipt	0.001 (0.001)		-0.003*** (0.001)		0.001 (0.001)		-0.007*** (0.001)	
DI amount	-9.903* (5.551)		* (5.720)		-2.629 (6.288)		* (6.454)	
Salary	21.049 (14.452)		45.298*** (15.289)		20.141 (19.657)		64.482*** (20.395)	
UB amount	5.544** (2.277)		15.630*** (2.358)		5.125* (2.844)		28.468*** (2.953)	
GA amount	-1.970 (2.637)		-6.059** (2.738)		-3.536** (1.662)		-3.996** (1.739)	
OB amount	1.071 (1.333)		-1.547 (1.312)		3.138 (1.918)		1.642 (1.939)	
Individuals	6,429,372				8,196,984			
Observations	34,186				43,583			

Notes: All regressions employ the linear regression model given by Eq. (4.1). Standard errors, in parentheses, are adjusted for clustering at the individual level. ***, **, * indicate statistical significance at the 0.01, 0.05, 0.10 levels, respectively. P-values based on the Wald statistic for the equality of the estimated reform effects for men and women are as follows. For the transitional WAO reform, the p-values are 0.596, 0.671, 0.480, 0.579, 1.000, 0.386, 0.970, 0.908, 0.615, and 0.376 for the ten outcome variables. For the WIA reform, the p-values are 0.003, 0.777, 0.480, 0.579, 0.005, 0.088, 0.452, 0.001, 0.525, and 0.173.

Age of falling sick

Age can also play a role. The elderly workers may have more difficulties in recovering from diseases and resuming working. Table 4.5 shows results by three age groups when individuals fell sick: individuals who are younger than 30, between 30 and 50, and older than 50.

Both reforms reduce DI more for older age group than other groups. This might reflect a composition effect. Older individuals more often hold a regular work contract (68.1 and 39.6 percent of the individuals in the oldest and youngest age

groups, respectively) which might make it more difficult to access the disability scheme for different reasons. For example, it might be easier to find potential jobs that these individuals can still perform, leading to a smaller disability grade for them (Section 4.2). However, this explanation is not supported by the results on labor participation and unemployment benefit receipt. Older individuals appear to be less able to compensate the decrease in disability benefit receipt or income with higher labor participation or wages, whereas they more often rely on unemployment benefit. Another possibility is that with the WIA reform employers have become reluctant to allow access to disability insurance on the basis of the age of their employees. The cost of a sick worker for the employer has increased since the introduction of experience rating in 1998 which punishes firms with many employees on disability benefits by charging them with a higher premium. Since older individuals are more likely to fall sick, cost of a sick worker is higher if the share of older workers is higher in the firm. Since the WIA reform extended the experience rating period from 5 to 10 years for partially disabled individuals (Section 4.2), employers might have become selective in allowing their older workers to access disability benefits.

The effect of the WIA reform found for the younger age group (2.9 and 2.7 percentage points for the group younger than 30 and the group between 30 and 50, respectively) is similar to the effect of the 1993 Dutch DI reform among the individuals of a similar age group. Borghans et al. (2014) find that the 1993 reform increased the probability of working by 2.9 percentage points for workers younger than age 45. During the years before the 1993 reform was introduced, exceptionally large numbers of people were receiving benefits due to the very generous disability scheme. There was therefore much potential for strong labor supply responses to the disability reform. We could expect that the potential for labor supply responses have become smaller in the aftermath of the 1993 reform due to the stricter eligibility criteria brought by this reform. The similar effect sizes of the 1993 and WIA reforms, therefore, suggest that the effect of the WIA reform on labor participation for the younger cohort is at least comparable to the 1993 reform.

Work status

Table 4.6 distinguishes results between individuals who hold a regular contract, temporary contract or a contract through a temporary work agency, and who are unemployed. Sick individuals with different work status can face difference

incentives under the reform measures, e.g. a worker holding a permanent contract with an employer actively urging him back to work may show different behavioral response in work resumption from a saddened unemployed individual claiming sickness benefit from the government.

Table 4.5: Main results by age of falling sick

Age of falling sick	< 30		[30,50)		≥ 50	
	Trans. WAO × Post	WIA × Post	Trans. WAO × Post	WIA × Post	Trans. WAO × Post	WIA × Post
DI receipt	-0.000 (0.006)	-0.031*** (0.006)	-0.007* (0.004)	-0.064*** (0.004)	-0.019*** (0.006)	-0.073*** (0.006)
Labor Participation	0.013* (0.007)	0.029*** (0.008)	0.008* (0.004)	0.027*** (0.004)	0.017*** (0.007)	0.022*** (0.007)
UB receipt	-0.001 (0.002)	0.004** (0.002)	0.002 (0.002)	0.012*** (0.002)	0.008** (0.003)	0.028*** (0.003)
GA receipt	-0.005 (0.004)	-0.010** (0.004)	-0.002 (0.002)	-0.002 (0.002)	-0.001 (0.002)	-0.002 (0.002)
OB receipt	0.001 (0.001)	-0.001 (0.001)	0.001 (0.001)	-0.006*** (0.001)	0.001 (0.002)	-0.007*** (0.002)
DI amount	-3.668 (7.015)	-25.740*** (7.432)	-1.645 (5.828)	-53.891*** (5.980)	-23.271** (9.920)	* (9.976)
Salary	41.470* (22.215)	88.133*** (24.291)	33.887** (16.037)	100.694*** (16.480)	40.406 (27.142)	79.531*** (27.567)
UB amount	-0.236 (2.213)	6.450** (2.318)	5.110** (2.285)	17.669*** (2.354)	11.345** (5.433)	47.564*** (5.545)
GA amount	-4.872 (3.953)	-10.522** (4.145)	-2.136 (2.054)	-2.293 (2.163)	-0.862 (2.035)	-2.193 (2.092)
OB amount	2.114* (1.239)	-0.204 (1.209)	1.767 (1.350)	-0.020 (1.289)	2.747 (3.890)	-0.214 (3.949)
Individuals	2,830,800		8,217,612		3,577,944	
Observations	15,031		43,695		19,043	

Notes: All regressions employ the linear regression model given by Eq. (4.1). Standard errors, in parentheses, are adjusted for clustering at the individual level. ***, **, * indicate statistical significance at the 0.01, 0.05, 0.10 levels, respectively. P-values based on the Wald statistic for the equality of the estimated reform effects for individuals who fell sick between age 30 to 50 and those who fell sick before age 30 are as follows: for the transitional WAO reform, the p-values are 0.332, 0.535, 0.289, 0.502, 1.000, 0.824, 0.782, 0.093, 0.539, and 0.850 for the ten outcome variables. For the WIA reform, the p-values are 0.000, 0.823, 0.005, 0.074, 0.000, 0.003, 0.669, 0.001, 0.078, and 0.917. P-values based on the Wald statistic for the equality of the estimated reform effects for individuals who fell sick older than 50 and those who fell sick before age 30 are as follows: for the transitional WAO reform, the p-values are 0.025, 0.686, 0.013, 0.371, 1.000, 0.107, 0.976, 0.048, 0.367, and 0.877 for the ten outcome variables. For the WIA reform, the p-values are 0.000, 0.510, 0.000, 0.074, 0.007, 0.000, 0.815, 0.000, 0.073, and 0.998.

The effects of reforms show substantial heterogeneity with respect to work status. Both reforms are least effective for the unemployed individuals, with least increase in labor participation and most increase in the unemployment benefit.

Table 4.6: Main results by work status

Age of falling sick	Regular		Temp.		Unemployed	
	Trans. WAO × Post	WIA × Post	Trans. WAO × Post	WIA × Post	Trans. WAO × Post	WIA × Post
DI receipt	-0.009** (0.004)	-0.058*** (0.003)	-0.016** (0.008)	-0.055*** (0.008)	-0.004 (0.008)	-0.068*** (0.008)
Labor Participation	0.008* (0.004)	0.032*** (0.005)	0.019** (0.008)	0.026*** (0.009)	-0.000 (0.008)	-0.027*** (0.008)
UB receipt	0.001 (0.001)	0.005*** (0.001)	0.007** (0.003)	0.023*** (0.003)	0.007* (0.004)	0.043*** (0.004)
GA receipt	-0.001 (0.001)	-0.003* (0.001)	-0.009* (0.005)	-0.014*** (0.005)	-0.006 (0.004)	-0.003 (0.005)
OB receipt	0.001 (0.001)	-0.006*** (0.001)	-0.002 (0.002)	0.002 (0.002)	0.003 (0.002)	-0.009*** (0.002)
DI amount	-4.681 (5.231)	-55.753*** (5.292)	-12.497 (10.595)	-39.572*** (11.544)	-9.387 (11.194)	-51.530*** (11.454)
Salary	30.114 (18.607)	125.315*** (19.240)	45.490* (25.687)	61.299** (28.120)	-9.080 (23.914)	-119.296*** (23.816)
UB amount	2.518 (2.007)	8.604*** (2.140)	10.605*** (3.937)	32.785*** (4.301)	18.325*** (6.169)	70.116*** (6.107)
GA amount	-1.532 (1.270)	-2.974** (1.345)	-9.038* (4.723)	-15.257*** (5.062)	-5.472 (4.468)	-3.515 (4.631)
OB amount	3.493** (1.753)	-3.611** (1.733)	-0.336 (2.254)	6.693** (2.818)	2.592 (2.677)	6.562*** (2.450)
Individuals	8,518,944		2,534,856		2,945,844	
Observations	45,312		13,445		15,682	

Notes: All regressions use model given by Eq. (4.1). Standard errors in parentheses are adjusted for clustering at the individual level. ***, **, * indicate statistical significance at the 0.01, 0.05, 0.10 levels, respectively. P-values based on the Wald statistic for the equality of the estimated reform effects for individuals hold a temporary work contract and those who hold a regular work contract are as follows: for the transitional WAO reform, the p-values are 0.434, 0.219, 0.058, 0.117, 0.180, 0.508, 0.628, 0.067, 0.125, and 0.180 for the ten outcome variables. For the WIA reform, the p-values are 0.725, 0.560, 0.000, 0.031, 0.000, 0.203, 0.060, 0.000, 0.019, and 0.002. P-values based on the Wald statistic for the equality of the estimated reform effects for individuals who are unemployed and those who hold a regular work contract are as follows: for the transitional WAO reform, the p-values are 0.576, 0.371, 0.146, 0.225, 0.371, 0.703, 0.196, 0.015, 0.396, and 0.778 for the ten outcome variables. For the WIA reform, the p-values are 0.242, 0.000, 0.000, 1.000, 0.180, 0.738, 0.000, 0.000, 0.911, and 0.001.

In particular, we find the following two notable results: first, the WIA reform decreases use of disability benefit among the unemployed individuals more than it does among those who hold a temporary or regular contract. It might be that the unemployed are more often sick, without necessarily having severe health problems, and therefore are more likely to be affected by a stricter disability benefit regime. Furthermore, the reform decreases labor participation and salary among the unemployed but increases them among those with regular or temporary work contracts. WIA reform brings a much larger increase in the receipt and amount of UB for the unemployed individuals than those with regular and temporary contract.

It might be that the work resumption incentives brought by the WIA reform lead employers to reintegrate their employees back to their job while they prove ineffective if there is no employer (Koning and Lindeboom, 2015). On the employees' side, the incentives to resume working may be smaller for the unemployed individuals insured under the WIA, compared to those insured under the WAO, since they spend an additional year in the sickness scheme before they become eligible for disability benefits. A longer unemployment spell may lead to more human capital loss or have a stronger "scarring effect" for the sick unemployed individuals, and decrease their prospects of finding a job (Arulampalam, 2001; Arulampalam et al., 2001). These results suggest that the reforms risk making the individuals who are unemployed particularly vulnerable to become disabled since they appear to have limited access to the disability scheme and lack the incentives other groups have to resume working.

A second notable result follows from comparing the results for workers with regular and temporary contracts. Due to the WIA reform, individuals with regular contracts show a slightly higher increase in labor participation compared to those with temporary contracts. This is expected because these individuals have an employer to return to after recovering from sickness within two years. However, these individuals are far better in coping with the loss of disability benefit as they increase their salary by a much larger amount and rely less on unemployment benefit. Possibly for the same reason, resuming work with their own employer gives them more flexibility in increasing their number of work hours to the former level compared to the individuals with temporary contracts who usually have to resume working with a new employer.

4.9 Sensitivity analyses

"Randomly" falling sick and the seasonality

One of the main assumptions for our empirical strategy is that sick individuals cannot manipulate the date of falling sick to select themselves into the control or treatment groups.

For transitional WAO this assumption is not likely to violate, because the stricter criteria to become eligible for disability benefit after a sickness period of one year

was announced much later, on 12 March 2004. This means that sickness reporting in reaction or anticipation to the transitional reform is impossible.

For WIA reform, the government only presented a general policy program outlining, among other targets, its plan to reform the WAO scheme on 15 September 2003. In particular, it is announced that the sickness period will be extended from one to two years, and a stricter disability insurance law will be introduced for the individuals who fall sick from 1 January 2004. The details of the law were announced much later on 18 August 2004. This means that, in Figure 4.4, which presents the distribution of individuals by the month of falling sick, individuals who fall sick in the peak month of September are unlikely to have selected themselves into the lenient WAO scheme since they have little time to react to the planned reform announced in mid September.

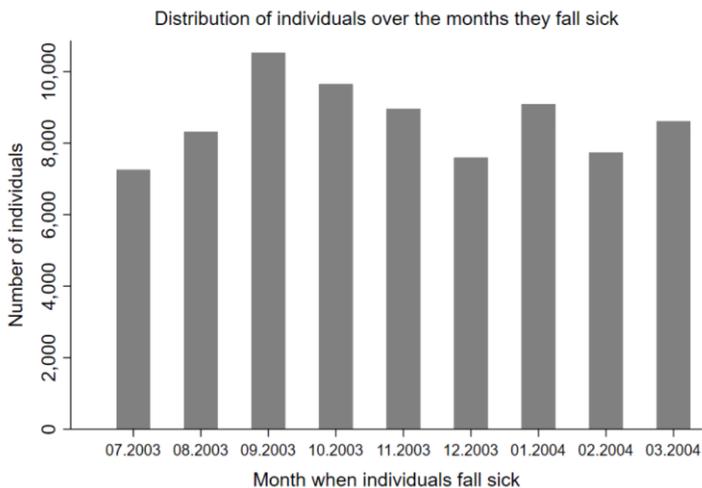


Figure 4.4: Distribution of individuals over the months they fall sick

If the individuals select themselves into transitional WAO to avoid a much stricter WIA regime, we would expect a peak in December right before the WIA reform commencing in January. However, we do not see such pattern in Figure 4.4. The number of sickness cases slight drop in December which rather reflects the seasonality of sickness cases.⁸⁶

⁸⁶ Van Sonsbeek and Gradus (2013) show that the DI inflow usually peaks in the first quarter and reaches the trough in the third quarter. The pattern of DI inflow reflects a lagged pattern of sickness cases.

We further do a “donut-hole” regression, dropping individuals falling sick in September and December, for two purposes: the first purpose is to check the strategic behavior around the commencing dates of the two reforms. For those who strategically select themselves into a more lenient system, we would expect these people are more likely to utilize the DI system. By dropping individuals falling sick right before the two reforms (i.e., September and December, we check if dropping these people would change the estimation results. Secondly, by removing individuals in the peak and trough months, we also check if the results are robust to the seasonality of the sickness cases. The Panel (1) of Table 4.7 shows that the results are rather robust to dropping individuals falling sick in special months.

Common trend assumption

Another main assumption of our identification strategy is that the potential outcomes of control and treatment groups (in absence of reforms) share common trends. We cannot directly test this assumption. But we can show evidence for parallel trends in the pre-sickness period. As mentioned in Section 4.4, Figures 4.1a and Figure 4.1b show that control and treatment groups share very similar time trends until individuals fall sick in most cases. But the lines of WAO group is slightly lower than other groups for the receipt and amount of general assistance between January 2002 and June 2003. And the lines of WAO group’s labor participation and the amount of unemployment benefit are slightly higher than other groups between January 2002 and January 2003, and between January and June 2003, respectively.

Following Abrammitzky and Lavy (2014) we use regression analysis to show the evidence and potential violation of the common pre-trend across groups. In particular, we use pre-sickness data from January 1999 to June 2003 to estimate a regression where labor participation, benefit receipt, labor income, or benefit income is the outcome, and calendar month dummies, interactions of treatment and calendar month dummies, and time-invariant individual fixed effects are controls. January 1999 is chosen as the base month for comparison. Coefficient estimates of interaction terms that are not statistically different from zero provide evidence in favour of the common trend assumption.

Figures 4.5a and 4.5b plot the coefficient estimates of the treatment and calendar month dummy interactions from regressions where labor participation, benefit receipt, labor income, and benefit income are the outcomes.⁸⁷

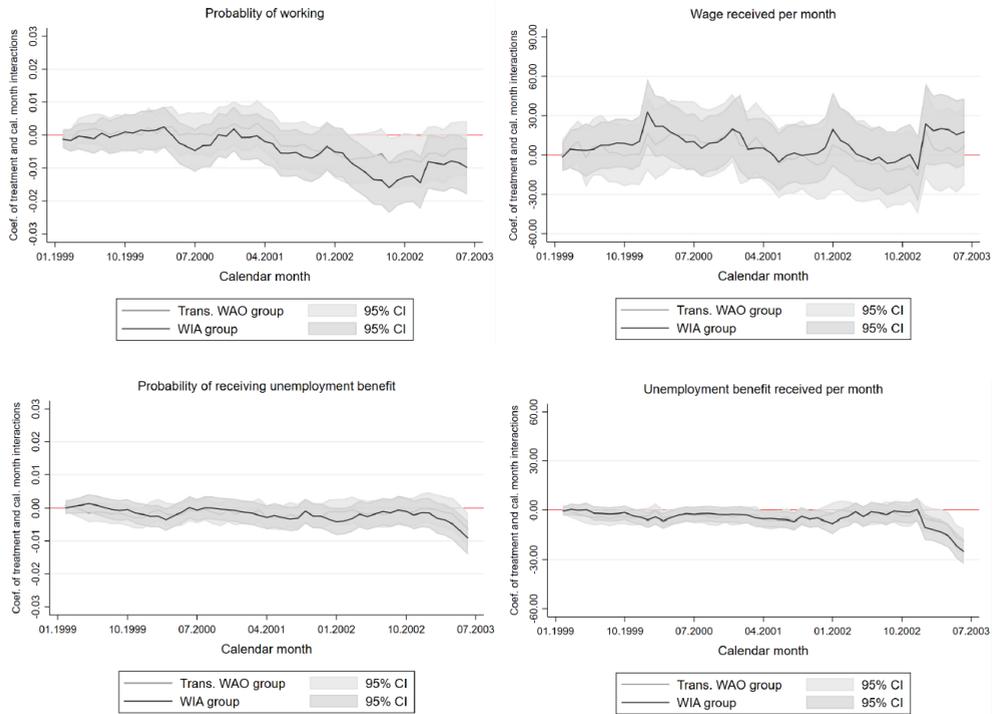


Figure 4.5a: Coefficient estimates of treatment groups and calendar month interactions in the pre-sickness period

Note: Around each estimate is a 95 percent confidence interval. Regressions are based on pre-sickness data from January 1999 to June 2003. January 1999 is the base month for comparison. Labor participation, salary, the receipt or the amount of unemployment benefit is the outcome, and calendar month dummies, treatment and calendar month dummy interactions, and time-invariant individual fixed effects are controls. Each regression uses 4,199,526 observations for 77,769 individuals who fell sick during the period from July 2003 until March 2004. Standard errors are adjusted for clustering at the individual level.

⁸⁷ We do not plot for the receipt and amount of DI because the outcomes equal 0 in the pre-sickness period.

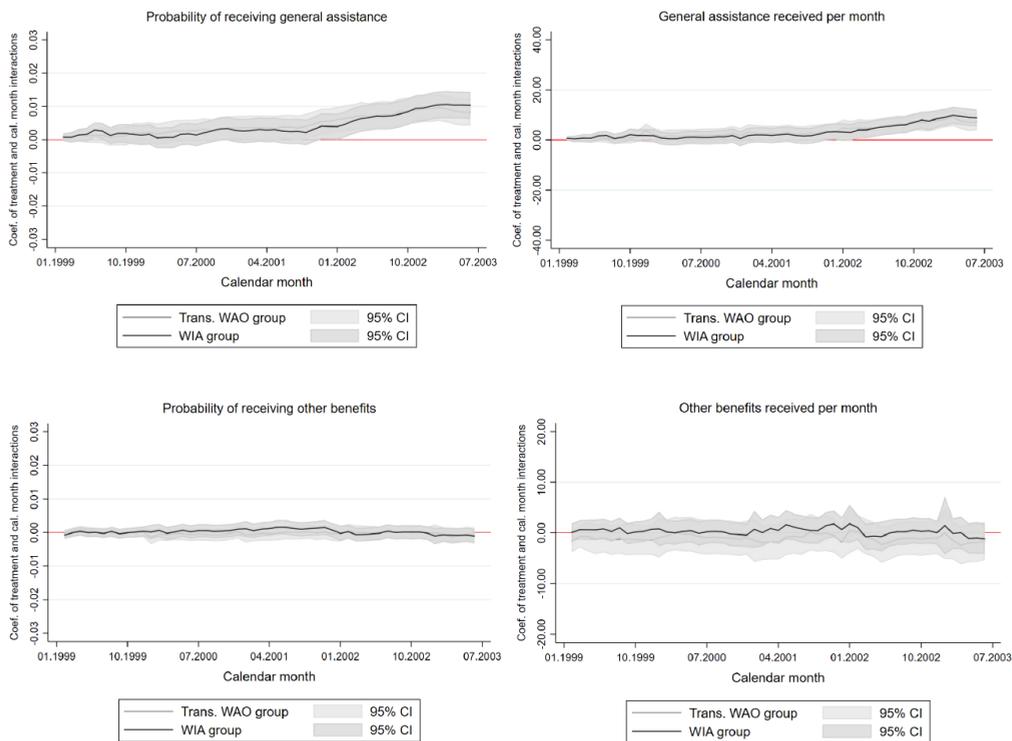


Figure 4.5b: Coefficient estimates of treatment groups and calendar month interactions in the pre-sickness period

Note: Around each estimate is a 95 percent confidence interval. Regressions are based on pre-sickness data from January 1999 to June 2003. January 1999 is the base month for comparison. The receipt or the amount of general assistance or other benefits is the outcome, and calendar month dummies, treatment and calendar month dummy interactions, and time-invariant individual fixed effects are controls. Each regression uses 4,199,526 observations for 77,769 individuals who fell sick during the period from July 2003 until March 2004. Standard errors are adjusted for clustering at the individual level.

For both treatment groups, the coefficient estimates are not statistically different in most months for most outcomes. However, for the receipt and amount of general assistance, we do see a small differential trend from January 2002 to June 2003. For labor participation, coefficients deviate slightly from 0 between January 2002 and January 2003. And the curve of the amount of unemployment benefit also slightly tilts down since January 2003. This pattern is consistent with the deviating curves of the WAO group in Figure 4.1a and 4.1b.

To check how these differential trends matter, we do the following two sets of sensitivity checks: the first set of analyses checks how much the differential trends in

the pre-sickness periods matter to the estimation results. We only keep observations from January 1999 to December 2001 in the pre-sickness periods when we do not observe violations of common trend. We run the same regressions as in the main analysis based on the sample with restricted “pre-sickness” periods. Results are summarized in Panel (2) of Table 4.7. They are very close to the main results in Table 4.3, indicating that results are robust to whether we include the periods with differential trends or not. Another check is to explicitly control for differential calendar time trends between the “problematic period”, that is, we add treatment group indicators interacting with calendar month dummies from January 2002 to June 2003 as extra controls to Equation (4.1). Panel (3) of Table 4.7 shows the estimation results. They are also very close to the main results in Table 4.3, suggesting that the existence of the occasional differential trends does not influence the estimation results.

Another concern is that the deviation from common trend at the end of the pre-sickness period may influence the potential outcomes after falling sick. To address this concern, we first check if the violation of common trend occurs randomly or systematically across outcomes when using different subgroups of WAO as control groups. Here subgroups refers to people falling sick in July, August, and September of 2003. Like what we do for Figure 4.5a and 4.5b, we keep data from January 1999 to June 2003, and regress each outcome on individual fixed effects, calendar month dummies, and calendar month dummies interacting with indicators for transitional WAO (T1) and WIA (T2), respectively. Each set of regressions are done on three samples where we use the July, August, and September group as the control group, respectively. We test for the null hypotheses that (1) the coefficients of $\text{pre-months} \times T1$ all equal 0; (2) the coefficients of $\text{pre-months} \times T2$ all equal 0. The F statistics and the p-values are summarized in Table 4.8. If the null hypothesis is rejected at the significance level of 0.05, we consider it as an indication of potential violation of the common pre-trend. The shaded grids indicate samples where no violation of common pre-trend is detected.

We find that the violation does not occur systematically across outcome variables and across different subgroups of WAO as the control group.

Next, we pick the “safe control group” where we do not detect a differential pre-trend for each outcome, i.e., the shaded groups in Table 4.8. For the receipt and

amount of DI whose pre-trends are always 0, any subgroup can be considered as the “safe group”. We just randomly pick August group as the control group for these two outcome variables. Based on the selected control group, we estimate the same model in Equation (1) and show results in Panel (1) of Table 4.9. The results are very similar to the main results in Table 4.3. Estimations with subsamples that have parallel trend between control and treatment groups give essentially the same results, suggesting that results are robust to the existence of the slight differential trends in some of the outcome variables.

Table 4.7: Sensitivity analysis on dropping special months and pre-sickness trend

Coefficient estimates	(1) Donut-hole		(2) Restricted "before" period		(3) Control for differential trend	
	Trans. WAO × Post	WIA × Post	Trans. WAO × Post	WIA × Post	Trans. WAO × Post	WIA × Post
DI receipt	-0.012*** (0.004)	-0.060*** (0.003)	-0.008** (0.003)	-0.059*** (0.003)	-0.008** (0.003)	-0.058*** (0.003)
Labor Participation	0.014*** (0.004)	0.025*** (0.004)	0.005 (0.004)	0.015*** (0.004)	0.005 (0.004)	0.015*** (0.004)
UB receipt	0.005*** (0.002)	0.015*** (0.002)	0.003** (0.001)	0.013*** (0.001)	0.002* (0.001)	0.014*** (0.001)
GA receipt	-0.001 (0.002)	-0.004** (0.002)	-0.001 (0.002)	-0.002 (0.002)	-0.001 (0.002)	-0.002 (0.002)
OB receipt	0.001 (0.001)	-0.005*** (0.001)	0.001 (0.001)	-0.005*** (0.001)	0.001 (0.001)	-0.005*** (0.001)
DI amount	-12.053** (5.318)	-52.319*** (5.106)	-5.618 (4.300)	-50.829*** (4.414)	-5.49 (4.301)	-50.803*** (4.413)
Salary	33.934** (16.124)	73.734*** (15.64)	16.998 (13.296)	55.893*** (13.836)	17.05 (13.153)	54.915*** (13.694)
UB amount	9.193*** (2.347)	24.271*** (2.247)	5.068*** (1.827)	20.480*** (1.894)	4.932*** (1.825)	22.160*** (1.895)
GA amount	-2.373 (1.835)	-5.053*** (1.779)	-1.236 (1.608)	-3.024* (1.671)	-1.400 (1.580)	-3.315** (1.643)
OB amount	1.111 (1.482)	0.221 (1.415)	2.194* (1.257)	-0.000 (1.239)	2.167* (1.249)	0.157 (1.235)
Individuals	11,145,396		12,916,711		14,626,356	
Observations	59,639		77,769		77,769	

Notes: All regressions employ the linear regression model given by Eq. (4.1) except for Panel (3). Standard errors, in parentheses, are adjusted for clustering at the individual level. ***, **, * indicate statistical significance at the 0.01, 0.05, 0.10 levels, respectively. Panel (1) is based on a sample dropping individuals falling sick in September and December 2003. Panel (2) is based on a sample with restricted pre-sickness period, i.e., in the pre-sickness periods, only observations in January 1999 to December 2001 are included. Panel (3) controls for differential calendar time trends during the period from 01.2002 to 06.2003, that is, we add treatment group indicators interacting with calendar month dummies from 01.2002 to 06.2003 as extra control to Equation (4.1).

Table 4.8: Tests for common pre-trend with subgroups of the WAO as alternative control groups

Control	July group		August group		September group	
	F stat.	p val.	F stat.	p val.	F stat.	p val.
Labor Participation						
Pre-months×Trans. WAO	1.28*	0.0810	1.14	0.2213	1.07	0.3416
Pre-months×WIA	1.41**	0.0261	1.41**	0.0267	1.32*	0.0597
UB receipt						
Pre-months×Trans. WAO	1.16	0.2027	1.04	0.3918	1.05	0.3715
Pre-months×WIA	1.29*	0.0749	1.27*	0.0911	0.74	0.9236
GA receipt						
Pre-months×Trans. WAO	1.08	0.3155	0.94	0.5947	0.88	0.7107
Pre-months×WIA	1.11	0.2668	1.32*	0.0571	1.33*	0.0527
OB receipt						
Pre-months×Trans. WAO	0.99	0.5022	1.29*	0.0776	0.73	0.9320
Pre-months×WIA	1.14	0.2215	1.38**	0.0349	0.95	0.5717
Salary						
Pre-months×Trans. WAO	1.1	0.2932	1.28*	0.0822	1.27*	0.0914
Pre-months×WIA	1.11	0.2743	1.26*	0.0968	1.19	0.1652
UB amount						
Pre-months×Trans. WAO	2.02***	0.0000	1.36**	0.0414	0.94	0.6048
Pre-months×WIA	2.16***	0.0000	1.64***	0.0023	0.91	0.6533
GA amount						
Pre-months×Trans. WAO	1.24	0.1151	1.60***	0.0037	1.03	0.4172
Pre-months×WIA	1.21	0.1375	1.80***	0.0003	1.36**	0.0408
OB amount						
Pre-months×Trans. WAO	1.1	0.2949	1.06	0.3627	0.83	0.8041
Pre-months×WIA	1.25	0.1083	0.85	0.7683	1.06	0.3656

Note: ***, **, * indicate statistical significance at the 0.01, 0.05, 0.10 levels. Each outcome is regressed on calendar month dummies, calendar month dummies interacting with group indicators (Trans. WAO and WIA), and individual fixed effects on a sample between January 1999 to June 2003 where the control group is the group falling sick in July, August, and September respectively. We report the F stat and the p-value of the joint tests for the null hypotheses that (1) the coefficients of pre-months×T1 all equal 0 (2) the coefficients of pre-months×T2 all equal 0. Shaded grids are samples where no violation of common pre-trend is detected.

Alternative sample restrictions

As explained in Section 4.3 and Appendix 4.A, we restrict sample to sickness cases longer than 180 days to ensure the comparability of groups, because there is under-reporting of short cases especially in the WAO group.

However, restricting sample to relatively longer duration of 180 days may overlook the behavioral responses early in the waiting period, that is, people may select themselves into a shorter or longer stay in the sickness period, which will lead to an under- or overestimation of the effects of the reforms. For example, healthier individuals may think that two years of waiting period is too long and they will

Table 4.9: Regressions with control groups where no differential trend is detected and placebo test

Coefficient estimates	(1) Safe control group			(2) Placebo test		
	Trans. × Post	WAO WIA × Post	Chosen group control	Trans. × Post	WAO WIA × Post	WIA × Post
DI receipt	-0.008*** (0.004)	-0.060*** (0.004)	August	-0.005 (0.005)	0.002 (0.005)	
Labor Participation	-0.003 (0.005)	0.008* (0.005)	September	-0.000 (0.006)	-0.009 (0.006)	
UB receipt	0.001 (0.002)	0.012*** (0.002)	September	-0.002 (0.002)	-0.001 (0.002)	
GA receipt	-0.004 (0.002)	-0.005** (0.002)	July	0.002 (0.003)	0.000 (0.003)	
OB receipt	0.001 (0.001)	-0.005*** (0.001)	September	0.002 (0.001)	0.000 (0.001)	
DI amount	-2.443 (6.169)	-48.631*** (6.267)	August	-3.600 (7.332)	5.723 (7.415)	
Salary	73.793*** (20.574)	110.880*** (21.004)	July	11.242 (21.747)	-21.101 (21.608)	
UB amount	2.737 (2.448)	20.385*** (2.508)	September	-4.912 (3.272)	-1.774 (3.173)	
GA amount	-3.100 (2.270)	-5.119** (2.320)	July	1.757 (2.612)	1.265 (2.590)	
OB amount	2.527 (1.606)	0.564 (1.623)	September	1.521 (1.960)	-0.046 (2.083)	
	July as control	August as control	September as control			
Individuals	11,006,388	11,210,676	11,635,380	5,013,312		
Observations	58,915	59,979	62,191	26,111		

Notes: All regressions employ the linear regression model given by Eq. (4.1). Standard errors, in parentheses, are adjusted for clustering at the individual level. ***, **, * indicate statistical significance at the 0.01, 0.05, 0.10 levels, respectively. Panel (1) is based on a sample with a selected subsample of WAO as the control group. The column “chosen group” indicates which subsample is chosen as the control group. July, August, September refers to the subgroup falling sick in July, August, and September of 2003. Panel (2) is based on WAO group only. The WAO group is randomly divided into three “placebo groups”, “placebo control”, “placebo transitional WAO”, and “placebo WIA”. The number of individuals for three “placebo groups” are 8703, 8704, and 8704, respectively. We then estimate the same model as Equation (4.1) on the placebo sample.

recover before two years anyway. So they may decide to exit waiting period and return to worker earlier in WIA than in WAO. In this case, dropping sickness cases shorter than 180 days will drop more of such healthier guys in WIA than in WAO, which will lead to an underestimation of the effects of the WIA reform. By the same token, the effects of the reforms can be over-estimated if individuals under WIA are inclined to stay longer in sickness period, being paid without working.

Here we examine how much the results would be influenced by the under-reporting issue and the potential overlook of the behavioral responses early in the waiting

period (self-selection issue). We estimate regressions given by Equation (4.1) on three samples of individuals who have spent at least 90, 120, and 150 days in sickness.

Table 4.10: Main regressions with alternative sample restrictions

Sick at least for:	90 days		120 days		150 days	
	Trans. WAO × Post	WIA × Post	Trans. WAO × Post	WIA × Post	Trans. WAO × Post	WIA × Post
DI receipt	-0.021*** (0.002)	-0.052*** (0.002)	-0.014*** (0.002)	-0.050*** (0.002)	-0.010*** (0.003)	-0.053*** (0.003)
Labor Participation	0.017*** (0.003)	0.028*** (0.003)	0.011*** (0.003)	0.021*** (0.003)	0.008** (0.003)	0.018*** (0.003)
UB receipt	0.002* (0.001)	0.010*** (0.001)	0.002*** (0.001)	0.011*** (0.001)	0.003*** (0.001)	0.013*** (0.001)
GA receipt	-0.001 (0.001)	-0.003*** (0.001)	-0.002 (0.001)	-0.004*** (0.001)	-0.002* (0.001)	-0.004*** (0.001)
OB receipt	0.000 (0.001)	-0.004*** (0.001)	0.000 (0.001)	-0.004*** (0.001)	0.000 (0.001)	-0.004*** (0.001)
DI amount	-24.601*** (3.009)	-52.127*** (3.084)	-14.913*** (3.362)	46.730*** (3.462)	-8.820** (3.804)	46.764*** (3.914)
Salary	41.900*** (10.100)	77.722*** (10.528)	23.891** (10.895)	59.118*** (11.407)	16.38 (11.799)	53.001*** (12.339)
UB amount	3.820*** (1.447)	16.603*** (1.489)	4.767*** (1.574)	18.725*** (1.627)	6.152*** (1.721)	21.714*** (1.783)
GA amount	-1.167 (1.147)	-4.193*** (1.187)	-2.180* (1.245)	-4.766*** (1.297)	-2.847** (1.367)	-4.952*** (1.426)
OB amount	0.977 (0.883)	-0.567 (0.888)	1.612* (0.971)	0.095 (0.983)	2.009* (1.081)	0.145 (1.088)
Individuals	23,579,556		20,078,508		17,097,660	
Observations	125,535		106,807		90,918	

Notes: All regressions employ the linear regression model given by Eq. (4.1). Standard errors, in parentheses, are adjusted for clustering at the individual level. ***, **, * indicate statistical significance at the 0.01, 0.05, 0.10 levels, respectively. The three panels correspond to samples of sick individuals who are at least sick for 90, 120, and 150 days.

For the WIA reform, as we include shorter sickness cases into our sample, on one hand, we expect the estimated effects to increase or decrease due to the alleviation of the aforementioned self-selection issue. On the other hand, we expect more severe under-reporting issue of short-term sickness, which will lead to further overestimation of the effects. For the transitional WAO reform, we expect the effects to be further overestimated as we allow for shorter sickness cases into the sample. The transitional WAO group should be less influenced by the self-selection issue than WIA group since transitional WAO reform did not change the length of the sickness period. The under-reporting issue should dominate and blow up the estimates.

Estimation results shown in Tables 4.10 are rather consistent with our expectation. Compared with the baseline results in Table 4.3, results are qualitatively similar for the WIA reform in almost all outcomes. The only exceptions are the estimated effects on labor participation and wage in the sample of individuals spending at least 90 days in sickness. The effects are larger, possibly due to a dominating under-reporting issue. For transitional WAO reform which is supposed to be mainly influenced by the under-reporting issue, we indeed observe the increasingly overestimation of the effects on the receipt and amount of disability benefit and wage as we allow for shorter sickness cases. For other variables, the results are rather robust.

To sum up, these results suggest that the baseline estimates of the effects are to a large extent robust to the sampling of sick individuals with respect to the number of days they spend in the sickness scheme.

Placebo test and Mortality effect

We run a placebo test to check if there is an “effect” when there is no reform. We randomly divide WAO group into three “placebo groups”, “placebo control group” with 8703 individuals, “placebo transitional WAO group”, and “placebo WIA group”, both with 8704 individuals. We estimate the same model as in Table 4.3. Results are shown in Panel (2) of Table 4.9. As expected, no effect is found in any outcome variable in the placebo sample.

We also check the mortality rates across groups. Note that individuals do not exit the sample because of death. But checking mortality rates gives an idea of the extreme health effect of the reform, and also partially explain why some individuals are neither working nor claiming any benefits. We do not have direct information on mortality. But the missing values in the dataset of demographic information usually indicate death or moving abroad. This information can therefore be seen as the upper bound of the mortality. We found 150 individuals out of 26111 individuals (0.57%) in WAO group “died” within the sample period. For transitional WAO group and WIA, the number is 131 out of 26217 individuals (0.50%) and 125 out of 25441 (0.49%), respectively. The mortality rates are small and similar across groups.

4.10 Conclusion

In the last decades disability insurance programs have grown in many western countries. Governments implement social security reforms to reduce enrollment in the disability insurance program and the cost of disability benefits. We evaluated a major disability insurance reform in the Netherlands, WIA reform, that introduced a basket of changes, including strong incentives for work resumption, tightening entrance criteria, extending the waiting period, etc. Using unique administrative data on agents who fall sick before and after the reform, we analyzed the effect of the reform on individuals' labor participation decisions and use of benefits from alternative social security programs.

Difference-in-difference analyzes provide notable findings. WIA reform decreased the amount and receipt of disability benefit substantially, and its impact had been persistent during the ten years of the study period. It increased labor participation and the unemployment benefit receipt to a sizable extent. On average, individuals fully compensate lost disability benefits by the increase in the salary and the unemployment benefit.

The impact of the WIA reform is substantially heterogeneous with respect to work status and age. Individuals who are unemployed are much more limited in their access to the disability insurance scheme and lack the incentives the group of individuals with regular or temporary contracts have to resume working. WIA reform even decreases the labor participation and income for the unemployed group. This suggests that unemployed individuals, who are already a vulnerable labor market group, face additional limitations in their access to the labor market when they fall sick. Besides the unemployed, older individuals appear as a second vulnerable group. They are less able to compensate the decrease in disability benefit receipt or income with higher labor participation or wages, whereas they more often rely on unemployment benefit.

These heterogeneous effects raise inequality concerns and call for extra attention to the vulnerable labor market groups under such universal reform. For policy makers, special care should be given to the naturally "left-out" group. Like in the WIA reform, employers are supposed to play an important role in facilitating work resumption. Then the unemployed individuals are naturally "left out" by the policy design. Extra measures need to be considered to protect them from welfare loss.

In the future research, it would be interesting to investigate the effect of WIA reform on individuals' health and health care utilization. The effects of the reform could be heterogeneous depending on the household structure. The within-household or inter-generational spillover effect of the reform, and decomposing the effect of each policy measure from the overall policy effect, would all be interesting next steps.

Appendix 4.A The under-reporting of sickness cases shorter than 180 days

According to the “Gatekeeper Protocol”, sickness cases longer than 13 weeks must be reported to UWV. For sickness cases shorter than longer than 90 days, there can be under-reporting of cases. Figure 4.A.1 shows the distribution of the sickness duration by the month of falling sick for cases between 0 to 360 days. Indeed the distributions are not comparable for cases between 0 to 90 days. It seems that as of December 2003, the share of sickness between 0 to 90 days suddenly increases, suggesting that employers are suddenly more active in reporting short cases.

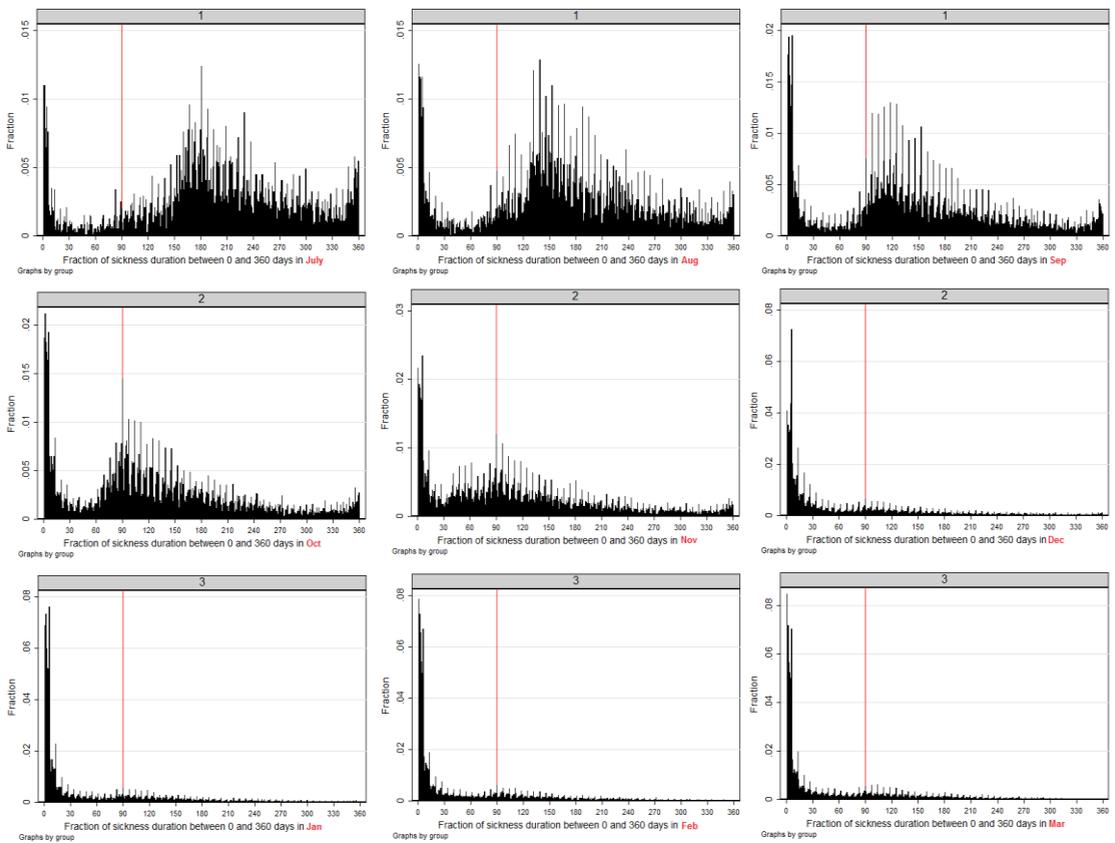


Figure 4.A.1: Distribution of the sickness duration by the month of falling sick for cases between 0 to 360 days

Note: group 1, 2, 3 refer to WAO, trans. WAO and WIA group, respectively.

For sickness cases longer than 90 days, we expect to see similar distributions of sickness duration across groups. Figure 4.A.2 shows the distribution of duration for cases between 90 to 360 days across groups. Groups 2 and 3 (transitional WAO and WIA groups) are indeed comparable. But group 1 (WAO group) has a slightly smaller share of shorter cases compared with groups 2 and 3.

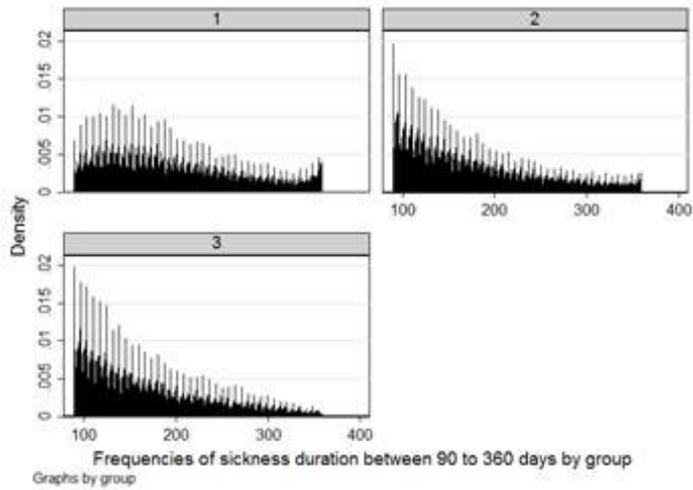


Figure 4.A.2: Distribution of the sickness duration across groups for cases between 90 to 360 days

Note: group 1, 2, 3 refer to WAO, trans. WAO and WIA group, respectively.

Figure 4.A.3 shows a zoomed-in plot of the distribution of the sickness duration by the month of falling sick for cases between 90 to 360 days. Distributions are comparable as of October. For cases falling sick in July, August, and September, we see a bulk of distribution “missing” at the left end (shorter than 180 days). The missing short cases are more obvious in July and August. We also noticed that the distributions are similar for cases longer than 180 days.

Table 4.A.1 shows the similar pattern that July and August may have “under-reporting” of cases between 90 to 180 days, while the frequencies for cases longer than 180 days are generally comparable across groups. (All the calculations in this Appendix 4.A are based on the original sickness data, where no sample restriction is imposed yet.)

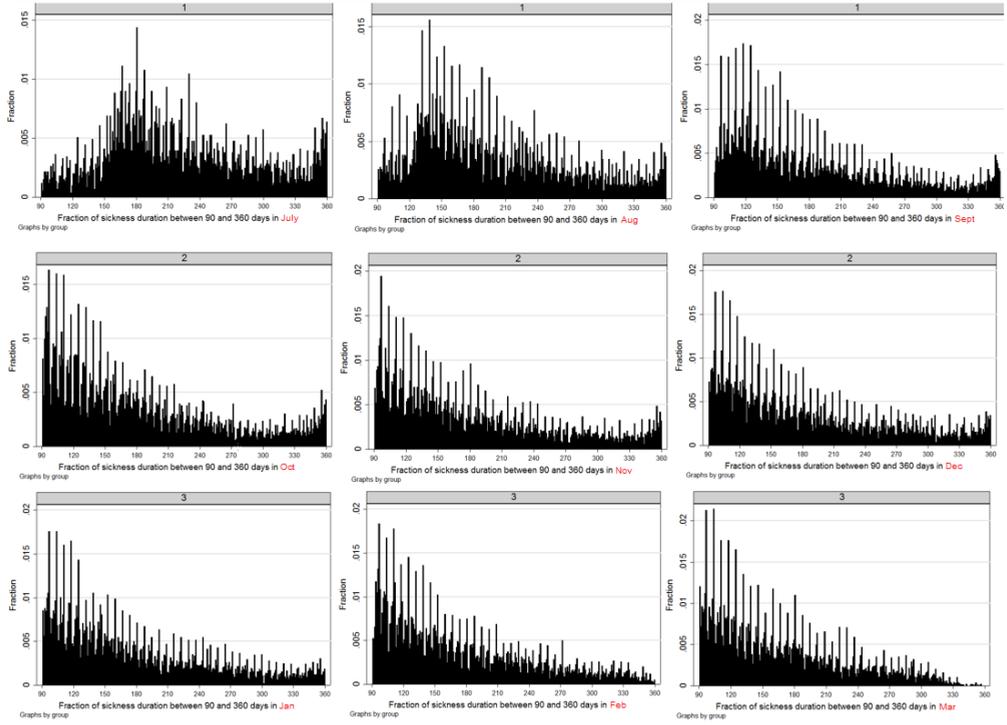


Figure 4.A.3: Distribution of the sickness duration in by the month of falling sick for cases between 90 to 360 days

Note: group 1, 2, 3 refer to WAO, trans. WAO and WIA group, respectively.

Table 4.A.1: Frequencies of cases by duration and by months of falling sick

The month when falling sick	Number of sickness cases between 90 to 180 days	Number of sickness cases longer than 180 days
07. 2003	1,970	9,109
08. 2003	4,248	10,284
09. 2003	7,734	12,539
10. 2003	8,223	11,870
11. 2003	7,495	11,216
12. 2003	5,936	9,682
01. 2004	7,883	12,052
02. 2004	6,860	10,628
03. 2004	8,200	12,430

Appendix 4.B Descriptive plots with yearly data

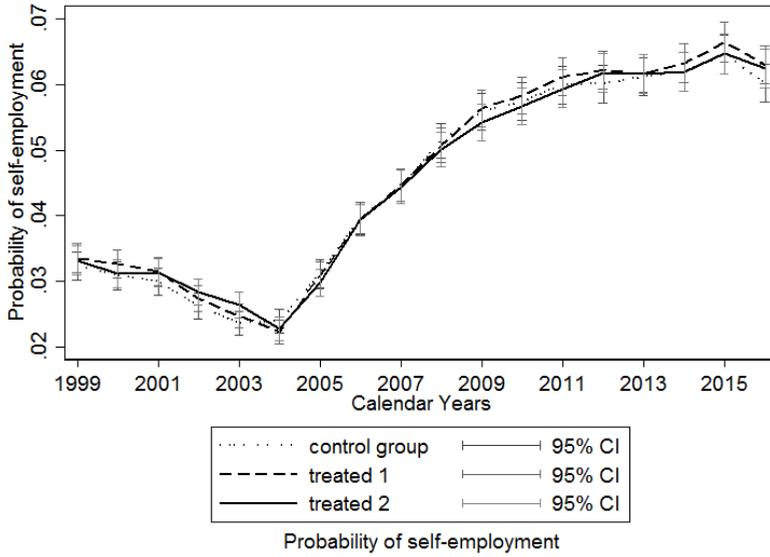


Figure 4.B.1 Yearly average probability of being self-employed

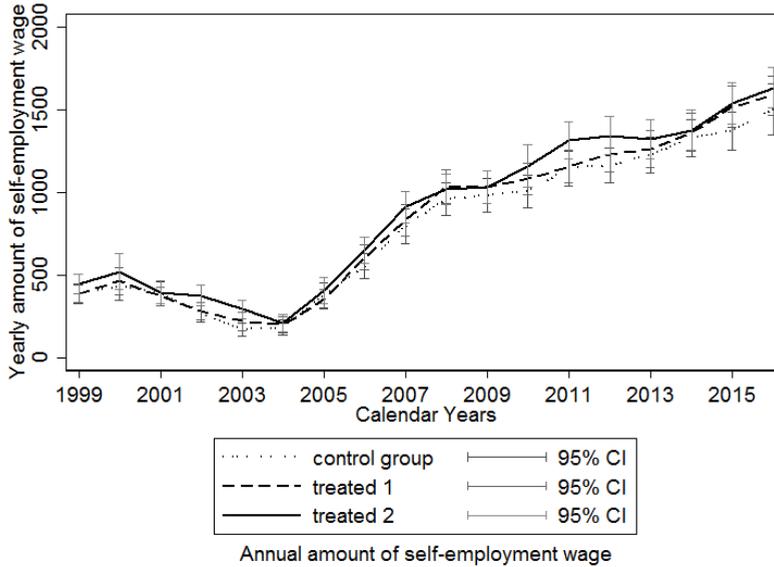


Figure 4.B.2 Yearly average self-employment earnings

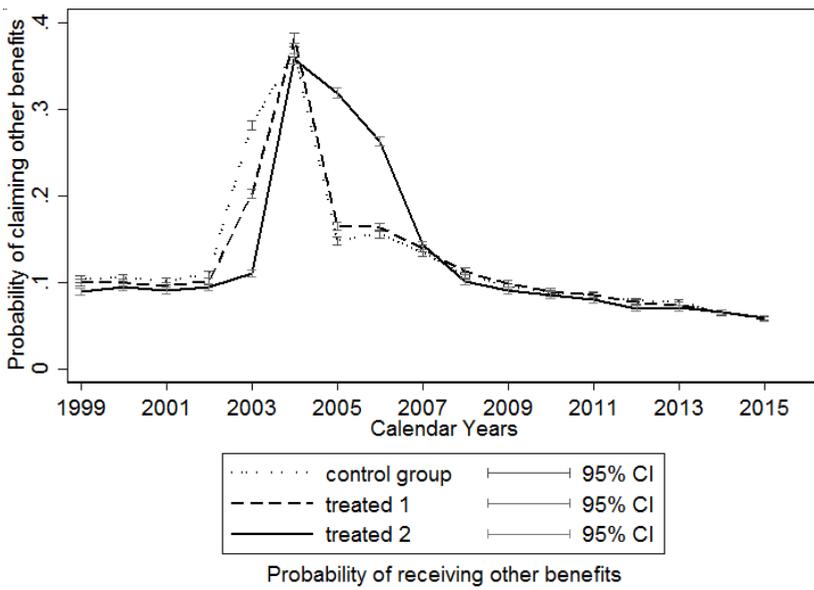


Figure 4.B.3 Yearly average probability of having other benefits and/or sickness benefits

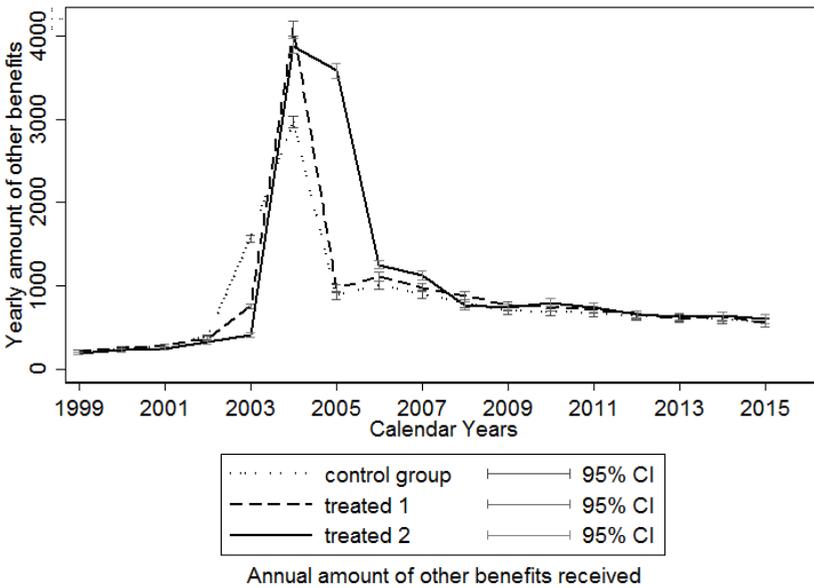


Figure 4.B.4 Yearly average amount other benefits and/or sickness benefits

Appendix 4.C A back-of-envelope calculation to decompose effects of WIA reform on the probability of claiming DI and the probability of working

Table 4.C.1: Share of observations by working and DI claiming status by group

	Pre		Post	
WAO	85.8% of WAO observations in the pre-periods work	→	48.6% fo WAO observations in the post-periods only work, not claiming DI.	1
			4.84% work and meanwhile receive DI.	2
	14.2% do not work	→	14.9% does not work but receive DI.	3
			31.6% does not work and have no DI.	4
Trans. WAO	85.6% of trans. WAO observations in the pre-periods work	→	49.2% fo trans. WAO observations in the post-periods only work, not claiming DI.	5
			4.38% work and meanwhile receive DI.	6
	14.4% do not work	→	14.7% does not work but receive DI.	7
			31.7% does not work and have no DI.	8
WIA	86% of WIA observations in the pre-periods work	→	51.7% fo WIA observations in the post-periods only work, not claiming DI.	9
			2.12% work and meanwhile receive DI.	10
	14% do not work	→	12% does not work but receive DI.	11
			34.2% does not work and have no DI.	12

Note: The shares are calculated using the number of observations in the corresponding work and DI claim status in the corresponding period and group divided by total number of observations in the corresponding period and group.

We only focus on WAO and WIA. WAO and transitional WAO can be calculated similarly.

Given the probability of working in the pre-period is very similar for all groups, we assume in the pre-period the probability of working is the same for all three groups. And the probability of DI is 0 for all groups. Then we can simply use numbers in the post-period to calculate where the effects come from.

(1) Where does the 5.8 percentage point of DI reduction come from?

DID estimates of the effect of WIA on prob. of DI = (WIA's prob. of DI post – WAO's prob. of DI post) – (WIA's prob. of DI pre – WAO's prob. of DI pre) {this pre-term is 0}

$$= \text{WIA's prob. of DI post} - \text{WAO's prob. of DI post}$$

$$= (\text{WIA's prob. of DI post while working} + \text{WIA's prob. of DI post while Not working}) - (\text{WAO's prob. of DI post while working} + \text{WAO's prob. of DI post while Not working})$$

while Not working)

= (WIA-WAO)'s DI post while working – (WIA-WAO)'s DI post while not working

$$= (2.12\% - 4.84\%) + (12\% - 14.9\%) = (-2.72\%) + (-2.9\%) = -5.62\%$$

(2) Where does the 1.8 p.p. of increase in probability of work come from?

DID estimates of the effect of WIA on prob. of working = (WIA's prob. of work post – WAO's prob. of work post) – (WIA's prob. of work pre – WAO's prob. of work pre) {this pre-term is assumed to be 0}

$$= \text{WIA's prob. of work post} - \text{WAO's prob. of work post}$$

= (WIA's prob. of work post while claiming DI + WIA's prob. of work post while Not claiming DI) – (WAO's prob. of work post while claiming DI + WAO's prob. of work post while Not claiming DI)

= (WIA-WAO)'s work post while claiming DI – (WIA-WAO)'s work post while Not claiming DI

$$= (2.12\% - 4.84\%) + (51.7\% - 48.6\%) = (-2.72\%) + (3.1\%) = 0.39\%$$

Note that because we assume away the pre-difference in prob. of working. The 0.39 p.p. looks smaller than our DID estimates in main regression which is 1.8 p.p.

But the idea of the above calculation is:

- (1) The 5.8 p.p. DI reduction come from two parts, 2.72 p.p. reduction of people who claim DI and working, and 2.9 p.p. reduction of people who claim DI and not working.
- (2) But the increase in probability of working is small because on the one hand more fraction of people work while having no DI (3.1 p.p.), but on the other hand, this increase is partly off-set by the fact that fewer fraction of people can partially work and partially claim DI.

Compared to WAO, WIA has:

3.15 p.p. more people only work and not claiming DI

2.72 p.p. fewer people work while claiming DI

2.9 p.p. fewer people does not work but claim DI

2.6 p.p. more people does not work and does not have DI

References for Chapter 4

- Abramitzky, R., Lavy, V. (2014). How responsive is investment in schooling to changes in redistributive policies and in returns?. *Econometrica*, 82(4), 1241-1272.
- Arulampalam, W., 2001. Is unemployment really scarring? Effects of unemployment experiences on wages. *The Economic Journal* 111 (475), F585–606.
- Arulampalam, W., Gregg, P., Gregory, M., 2001. Introduction: unemployment scarring. *The Economic Journal* 111 (475), F577–584.
- Autor, D., Kostøl, A., Mogstad, M., Setzler, B., 2019. Disability benefits, consumption insurance, and household labor supply. *American Economic Review* 109 (7), 2613–54.
- Autor, D. H., Duggan, M., Greenberg, K., Lyle, D. S., 2016. The impact of disability benefits on labor supply: Evidence from the VA's disability compensation program. *American Economic Journal: Applied Economics* 8 (3), 31–68.
- Autor, D. H., Duggan, M. G., 2003. The rise in the disability rolls and the decline in unemployment. *The Quarterly Journal of Economics* 118 (1), 157–206.
- Borghans, L., Gielen, A. C., Luttmer, E. F. P., 2014. Social support substitution and the earnings rebound: evidence from a regression discontinuity in disability insurance reform. *American Economic Journal: Economic Policy* 6 (4), 34–70.
- Burkhauser, R., Marc, D., McVicar, D., Wilkins, R., 2014. Disability benefit growth and disability reform in the US: lessons from other OECD nations. *IZA Journal of Labor Policy* 3 (4).
- Campolieti, M., 2004. Disability insurance benefits and labor supply: some additional evidence. *Journal of Labor Economics* 22 (4), 863–889.
- Campolieti, M., Riddell, C., 2012. Disability policy and the labor market: evidence from a natural experiment in Canada, 1998-2006. *Journal of Public Economics* 96 (3-4), 306–316.
- Chen, S., van der Klaauw, W., 2008. The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics* 142 (2), 757–784.
- Dahl, G. B., Gielen, A. C., 2018. Intergenerational spillovers in disability insurance. *NBER Working Paper* No. 24296.
- De Jong, P., Lindeboom, M., 2004. Privatisation of sickness insurance: evidence from the Netherlands. *Swedish Economic Policy Review* 11, 121–144.
- De Jong, P., Lindeboom, M., van der Klaauw, B., 2011. Screening disability insurance applications. *Journal of the European Economic Association* 9 (1), 106–129.
- Deshpande, M., 2016. The effect of disability payments on household earnings and income: Evidence from the SSI children's program. *Review of Economics and Statistics* 98 (4), 638–654.

- Fevang, E., Hardoy, I., Red, K., 2017. Temporary disability and economic incentives. *The Economic Journal* 1127 (603), 1410–1432.
- French, E., Song, J., 2014. The effect of disability insurance receipt on labor supply. *American Economic Journal: Economic Policy* 6 (2), 291–337.
- García-Gómez, P., Gielen, A., 2018. Mortality effects of containing moral hazard: evidence from disability insurance reform. *Health Economics* 27 (3), 606–621.
- Gelber, A., Moore, T. J., Strand, A., 2017. The effect of disability insurance payments on beneficiaries' earnings. *American Economic Journal: Economic Policy* 9 (3), 229–61.
- Groot, N. D., Koning, P. W. C., 2016. Assessing the effects of disability insurance experience rating. The case of the Netherlands. *Labour Economics* 41, 304–317.
- Gruber, J., 2000. Disability insurance benefits and labor supply. *Journal of Political Economy* 108 (6), 1162–1183.
- Karlström, A., Palme, M., Svensson, I., 2008. The employment effect of stricter rules for eligibility for DI: Evidence from a natural experiment in Sweden. *Journal of Public Economics* 92 (10-11), 2071–2082.
- Koning, P.W. C., 2009. Experience rating and the inflow into disability insurance. *De Economist* 157 (3), 315–335.
- Koning, P. W. C., Lindeboom, M., 2015. The rise and fall of disability insurance enrollment in the Netherlands. *Journal of Economic Perspectives* 29 (2), 151–172.
- Koning, P. W. C., van Sonsbeek, J.-M., 2017. Making disability work? The effects of financial incentives on partially disabled workers. *Labour Economics* 47, 202–215.
- Koning, P. W. C., van Vuuren, D. J., 2010. Disability insurance and unemployment insurance as substitute pathways. *Applied Economics* 42 (5), 575–588.
- Kostøl, A. R., Mogstad, M., 2014. How financial incentives induce disability insurance recipients to return to work. *American Economic Review* 104 (2), 624–655.
- Maestas, N., Mullen, K., Strand, A., 2015. Does delay cause decay? The effect of administrative decision time on the labor force participation and earnings of disability applicants. *IZA Discussion Papers* No. 8788.
- Maestas, N., Mullen, K. J., Strand, A., 2013. Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review* 103 (5), 1797–1829.
- Maestas, N., Song, J., 2011. The labor supply effects of disability insurance: Evidence from automatic conversion using administrative data. *Michigan Retirement Research Center Research Paper* No. 2010-247.
- Mandicó, S. G., García-Gómez, P., O'Donnell, O., 2018. Earnings responses to disability benefit cuts. *IZA Discussion Paper* No. 11410.

- Moore, T. J., 2015. The employment effects of terminating disability benefits. *Journal of Public Economics* 124, 30–43.
- Mullen, K. J., Staubli, S., 2016. Disability benefit generosity and labor force withdrawal. *Journal of Public Economics* 143, 49–63.
- OECD, 2010. *Sickness, Disability and Work: Breaking the Barriers*. OECD Publishing, Paris.
- Ruh, P., Staubli, S., 2018. Financial incentives and earnings of disability insurance recipients: evidence from a notch design. *CEPR Discussion Papers* (12979).
- Staubli, S., 2011. The impact of stricter criteria for disability insurance on labor force participation. *Journal of Public Economics* 95 (9-10), 1223–1235.
- Van Sonsbeek, J.-M., Gradus, R. H. J. M., 2013. Estimating the effects of recent disability reforms in the Netherlands. *Oxford Economic Papers* 65 (4), 832–855.

Chapter 5

Measuring Non-cognitive Skills Exploiting Log-files on Online Behavior⁸⁸

5.1 Introduction

Non-cognitive skills, such as social skills, perseverance, and deep learning, are important components of individuals' human capital. They appear to be good predictors of schooling and labor market career success (see, e.g., Weiss 1988; Heckman, Stixrud, and Urzua 2006; Duckworth et al. 2007). Moreover, their malleability in early life opens the door for interventions through education policy (García 2016; West et al. 2016).

On the other hand, measuring non-cognitive skills is challenging. Conventional measures based upon self-reports using Likert scales are criticized for the lack of measurement comparability, in both the economics and the psychology literature. Bond and Lang (2019) point out that measures reported in ordered intervals (e.g. happiness) are only comparable across groups under a rather strong assumption that all individuals share a common reporting scale. Similarly, van de Gaer et al. (2012) show that self-reported Likert-scale measures are not necessarily comparable in cross-cultural analysis due to heterogeneous reporting behavior. A classic example comes from the motivation-achievement paradox. With data from the Programme for International Student Assessment (PISA), students' self-reported learning motivation is often found to be positively related to academic achievement within each participating country. However, when scores are aggregated at the country level and the correlation is computed between countries' average levels of motivation and achievement, a negative correlation is found. For example, East Asian countries, such as China, Korea, and Japan typically show *high* scores on achievement in the PISA studies, but tend to have *low* scores on learning motivation. Such a paradox is partially attributable to the reference group effect, implying that respondents use different implicit standards (influenced by their immediate social context) in their

⁸⁸ This chapter is coauthored with Jia He. We thank PIAAC team and Leibniz Institute for Research and Information in Education (DIPF) for providing the PIAAC log data. We thank Jochem de Bresser, Marike Knoef, Jan van Ours, Arthur van Soest, and seminar participants at IAIR 2019 Shanghai Conference and CIES 2019 San Francisco for their helpful comments.

self-evaluations, or due to their different styles to present themselves (e.g., response amplification through the tendency to endorse the end points of a scale, or response moderation through the endorsement of the midpoint of a scale).

Various strategies have been proposed to alleviate measurement incomparability of self-reported Likert-scale measures. Correction procedures such as anchoring vignettes and alternative item formats such as forced-choice responses are employed to enhance their comparability (e.g., Kapteyn, Smith, and van Soest 2007; Kyllonen and Bertling 2014; Leising et al. 2015; Robert et al. 2015). Assessment by a third party (e.g. peers) can be used to mitigate the effect of self-presentation styles (e.g., Konstabel, Aavik, and Allik 2006).

Another strand of the literature explores alternative measures of non-cognitive skills. Heckman and Rubinstein (2001) use a behavioral indicator (the “General Educational Development” testing program), as a proxy for (low) non-cognitive skills. Lindqvist and Vestman (2011) utilize administrative records of suitability assessment for military service as a measure of non-cognitive skills. Other alternatives include behavioral observation and coding of survey respondents by interviewers or experimenters (e.g., Renninger and Bachrach 2015) and measuring non-cognitive skills on the basis of task performance (e.g., Reynolds et al. 2006). These behavioral measures are more objective, therefore less plagued by incomparable reporting styles. But they are usually context-specific and not easily generalizable. Moreover, they can be more sensitive to incentives and situational factors (Lundberg 2015).

We add to the literature by proposing a new source of behavioral measures: we propose to use computer-generated logs to construct behavioral indicators for non-cognitive skills. In computer-based assessments (such as the well-known PISA test), log files record respondents’ sequences of actions like keystrokes and mouse clicks etc., from which we extract behavioral measures to quantify certain non-cognitive skills. The objective and unobtrusive nature of such measures makes them immune to respondents’ self-presentation styles or to reference group effects. They thus hold promise in validating self-report data and predicting achievement in a cross-group (esp. cross-cultural) context.

We analyze two examples of non-cognitive skills that are considered important to education policy: perseverance and deep learning (García 2016). The latter refers to

the ability of actively seeking meaning and integrating information in order to understand the material that is taught (Marton and Säljö 1976). We use the log files from the Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC) to construct our behavioral indicators for the two skills, respectively. We show that log-based behavioral measures have higher cross-country comparability than self-assessments, as they predict the performance of tests consistently at individual and country levels, while conventional self-reported measures do not. We also discuss the methodological implications of log-file based behavioral measures, and encourage researchers to apply them in combination with conventional self-reported measures.

5.2 Two Examples: Perseverance and Deep Learning

5.2.1 Example 1: Perseverance in PISA

Data source

We use the log data and background questionnaire of PISA in 2012. The PISA test has a computer-based assessment targeting 15-year-old students in 42 countries. It includes a cognitive test on math, digital reading, and problem solving, and a background questionnaire on various attitudes and behaviors related to learning. Data on the background questionnaire, cognitive tests, and log files of sampled cognitive items are published for public research use on the OECD website (OECD 2013a, 2013b).

Sample restriction

To compare with our log-based behavioural measure, we need to construct a self-reported measure of perseverance from Likert-scale items relating to perseverance. In order to make sure the self-reported measure of perseverance measures the same construct across countries, we perform a multi-group confirmatory factor analysis and country-wise internal consistency checks. We find that the perseverance construct is different in United Arab Emirates, Brazil, Bulgaria, Columbia, Hungary, Malaysia, Montenegro, Slovenia and Serbia. We therefore drop observations from these 9 countries, retaining a sample of 33 countries and 14,888 observations.

Appendix 5.A provides details on the multi-group confirmatory factor analysis and the sample restriction.

Measures for Perseverance

Self-reported perseverance: A continuous measure extracted with factor analysis from 4 perseverance related items. These items are: “When confronted with a problem I give up easily”, “I remain interested in the tasks that I start”, “I continue working on tasks until everything is perfect”, and “When confronted with a problem I do more than what is expected of me.” They all have the same response options ranging from 1 (*very much like me*) to 5 (*not at all like me*).⁸⁹ The value of self-reported perseverance ranges from -2.26 to 1.67, with mean 0 and standard deviation 0.90.⁹⁰ A larger value indicates a higher level of perseverance.

Log-file based behavioral measure of perseverance: We use the total number of clicking the “RESET” button in the “Traffic” unit in the problem-solving cognitive assessment as the behavioral measure of perseverance. The “Traffic” unit provides a map connecting different areas with the following description:

“Here is a map of a system of roads that links the suburbs within a city. The map shows the travel time in minutes at 7:00 am on each section of road. You can add a road to your route by clicking on it. Clicking on a road highlights the road and adds the time to the Total Time box. You can remove a road from your route by clicking on it again. You can use the RESET button to remove all roads from your route.”

The map can be found in Figure 5.1. Respondents are asked to utilize the interactive map to answer three questions: a calculation of the time needed for the shortest route between two specific areas (Question 1), highlighting the shortest route from two areas on the map (Question 2), and selecting the best place for three persons living in different areas to meet, given that no one needs to travel more than 15 minutes (Question 3). The answers to these questions are not obvious without trying different routes. The respondent’s trials and errors until reaching the answer naturally proxy the extent to which the respondent does not give up and persists in pursuing their goal. Perseverance here is therefore operationalized as the total number of trials and

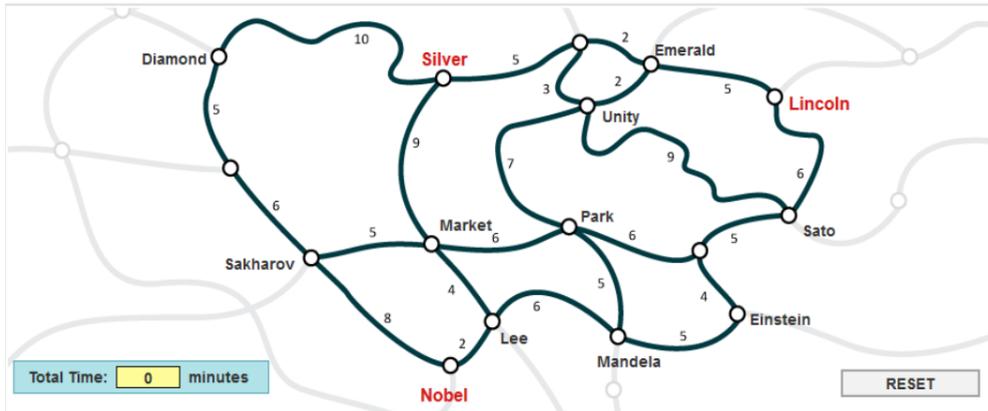
⁸⁹ Item “I put off difficult problems” is dropped due to incomparability. See Appendix 5.A.

⁹⁰ The standard deviation would be 0 if we extract the factor scores in a single CFA model, but we do it in the metric invariance model of the multigroup CFA, where each country gets its own mean of 0 and SD of 1, and pooling them together, the SD is slightly deviating from 1.

errors by clicking the RESET button – we extract the number of resets for each question and add them up to obtain the behavioral indicator of perseverance for each individual. The average ranges from 1.98 in Croatia to 6.23 in South Korea.

TRAFFIC

Here is a map of a system of roads that links the suburbs within a city. The map shows the travel time in minutes at 7:00 am on each section of road. You can add a road to your route by clicking on it. Clicking on a road highlights the road and adds the time to the **Total Time** box. You can remove a road from your route by clicking on it again. You can use the RESET button to remove all roads from your route.



Source: <http://www.oecd.org/pisa/pisaproducts/pisa2012problemsolvingquestions.htm>

Figure 5.1: Screenshot of the Map of the “Traffic” Unit

Outcome variable

Traffic Unit Performance: The total grades for three questions in the Traffic unit. The value ranges from 0 to 3, one point for each correct answer. We use our measures of perseverance to predict this Traffic Unit performance

Control variables

Math achievement score: a larger total number of resets may reflect a lower innate ability or IQ instead of measuring a higher level of perseverance. We use individuals’ math achievement score as a proxy for the respondent’s ability. We randomly take the first of the five plausible values of the math achievement scores provided in the data.⁹¹ It has a mean score of 503.82 and standard deviation of 98.52.

⁹¹ Plausible values are imputed values that can be used to estimate population characteristics correctly (De Leeuw, Hox, and Dillman 2008). In PISA test, each student was given only a subset of the overall cognitive test. To get a comparable math score and also to account for the missing data, PISA test

Total time: respondent could simply have not enough time left for this unit and decide to guess an answer. Then a lower number of resets associate with a low performance score could be confounded by the strategic time planning. We therefore extract and control for the total time spent on the Traffic unit in seconds. Average time spent on the unit is 268.38 seconds and the standard deviation is 113.13 seconds.

Predictive performance

We expect a positive correlation between the level of perseverance and Traffic unit performance. A good measure of perseverance should predict the test performance consistently both at the individual level and at the country level.

Table 5.1: Perseverance and Performance at the Individual Level

Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Traffic unit performance							
Self-reported persev. measure	0.029*** (0.008)	0.057*** (0.008)	-0.005 (0.007)	-0.007 (0.007)				
Behavioral persev. measure					0.011*** (0.001)	0.009*** (0.001)	0.007*** (0.001)	0.002** (0.001)
Math score			0.004*** (0.000)	0.004*** (0.000)			0.004*** (0.000)	0.004*** (0.000)
Total time				0.001*** (0.000)				0.001*** (0.000)
Country Fixed Effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	14,826	14,826	14,826	14,826	14,888	14,888	14,888	14,888

Note: All the columns report OLS estimates of linear regression models. *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. A constant term is also included in all regressions. In columns (1) to (4), there are fewer observations due to missing values in perseverance items.

Table 5.1 reports OLS estimates of linear regression models at the individual level. Columns (1) and (5) both show a positive correlation between perseverance and performance for both the self-assessment measure and the log-file based measure. When adding country fixed effects in column (2) and (6), this positive correlation remains for both measures. For the self-reported measure of perseverance in Column (2), the coefficient becomes larger, possibly because utilizing the variation within a country help to avoid the heterogeneous reporting behavior common to a country. Columns (3), (4), (7), and (8) add controls for general cognitive ability (match

imputes and reports 5 plausible values of test scores for each individual. The results in our analysis are insensitive to the choice of the plausible values.

achievement score) and strategy (total time spent on the unit). The log-based behavioral measure remains positively correlated with Traffic unit performance, while self-reported perseverance no longer has a significantly positive coefficient.

We then aggregate all variables to the country level, i.e. each observation is the country average of a variable. We perform similar regressions as in Table 5.1 at the country level and report OLS estimates in Table 5.2. In column (1), the correlation between the self-reported perseverance and the Traffic unit performance reverses into negative. Similar to the “motivation-achievement paradox”, cross-country comparisons of the self-reported measure may suffer from the reference group effect and the bias from heterogeneity in reporting behavior, leading to a counter-intuitive correlation pattern. On the other hand in column (4), the unobtrusive objective behavioral measure, which is immune to such concerns, still consistently shows a positive relationship with the test performance. Moreover, this positive correlation is robust to adding controls for innate ability and strategic time planning (columns (5) and (6)).

Table 5.2: Perseverance and Performance at the Country Level

Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)
	Avg. Traffic unit performance					
Avg. Self-reported persev. measure	-0.173* (0.097)	-0.106 (0.066)	-0.104 (0.070)			
Avg. behavioral persev. measure				0.057** (0.025)	0.029 (0.018)	0.053** (0.022)
Avg. Math score		0.003*** (0.000)	0.003*** (0.001)		0.003*** (0.000)	0.001* (0.001)
Avg. Total time			-0.000 (0.001)			-0.002* (0.001)
Observations	33	33	33	33	33	33

Note: All the columns report OLS estimates of linear regression models. *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. All variables are country averages. A constant term is also included in all regressions.

5.2.2 Example 2: Deep Learning in PIAAC

To show that the predictive advantage of our proposed log-based behavioral measure is not just by chance, we construct behavioral measures for another desirable non-cognitive skill, namely deep learning, using a different data source and perform a similar predictive exercises as in Example 1.

Data source

We use the data of log files and cognitive test scores from PIAAC 2013. PIAAC has a computer-based assessment in 24 countries targeting working adults aged between 16 and 65 years old. Data of the background questionnaire, cognitive test scores, and log file data of the cognitive assessment in 16 countries are published for research use (OECD, 2017).

Sample restriction

Given the booklet design (planned missing) in the cognitive assessment and the availability of specific unit log file information, we restricted our analysis to respondents in 13 countries with at least one valid response in the targeted log files of problem solving tasks.⁹² These countries are Austria, Belgium, Germany, Estonia, Finland, Ireland, Netherlands, Norway, Poland, Slovak Republic, UK, and USA. The resulting total sample size is 20,167 observations.

Measures for deep learning

Self-reported deep learning: We use the deep learning strategy item “looking for additional information” (I_Q04m) as the self-reported measure of deep learning. The value ranges from low level of deep learning: 1 (*not at all*) to high level: 5 (*to a high extent*). It is a Likert scale measure, but we treat it as continuous in the regressions. The mean level of this variable is 4.05. We do not use other self-reported items (e.g., relate new ideas into real life, get to the bottom of things, like learning new things) because each item in the scale refers to a very different strategy and a combination of these items does not have a clear link to specific behavioural indicators.

Two Log-based behavioral measures of deep learning: We use the *total numbers of different page visits*, extracted from log files of two problem solving units, to construct two behavioral indicators of deep learning. The total number of different page visits reflects the ability of looking for information on different websites in order to make a sound judgment, thus serving as a plausible proxy for deep learning skills. **Behavioral measures of deep learning 1:** The total number of different page visits extracted from the task of “The Sprained Ankle - Reliable/Trustworthy Site”

⁹² Similar as PISA, each respondent in PIAAC only gets a subset of questions. So in some countries, certain questions can be missing.

(PS_u06b).⁹³ This task shows links to five websites recommended by a friend on how to treat sprained ankle. The respondents are supposed to read through these websites and find the most reliable and trustworthy site. In addition to the one-page content on each webpage, three webpages provide an additional tab for links to read more about the site or the author. These additional webpages provide information such as opinions expressed by individual writers, certified surgeon, or president of a commercial equipment supplier, which can assist judgment on the credibility and reliability of the sources. The variable has a mean of 4.51 page visits with a standard deviation of 2.39 visits. **Behavioral measures of deep learning 2:** The total number of different page visits extracted from the task of “The Digital Photography Book Order” (PS_u07). This task provides six links to different vendors of or information on digital photography books. Respondents are asked to buy a book for beginners while staying within a budget of 40 USD in time for a friend's birthday in two weeks. Respondents need to find the most suitable website and place an order. Some websites have additional tabs to check availability, check shipping cost, etc. These are necessary to make the correct decision on vendor choice. The variable has a mean of 8.19 page visits with a standard deviation of 4.14 visits.

Outcome variables

The performance measures for two units: Score for “The Sprained Ankle” and Score for “The Book Order”: Both units have only one correct answer. 1 as correct and 0 as incorrect.

Control variable

Numeracy scale score: Similar to the math achievement score in Example 1, we randomly take the first of the ten plausible values of the numeracy scale score provided in the data, as the proxy for the respondent's innate ability. It has a mean score of 282.12 and standard deviation of 45.52.

Predictive Performance

We do similar predictive checks as in Example 1. We expect positive correlations between the level of deep learning and the performance of “The Sprained Ankle” and “The Book Order” units.

⁹³ We cannot provide any screenshot of PIAAC questions due to confidentiality reasons.

Table 5.3 reports regression results at individual level. In columns (1), (3), (5), and (7) where the self-reported/behavioral measure of deep learning is the only regressor, we do observe that both self-reported and behavioral measures are positively associated with the two performance measures. Columns (2), (4), (6), and (8) add further controls - country fixed effects and numeracy scores. Unlike in the PISA data, we do not have total time here. But in Tables 5.1 and 5.2, results are generally insensitive to controlling for total time, but rather sensitive to controlling for innate ability. So we control for numeracy scores to proxy innate ability. We find that the positive correlation between behavioral measures and performance measures is robust to adding further controls, while the coefficient of self-reported deep learning on the “Sprain Ankle” score is no longer significantly different from 0.

Table 5.3: Deep Learning and Performance at the Individual Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Outcome variable	"Sprain Ankle" score				"Book Order" score			
Self-reported deep learning	0.020*** (0.004)	0.004 (0.004)			0.040*** (0.004)	0.011*** (0.004)		
Behavioral deep learning measure 1			0.052*** (0.001)	0.037*** (0.002)				
Behavioral deep learning measure 2							0.079*** (0.001)	0.072*** (0.001)
Numeracy score		0.003** * (0.000)		0.002*** (0.000)		0.005*** (0.000)		0.001*** (0.000)
Country Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	20,015	20,015	19,996	19,996	20,167	20,167	20,149	20,149

Note: All the columns report OLS estimates of linear regression models. *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. A constant term is also included in all regressions. In columns (1) to (4), (7), and (8), there are fewer observations.

Aggregating variables to the country level, we only have 13 observations. As seen in Table 5.4, none of the self-reported measures are significant, possibly due to the small sample size. On the other hand, we still observe a significant positive relation between behavioral measures and performance measures, indicating that our log-based behavioral measures have better predictive performance than the self-reported measure.

Table 5.4: Deep Learning and Performance at the Country Level

Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Avg. "Sprain Ankle" score				Avg. "Book Order" score			
Avg. Self-reported dp. ln.	0.011 (0.151)	-0.006 (0.165)			-0.035 (0.119)	0.020 (0.118)		
Avg. Behavioral dp. ln.1			0.066* (0.034)	0.104** (0.037)				
Avg. Behavioral dp. ln.2							0.045** (0.015)	0.042** (0.018)
Avg. Numeracy score		-0.001 (0.002)		-0.003* (0.002)		0.002 (0.002)		0.000 (0.001)
Observations	13	13	13	13	13	13	13	13

Note: All the columns report OLS estimates of linear regression models. *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. All variables are aggregated at the country average level. A constant term is also included in all regressions.

5.3 Discussion

We set out to introduce an innovative source of behavioral measures, using log-file data to construct behavioral measures. Log-based behavioral measures are unobtrusive to collect and immune to self-presentation styles and reference group effects. We compare them with self-reported measures for predictive performance in cross-cultural contexts, using existing data from the large scale PISA and PIAAC studies. We show that log-based behavioral measures have better predictive performance than self-assessments, as they consistently have power to predict the test performance both at individual and country level, whereas self-reported measures produce weaker, and sometimes counter-intuitive correlations, especially at the country level.

Complementing Self-Reports with Unobtrusive Behavioral Measures

Log-based behavioral measures have clear advantages over self-reports and lab-based performance tasks. First, they are less sensitive to self-presentation styles and thus more objective; second, they occur in a natural environment instead of in contrived lab-settings and are thus unobtrusive, and third, they are easy to implement (with well-developed and validated tasks) in computer-based assessment, allowing to reach a number of respondents significantly larger than could be achieved in lab settings. The log-based behavioral measures will be especially useful in cross-group/country studies, particularly for data where vignette questions to correct for differences in subjective response scales used in self-assessments are not available.

Log-based behavioral measures can also be used to cross-check the validity of the self-reported measures.

Still, log-based behavioral measures are limited by their availability and specificity. First, they are, obviously, only available when log-file data are available. Given the progress of online data collection methods and the increasing number of new released datasets, we expect to see more log-file data becoming available in the future. Second, the log-based behavioral measure is restricted by the specificity of the task and usually only speaks to one facet of non-cognitive skills, which may limit the generalization of the empirical findings. This limitation of specificity also applies to other behavioral measures (including observational behaviors and lab-based performance task measures). Self-reported measures, on the other hand, can easily capture the multi-dimensional nature of non-cognitive skills, either by asking respondents to give overall rating of non-cognitive skills, or by using dimension reduction techniques to construct a measure from multiple Likert-scale items. It therefore makes sense to use both log-based behavioral measures and self-reported measures complementarily.

Future Directions

Our study is a first step to investigate the potential of this type of measures. We use test performance as the outcome variable. A natural next step would be to investigate the relationship between log-based behavioral measures and education or even labor market outcomes (e.g. linking PISA data to education and labor market related administrative data to investigate the long-run effects of non-cognitive skills). Furthermore, all of our behavioral measures are one-dimensional indicators from the domain of problem-solving. Future studies can design and validate a large variety of tasks in different domains (e.g., reading, numeracy, problem solving) and extract multiple indicators. Dimension reduction techniques can then be considered for constructing a multifaceted log-based behavioral measure.

Appendix 5.A Details on Sample Restrictions of Example 1

Construct Equivalence Check for the Self-Reported Perseverance Scale

For self-reported Likert measures in large-scale assessments, an empirical demonstration of construct equivalence is generally required. Construct equivalence indicates that the same theoretical construct is measured across countries. Without construct equivalence, there is no basis for any cross-cultural comparison of a self-reported measure using a Likert scale (Meredith, 1993).

To check the construct equivalence of our self-reported scale of perseverance, a multigroup confirmatory factor analysis was first conducted with all five perseverance items in all 42 countries, using robust maximum likelihood estimation in Mplus 7 (Muthen & Muthen, 1998-2012). The configural invariance model showed very poor model fit [χ^2 (210, $N = 19,382$) = 6620.71, $p < .01$, CFI = .63, RMSEA = .26], indicating that these items do not measure the same construct across all countries. Therefore, they do not provide a basis for any quantitative comparison.

Sample Restriction

To enable valid comparisons, it is necessary to identify a cluster of countries and a subset of items that show a higher level of comparability (i.e., metric invariance, which implies that factor loadings across countries are identical. With metric equivalence, associations of constructs can be compared across countries, but scale mean scores cannot.). A careful check of factor analysis solution and internal consistency per country revealed that items loaded very differently in nine countries: close-to-zero and reversed loadings in United Arab Emirates, Brazil, Bulgaria, Columbia, Hungary, Malaysia, Montenegro, Slovenia and Serbia). The fact that these countries the self-assessment does not measure the same concept in these countries was also confirmed by the low internal consistency in these countries (with Cronbach's Alpha values ranging from .51 to .59, below 0.7). The poor psychometric properties may in part be attributable to the unfamiliarity of computerized assessment among students in these countries. It was identified that the item "I put off difficult problems" showed sharply varying factor loadings in different countries, possibly due to the different understanding of the translation of "put off". Therefore, these nine countries and this item were excluded from the main analysis.

References for Chapter 5

- Bond, T. N., & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, 127(4), 1629-1640.
- De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008). *International handbook of survey methodology* (p. 399). Taylor & Francis Group/Lawrence Erlbaum Associates.
- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92, 1087. doi:10.1037/0022-3514.92.6.1087
- Garcia, E. (2016). The need to address non-cognitive skills in the education policy agenda. In *Non-cognitive skills and factors in educational attainment* (pp. 31-64). Brill Sense.
- Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, 24(3), 411-482.
- Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review*, 91(2), 145-149.
- Kapteyn, A., Smith, J. P., & Van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, 97(1), 461-473..
- Konstabel, K., Aavik, T., & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality*, 20, 549-566. doi:10.1002/per.593
- Kyllonen, P. C., & Bertling, J. J. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277-286). Boca Raton, FL: CRC Press.
- Leising, D., Locke, K. D., Kurzius, E., & Zimmermann, J. (2015). Quantifying the association of self-enhancement bias with self-ratings of personality and life satisfaction. *Assessment*, 23, 588-602. doi:10.1177/1073191115590852
- Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics*, 3(1), 101-28.
- Lundberg, S. (2015), Non-cognitive skills as human capital, University of California, Santa Barbara.
- Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British journal of educational psychology*, 46(1), 4-11.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543. doi:10.1007/BF02294825

- Muthen, L. K., & Muthen, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthen & Muthen.
- OECD. (2013a). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD Publishing.
- OECD. (2013b). *PISA 2012 technical report*. Paris, France: OECD Publishing.
- OECD. (2017). Programme for the International Assessment of Adult Competencies (PIAAC) log files (Publication no. 10.4232/1.12955). (ZA6712 Data file Version 2.0.0). from GESIS Data Archive
- Renninger, K. A., & Bachrach, J. E. (2015). Studying triggers for interest and engagement using observational methods. *Educational Psychologist, 50*, 58-69. doi:10.1080/00461520.2014.999920
- Reynolds, B., Ortengren, A., Richards, J. B., & de Wit, H. (2006). Dimensions of impulsive behavior: Personality and behavioral measures. *Personality and Individual Differences, 40*, 305-315. doi:<https://doi.org/10.1016/j.paid.2005.03.024>
- Robert, A. A., Donnellan, M. B., Brent, W. R., & Fraley, R. C. (2015). The effect of response format on the psychometric properties of the Narcissistic Personality Inventory: Consequences for item meaning and factor structure. *Assessment, 23*, 203-220. doi:10.1177/1073191114568113
- van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries *Journal of Cross-Cultural Psychology, 43*, 1205-1228. doi:10.1177/0022022111428083
- Weiss, A. (1988). High school graduation, performance, and wages. *Journal of Political Economy, 96*(4), 785-820.
- West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis, 38*, 148-170. doi:10.3102/0162373715597298

CENTER DISSERTATION SERIES

Center for Economic Research, Tilburg University, the Netherlands

No.	Author	Title	ISBN	Published
579	Julius Rüschenpöhler	Behavioural Perspectives on Subsistence Entrepreneurship in Emerging Markets	978 90 5668 580 5	January 2019
580	Khulan Altangerel	Essays on Immigration Policy	978 90 5668 581 2	January 2019
581	Kun Zheng	Essays on Duration Analysis and Labour Economics	978 90 5668 582 9	January 2019
582	Tatiana Zabara	Evolution of Entrepreneurial Teams in Technology-Based New Ventures	978 90 5668 583 6	February 2019
583	Yifan Yu	Essays on Mixed Hitting-Time Models	978 90 5668 584 3	April 2019
584	Daniel Martinez Martin	Unpacking Product Modularity, Innovation in Distributed Innovation Teams	978 90 5668 585 0	April 2019
585	Katalin Katona	Managed Competition in Practice Lessons for Healthcare Policy	978 90 5668 586 7	April 2019
586	Serhan Sadikoglu	Essays in Econometric Theory	978 90 5668 587 4	May 2019
587	Hoang Yen Nguyen	Emotions and Strategic Interactions	978 90 5668 588 1	May 2019
588	Ties de Kok	Essays on reporting and information processing	978 90 5668 589 8	May 2019
589	Yusiyu Wang	Regulation, Protest, and Spatial Economics	978 90 5668 590 4	June 2019
590	Ekaterina Neretina	Essays in Corporate Finance, Political Economy, and Competition	978 90 5668 591 1	June 2019
591	Ruth Wandhöfer	Technology innovation in Financial Markets: Implications for Money, Payments and Settlement Finality	978 90 5668 592 8	June 2019
592	Andinet Worku Gebreselassie	On communicating about taboo social issues in least developed countries: The case of Ethiopia	978 90 5668 593 5	June 2019
593	Filip Bekjarovski	Active Investing	978 90 5668 594 2	June 2019
594	Miguel Sarmiento	Essays on Banking, Financial Intermediation and Financial Markets	978 90 5668 595 9	June 2019

No.	Author	Title	ISBN	Published
595	Xiaoyin Ma	Essays on Alternative Investements	978 90 5668 596 6	June 2019
596	Victor van Pelt	A Dynamic View of Management Accounting Systems	978 90 5668 597 3	June 2019
597	Shuai Chen	Marriage, Minorities, and Mass Movements	978 90 5668 598 0	July 2019
598	Ben Gans	Stabilisation operations as complex systems: order and chaos in the interoperability continuum	978 90 5668 599 7	July 2019
599	Mulu Hundera	Role Conflict, Coping Strategies and Female Entrepreneurial Success in Sub-Saharan Africa	978 90 5668 600 0	August 2019
600	Hao Hu	The Quadratic Shortest Path Problem – Theory and Computations	978 90 5668 601 7	September 2019
601	Emerson Erik Schmitz	Essays on Banking and International Trade	978 90 5668 602 4	September 2019
602	Olga Kuryatnikova	The many faces of positivity to approximate structured optimization problems	978 90 5668 603 1	September 2019
603	Sander Gribling	Applications of optimization to factorization ranks and quantum information theory	978 90 5668 604 8	September 2019
604	Camille Hebert	Essays on Corporate Ownership and Human Capital	978 90 5668 605 5	October 2019
605	Gabor Neszveda	Essays on Behavioral Finance	978 90 5668 606 2	October 2019
606	Ad van Geesbergen	Duurzame schaarste - Een kritische analyse van twee economische duurzaamheids-paradigma's geïnspireerd door de filosofie van Dooyeweerd	978 90 5668 6079	October 2019
607	Richard T. Mason	Digital Enrollment Architecture and Retirement Savings Decisions: Evidence from the Field	978 90 5668 608 6	November 2019
608	Ron Triepels	Anomaly Detection in the Shipping and Banking Industry	978 90 5668 609 3	November 2019
609	Feng Fang	When performance shortfall arises, contract or trust? A multi-method study of the impact of contractual and relation governances on performance in Public-Private Partnerships	978 90 5668 610 9	November 2019
610	Yasir Dewan	Corporate Crime and Punishment: The Role of Status and Ideology	978 90 5668 611 6	November 2019

No.	Author	Title	ISBN	Published
611	Mart van Hulten	Aiming for Well-Being through Taxation: A Framework of Caution and Restraint for States	978 90 5668 612 3	December 2019
612	Carlos Sandoval Moreno	Three essays on poverty measurement and risk protection	978 90 5668 613 0	December 2019
613	Harmke de Groot	Core strength or Achilles' heel: Organizational competencies and the performance of R&D collaborations	978 90 5668 614 7	December 2019
614	Peter Brok	Essays in Corporate Finance and Corporate Taxation	978 90 5668 615 4	December 2019
615	Pascal Böni	On the Pricing, Wealth Effects and Return of Private Market Debt	978 90 5668 616 1	December 2019
616	Ana Martinovici	Revealing Attention: How Eye Movements Predict Brand Choice and Moment of Choice	978 90 5668 617 8	December 2019
617	Matjaz Maletic	Essays on international finance and empirical asset pricing	978 90 5668 618 5	January 2020
618	Zilong Niu	Essays on Asset Pricing and International Finance	978 90 5668 619 2	January 2020
619	Bjorn Lous	On free markets, income inequality, happiness and trust	978 90 5668 620 8	January 2020
620	Clemens Fiedler	Innovation in the Digital Age: Competition, Cooperation, and Standardization	978 90 5668 621 5	June 2020
621	Andreea Popescu	Essays in Asset Pricing and Auctions	978 90 5668 622 2	June 2020
622	Miranda Stienstra	The Determinants and Performance Implications of Alliance Partner Acquisition	978 90 5668 623 9	June 2020
623	Lei Lei	Essays on Labor and Family Economics in China	978 90 5668 624 6	May 2020
624	Farah Arshad	Performance Management Systems in Modern Organizations	978 90 5668 625 3	June 2020
625	Yi Zhang	Topics in Economics of Labor, Health, and Education	978 90 5668 626 0	June 2020

YI ZHANG (Hubei, China, 1985) obtained her Bachelor degree and Master degree in Economics in 2007 and 2010 at Shanghai International Studies University. She worked there as a student advisor for four years until she was enrolled in the research master program in Economics at Tilburg University in 2014. In 2016, she started as a PhD candidate at the department of Econometrics and Operations Research at TiSEM, Tilburg University.

This dissertation contains four essays, each with a different topic in Economics of labor, health, and education. The first essay studies the impact of training on individuals' perceived job match quality. The second investigates the causal effect of retirement on healthcare utilization in China. The third essay analyzes how a major disability insurance reform in the Netherlands influences sick individuals' labor market participation and social benefit claiming. The fourth essay proposes a new measure of non-cognitive skills derived from computer-generated log files on the online test takers' behavior, to address the non-comparability issue of self-reported non-cognitive skill measures.

ISBN: 978 90 5668 626 0

DOI: 10.26116/center-lis-2005