Netspar

Network for Studies on Pensions, Aging and Retirement

NETSPAR ACADEMIC SERIES

Essays on wealth, health, and data collection

Lieke Kools

PhD001/2020-003

E.M. MEIJERS INSTITUUT

INSTITUUT VOOR RECHTSWETENSCHAPPELIJK ONDERZOEK

Essays on wealth, health, and data collection

L. KOOLS

Essays on wealth, health, and data collection

Essays on wealth, health, and data collection

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 21 november 2018
klokke 15.00 uur

door

LIEKE KOOLS
geboren te Boarnsterhim
in 1990

Promotores:	prof.dr. K.P. Goudswaard
	prof.dr. C.L.J. Caminada
Co-promotor:	prof.dr. M.G. Knoef

Promotiecommissie:	prof.dr. R.J.M. Alessie (Rijksuniversiteit Groningen)
	dr. V. Angelini (Rijksuniversiteit Groningen)
	prof.dr. P.W.C. Koning (VU Amsterdam en Universiteit Leiden)
	prof.dr. J.I. van der Rest

Preface

Even though I have often complained that a PhD is written in solitary confinement, in reality the existence of this thesis should be accredited to many others than myself.

First of all, I would like to thank my promotors, Kees Goudswaard and Koen Caminada, for taking the gamble of appointing me a PhD position and providing me with ample freedom to explore outside the boundaries of both the country and the profession. I am not sure if this thesis would have existed had I not been granted such possibilities, but I am absolutely sure this thesis would not have existed in absence of my daily supervisor Marike Knoef. Marike, no matter how pessimistic I entered a meeting with you, your enthousiasm and constant stream of new ideas always made me leave fueled up and ready to tackle the challenges ahead. I would also like to thank the members of the committee, Rob Alessie, Viola Angelini, Pierre Koning, and Jean-Pierre van der Rest for taking the time to read the manuscript of this thesis and providing helpful comments.

To my colleagues, thank you for the daily lunch breaks, trips to the Hoogvliet, Chinese language exhange, and interesting discussions. To Jim in particular, thank you for sharing your LaTeX template. A special thanks goes out to my co-authors Jochem de Bresser, Pierre Koning and Dana Thomson. Thank you for sharing your wisdom with me. Dana, thank you for bringing me inspiration in Southampton.

I leave this PhD not only with new knowledge, but also with many new friends. I am grateful to both the members of the PhD committee and the members of SMO promovendi for providing me the necessary work-related distraction. You made the infamous valley of shit smell a lot

nicer. Daphne, I have learned a lot from your can-do mentality. Gintare and Muied, I am so happy to have met you at Wooldridge's course and to have shared all the fun times and struggles of PhD life with you. I wish I could have chosen you both as a paranimph, but well, Gintare saw me first.

I could not have thought of a better complement to the quiet PhD life, than the chaos at 'de Schans'. All the delicious food, dancing, dating, and long conversations kept me energized and ready to focus. Hanna, who but you, who has known me for most of my life and knows my *hersenspinsels* better than I do, could be my paranimph?

Thank you to all my friends for supporting me throughout. And thank you Tomas for teaching me not to take myself too serious.

Contents

List of Tables

List of Figures

1 Introduction

One of the aims of social insurance programs is to provide a financial safety net to households when encountering adverse circumstances. However, apart from offering mere protection, a system of social insurance can also be designed with the aim to increase overall welfare. In order to make the appropriate design decisions one needs to understand how individuals react to both negative shocks, such as health and wealth shocks, and the system put in place to protect them from these shocks. For example, in order to determine appropriate levels of contributions and benefits in social insurance contracts, one needs to understand how individuals prefer to move resources between different potential life outcomes and how consumption patterns are affected by negative shocks such as illness. Moreover, one needs to understand which (negative) behavior can be provoked by income protection and how such moral hazard can be counteracted by complementary efforts to income support. To gain understanding on such behavioral effects, access to high quality microdata is crucial. Innovations in data collection methods are thus key to a better understanding of the workings of social insurance systems.

This thesis contains four essays. The first two essays are aimed at gaining a better understanding of optimal levels of old-age income protection, by first providing insight in how individuals spread negative wealth shocks over the life course and second estimating how consumption patterns are affected by health declines. The third essay aims to measure the effects of a complementary intervention to sickness benefits, aimed to avoid unnecessary inflow into disability insurance leading up to long-term income losses. The fourth essay evaluates alternative survey data

collection methods, aimed at generating high quality individual-level data for countries that do not have an up-to-date Personal Records Database to sample from. The essays can be read independently and all contain an extensive introduction. This introductory chapter aims to summarize the motivations, research questions, and outcomes of the four essays.

1.1 Spreading wealth shocks over the life course

In the recent years the pension system in the Netherlands has been subject to many changes. In order to counteract the rise in government debt due to an increasing dependency ratio, possibilities for tax-advanced pension savings have been decreased and the statutory retirement age has been increased. At the same time pension funds got into financial problems. Life expentancy increased faster than anticipated, leading to increases in liabilities, and interest rates decreased. Moreover, increases in the old-age dependency ratio limited the possibility of counteracting disappointing stock market returns with increases in pension contributions. In the recent financial crisis pension funds had no choice but to forgo on inflation corrections on pension benefits or even cut pension benefits in nominal terms. At the same time, house prices fell sharply. These developments have raised concerns about pension adequacy: Do households still have sufficient funds to finance their future retirement?

The adequacy of post-retirement income is often assessed in a rather pragmatic way, by holding post-retirement gross income against the benchmark of 70% of average pre-retirement gross income. The idea behind the 70%-benchmark is that retirees no longer need to save, no longer have work-related expenses, and have more time to engage in home-production, thereby reducing their expenses. However, optimal replacement rates may change over time. According to the optimal life cycle model unexpected wealth losses should reduce consumption today ánd in the future. In that sense, individuals may cushion the effects of a shock in future pension income by spreading the loss over several years, leading to lower optimal replacement rates.

Chapter 2 of this thesis aims to answer the question: *"What is the effect of declines in Dutch pension annuities over the period 2008 - 2014 on retirement expenditure goals?"* By answering this question we hope to provide insight into both the extent to which individuals spread wealth losses over the life cycle and how pension adequacy has been affected by the Great Recession. The Dutch institutional framework offers the ideal context to answer such a question, because individuals can exert no influence on how much they contribute to their pension plan, nor how the money is invested. Moreover, the declines in pension annuities came completely unexpected. We make use of linked survey and administrative data, providing us with information on wealth, income, and consumption goals on an individual level.

The results show that indeed individuals react to a drop in pension annuities by lowering their planned post-retirement consumption. However, they also react to a general change in sentiment. The young mainly react to drops in housing wealth, while older individuals react stronger to reductions in pension wealth. Moreover, individuals with high incomes tend to change their planned consumption more than individuals with low incomes. Simulations predict that if individuals would not have lowered their planned expenditures, pension adequacy would have dropped substantially. However, by adapting their goals, individuals counteracted this drop at large part. This implies that individuals did anticipate the future drops in pension income to some extent and are unlikely to encounter unexpected income losses at retirement. However, the drop in (pension) wealth has left them in a worse situation, thereby decreasing overall welfare.

The relationship between health and consumption 1.2

The optimal life cycle framework is a useful tool for evaluating welfare effects of programs such as health insurance and pensions. According to the life cycle model an individual's lifetime utility is maximized if the expected marginal utility of consumption is kept constant over the life cycle, while taking into account factors such as impatience and risk aversion.

The expected marginal utility depends on the likelihood of life events such as job loss or family formation, which either affect future income streams (in the case of job loss) or the utility received from spending an extra euro (in the case of family formation), both altering the optimal level of contributions and benefits in social insurance schemes.

Also ones health status may be included in the model, not only because a health decline affects ones potential earnings capacity, but also because the marginal utility of consumption may depend on health. For example, the utility gained from spending on adventurous holidays may decrease in bad health, whereas the utility gained from spending money on a cleaner may increase. Whether this on average results in an increase or decrease of marginal utility cannot be theoretically determined and may depend on factors such as age, country of residence, and type of health problems. The body of empirical work on this subject has to date not led to conclusive results and is solely focused on the US context. Therefore, chapter 3 of this thesis aims to answer the following question: *"What is the effect of health on the marginal utility of consumption for elderly in Europe?"*

In order to answer this question we develop a methodological framework which relates subjective statements on income adequacy to an intertemporal utility model. The question that we use is common in many different household surveys. The advantage of this framework over closely related methods, is that precise results can be achieved even with relatively short panel data sets. This is especially relevant in the European context, where harmonized panel data sets can still be considered a novelty.

The empirical results indicate that a decrease in physical health positively affects the marginal utility of consumption for the average European. This implies that welfare can be increased by transferring income from periods in good health to periods in bad health. However, a worsening of cognitive health leads to a decrease in the marginal utility of consumption, possibly due to a decrease in the ability to plan and take initiative.

1.3 Counteracting moral hazard problems

Up to 2004 all employees in the Netherlands were covered by Disability Insurance (DI), which would compensate for 70% of the income foregone due to disability. That systems of income protection can be prone to moral hazard was demonstrated by the incredibly high DI enrollment rates in the Netherlands during the 1980s and 1990s, also known as 'the Dutch disease'. The system turned out to offer an attractive alternative for regular dismissals, so that many of the DI beneficiaries actually did not have a work impairment. The trend of increasing DI enrollment rates has been curbed by a system of gate keeping and increased employer responsibilities. Employers are now obliged to continue wage payments during the first two years of sickness. At the same time, both employer and employee are obliged to make efforts towards reintegrating the employee into the workplace. If an employee cannot return to work within two years, he or she enters a DI program with lower levels of income protection.

An example of the efforts undertaken by employer and employee to prevent long term dependency on DI is graded return-to-work. Engaging in (adapted) work during a fraction of the regular working hours may prevent loss of human capital and may even facilitate a quicker recovery of injuries. However, there is also the risk of working too much, leading to stress and strain on the body and thereby slowing down the rehabilitation process. The overall picture derived from the academic literature is that graded return-to-work is an effective measure for shortening sick spells and avoiding permanent work disabilities. However, little is known about how these graded return-to-work trajectories should be set up. Therefore, chapter 4 of this thesis aims to answer the question: *"Does the effectiveness of graded return-to-work depend on (1) weeks waited until start of graded return-to-work; (2) intensity of graded return-to-work; (3) type of work disability?"*

We aim to answer this question using registry data from a Dutch private workplace reintegration provider. This provider helps firms with executing the obligations of the gatekeeper protocol and setting up return-to-work plans for their sick employees. Whether and when sick employees start a graded return-to-work trajectory partly depends on their probability

of recovery, so that ordinary regression results are likely to be biased. The case managers have ample freedom in setting up their treatment plans and some may be more in favor of starting graded return-to-work early, while other prefer to wait. We construct instrumental variables measuring the preference of case managers to start graded return-to-work, to start graded return-to-work early, or to start graded return-to-work at a high intensity and use these to correct for the selection bias.

Graded return-to-work turns out to be even more effective when it is started early on in a sick spell and when it encompasses a substantial amount of hours. Possibly this provides more opportunities for the employee to engage in work processes and to be treated as a 'regular' employee. Graded return-to-work is a less effective tool for individuals who suffer from psychological or psychiatric problems. In those cases it is better to wait a bit longer with starting the graded return-to-work trajectory. In these types of circumstances a presence at the workplace may induce stress, thereby hampering the recovery process. Contrary to earlier literature the results show that even though graded return-to-work is an effective tool to speed up the rehabilitation of sick-listed workers, it does not improve the long-run probability to return to work. This may be because the circumstances for the 'control group' are quite different in the Netherlands than in other countries. Regardless of their participation in graded return-to-work arrangements employees and employers need to stay in contact and agree on ways to return to the workplace. If there is a possibility of recovery this is also likely to be achieved in absence of graded return-to-work.

1.4 Innovations in data collection methods

To evaluate or compare the effectiveness of social insurance programs, comparable and statistically accurate information on unemployment, income inequality, poverty, and health is necessary. Preferably this information is updated on a regular basis. Although in the Netherlands there is a substantial amount of administrative data and large scale surveys, it is harder to obtain such data in low and (some) middle income countries. In these

contexts one is dependent on information from face-to-face household surveys executed once every couple of years. Fielding such surveys is a time-consuming and costly task, such that innovations in data collection methods are essential to facilitate the analysis of social policy in these countries.

The most commonly used method for face-to-face household surveys at this moment is two-stage cluster sampling. When using this method first a number of random geographical areas are selected and next a number of random households within these areas are selected. This second step is necessary, because the defined regions are often too large to canvas the whole area within a day. By using two-stage cluster sampling field work can be concentrated in relatively small regions, however it requires several revisits to the region, thereby increasing both the costs of fieldwork and the risk of missing mobile and non-standard populations. Novel sampling methods such as gridded sampling allow for the definition of smaller geographical areas, such that it becomes possible to enroll all households within the selected areas in the sample. This method, called one-stage cluster sampling, may lead to substantial cost savings, since listing and survey phases can be combined in one day and the region to cover is much smaller. Moreover, non-standard and mobile households are less likely to be excluded from the sample. However, if households with similar characteristics tend to live close together, it may require increasing the total sample size, thereby increasing costs.

In chapter 5 of this thesis we aim to answer the question: *"How many more clusters should be sampled when using one-stage cluster sampling compared to two-stage cluster sampling?"* To answer this question we first develop a synthetic geo-coded micro-dataset covering all households in Oshikoto (Namibia), based on information from recent surveys, census, and spatial covariates. This information is combined using both model-based population generation methods, and clustering and prediction methods. The resulting data have the same statistical properties as the real population, however, real-world geo-coded datasets are rarely publicly released because of the risks of disclosing personal information of households. Based on simulated outcomes of the two survey sampling methods ap-

plied to these data, we determine the required number of clusters for both one-stage and two-stage cluster sampling under different scenario's for clustering of characteristics in the population.

Based on the results of the analysis we conclude that one-stage cluster sampling does not necessarily require increased sample sizes, unless we are in a situation where there is complete socio-spatial segregation for one of the characteristics of interest to the survey. When we do encounter this type of situation, sample sizes may increase by up to thirteen times. However, in most cases the increases are only moderate, with increases of about 1.3 times compared to two-stage cluster sampling, so that one-stage cluster sampling can be a viable alternative to two-stage cluster sampling.

2 Cutting One's Coat According to One's Cloth - How did the Great Recession affect retirement resources and expenditure goals?

Abstract

Pension and housing wealth fell substantially during the Great Recession in many industrialized countries. This raised questions about the development of retirement savings adequacy. Using a unique combination of survey and administrative panel data from before and after the Great Recession in the Netherlands, we investigate co-movements between wealth and retirement expenditure goals. We separate 'pure' wealth effects from common factors such as general pessimism. The estimates show that a shock in annuitized pension wealth of 100 euros reduced retirement expenditure goals with 23-33 euros. Whereas pensions drive the revision of goals for older individuals, the results indicate that individuals between the ages of 25 and 49 are more sensitive to housing wealth. Furthermore, while other studies find that the reaction of *current* consumption to financial shocks is relatively strong for low-income households, we document that *long term* expenditure goals are adjusted more by high-income households. Simulations show that the fraction of individuals falling short with regard to their own retirement expenditure goal would almost have doubled during the Great Recession if individuals would not have adjusted their retirement expenditure goals downward.

A working paper version of this chapter is published as De Bresser et al. (2018) and is currently under review. The chapter is co-authored by Jochem de Bresser and Marike Knoef. The authors thank Rob Alessie, Nicole Jonker, Arthur van Soest, and participants of the Netspar Pension Day 2016, the "Future well-being of the elderly" conference in Montreal, IIPF doctoral school 2016, EALE conference 2017, and the HSZ lunch seminar series at Leiden University Department of Economics.

2.1 Introduction

The recent Great Recession had a detrimental impact on household wealth in Western countries. Disappointing stock market returns had a negative effect on wealth accumulated in funded pension plans and austerity measures increased public pension eligibility ages. Moreover, during the crisis residential property prices declined sharply. The rapid decline of wealth during the crisis raised a host of questions for economic analysis, such as: what is the effect of a wealth shock on consumption, on labor supply, and on retirement behavior? Since household portfolios in Western countries are dominated by pension and housing wealth, concerns have also been expressed about the adequacy of retirement resources.

The life cycle model of household spending, developed by Modigliani and Ando (1960) and Ando and Modigliani (1963), predicts that individuals smooth exogenous wealth shocks over their remaining lifetime. Furthermore, the original life cycle model predicts that the effects of wealth shocks are the same for all asset types. Modern models, however, differentiate between asset classes, since pension and housing wealth differ in many dimensions. For example, there may be transaction costs related to borrowing against illiquid assets such as housing equity. Households may also develop 'mental accounts' that dictate that certain assets are more appropriate to use for current expenditure and others for long-term saving (Thaler 1990).

There is a large body of literature on the effect of wealth shocks on consumption. Several studies find evidence for a substantial causal effect of wealth on consumption,[1] while others find only small effects (e.g. Disney et al. 2010), or conclude that co-movements in consumption and wealth are not generated by a causal relationship, but by common factors such as general optimism or pessimism (Attanatio et al. 2009). Christelis et al. (2015) find that for every loss of 10% in housing and financial wealth, current household expenditures drop by about 0.6% and 0.9%, respectively. Such order of magnitude was also found by Mian et al. (2013) and by Angrisani et al. (2015), and is in line with the prediction of a life-cycle

[1] Paiella (2009) provides an overview of the literature.

model (Poterba 2000). Most studies examining the effect of wealth on retirement behavior find little or no evidence that wealth shocks have a causal effect on the (planned) timing of retirement (e.g. Coile and Levine 2006; Hurd et al. 2009; Goda et al. 2011; Goda et al. 2012 and Crawford 2013).

This chapter contributes to our understanding of wealth effects on household behavior. Instead of investigating the effect of wealth shocks on current consumption or on the (planned) retirement age, we examine the effect of wealth shocks on self-reported minimal retirement expenditure goals. Such goals are important determinants of retirement savings adequacy, which we measure as the difference between annuitized wealth at retirement and retirement expenditure goals on the individual level (De Bresser and Knoef 2015). Because of the aging society, understanding the relationship between wealth and retirement expenditure goals becomes even more crucial, as the generosity of public pensions declines and households become more dependent on financial markets and housing wealth.

This study estimates the 'pure' wealth effect that is the response of retirement expenditure goals to unanticipated wealth shocks. This 'pure' effect is the causal effect that is of interest in most of the literature (Paiella and Pistaferri 2017) and that can, in the context of retirement expenditure goals, mitigate the negative effect of a crisis on retirement savings adequacy. We separate this 'pure' effect from the effects of common macro factors that may be correlated with negative wealth shocks, such as general pessimism and negative expectations about future labor market conditions. The last part of this chapter shows the degree to which co-movements between wealth and retirement expenditure goals were able to compensate a decline in retirement savings adequacy brought about by the Great Recession.

We estimate the effects of shocks to both pension wealth and housing wealth by regressing first differences in retirement expenditure goals on differences in annuitized wealth from before and after the Great Recession. In the Netherlands shocks to pensions are exogenous, since workers cannot choose which pension fund to contribute to, how much to contribute, or which investment strategy to follow. All aspects of participation in

industry-wide funds are outside the control of participants – occupational pensions are a fixed aspect of work in a given industry. The variation in shocks to pension annuities is driven by past investment decisions of pension funds. Moreover, the pension cuts came unexpected, as almost all funds appeared financially fit before the crisis. While the institutional framework renders changes to pension annuities exogenous, home owners could react to the decline in house prices by increasing their mortgage down payments. Therefore, we instrument shocks to net housing wealth, which may be influenced by endogenous mortgage down payments, with shocks in gross housing wealth. Finally, simulations based on a seemingly unrelated regression (SUR) model are used to evaluate retirement savings adequacy with and without co-variation between expenditure goals and resources.

We bring the model to the data using matched administrative and survey data. The survey data contain self-reported retirement expenditure goals for a representative sample of the Dutch population, collected in January 2008 at the eve of the downturn in the financial markets and in December 2014, after some years of recession. A unique feature of these data is that individual panel members can be linked to tax records and administrative data from pension funds and banks. This allows us to construct a complete and precise measure of the financial resources available to households.

The contribution of this chapter to the literature is twofold. First, as far as we know we are the first to investigate the effect of unanticipated wealth shocks on retirement expenditure goals. Although effects of wealth shocks on (short term) consumption are often studied, analyses about the long term are scarce. However, the long term relation is highly important for the development of retirement savings adequacy. The analysis relies on administrative individual-level data on unanticipated wealth shocks, instead of aggregate measures of house and stock price changes that are often used in the literature. Second, this chapter studies to what extent co-movements in wealth and retirement expenditure goals during the Great Recession affected the development of retirement savings adequacy. Retirement savings adequacy is defined by the difference between individ-

ual retirement expenditure goals and annuitized wealth. It is common to measure readiness against a single universal threshold, e.g. a poverty line or a replacement rate of 70% of prior income or expenditures, or using a life-cycle model.[2] However, universal thresholds fail to capture relevant differences in coping strategies, which may have changed after some years of recession. Benchmarks based on the life cycle model are able to take into account differences between households, but have difficulty to accurately reflect heterogeneous preferences without excessive computational burden. This makes an alternative and complementary analysis useful.[3]

The estimation results show that a decrease of 100 euros per month in pension annuities reduced retirement expenditure goals by 23-33 euros. Splitting the sample by age, the estimation results suggest that expenditure goals of older individuals were primarily affected by pensions, while for younger individuals real estate played a more important role. Older individuals may be more likely to see their house as a bequest. For them, a higher house price may simply be a compensation for a higher implicit rental cost of living in the house, but has no real wealth effect (Sinai and Souleles 2005, and Campbell and Cocco 2007). Pensions, on the other hand, may not be salient to young individuals who have yet to accumulate a large part of their pension wealth. Another split, based on the median household income in 2008, reveals that individuals in high-income households adjust their expenditure goals more after a shock to pension wealth than do those with lower incomes. This suggests that while literature on the marginal propensity to consume shows that current consumption of low income individuals is more sensitive to shocks than current consumption of high income individuals, we find that in the long run low income individuals may prefer to use different margins to adapt to changing circumstances. They may, for instance, choose to work longer rather than cut their desired spending (which is in line with the results of Lindeboom and Montizaan

[2]For examples of universal standards of sufficiency see Haveman et al. (2007), Mitchell and Moore (1998) and Skinner (2007). Engen et al. (2005) and Scholz et al. (2006) use life-cycle models to assess preparedness.

[3]Our focus on attaining retirement expenditure goals after retirement means that we do not take into account other reasons to save, such as precautionary or bequest motives. If such additional rationales exist, our analysis should be interpreted as an upper bound on preparedness.

(2018), on planned retirement dates). Comparisons between log-log and level-level estimates suggest that only large drops in annuities resulted in a 'pure' wealth effect. As noted by Browning and Collado (2001), consumers may be less likely to smooth consumption when changes are small and the cost of adjusting consumption is not trivial.

Simulation results indicate that co-movements between wealth and retirement expenditure goals tempered the adverse effect of the Great Recession on retirement savings adequacy considerably. The fraction of individuals who are expected not to be able to afford their minimum retirement expenditure goal increased from 27% to 32%, if we only take pension wealth into account. In case individuals would not have revised their goals, around 50% would not have been able to finance their retirement expenditure goals, based on pensions alone.[4]

The remainder of this chapter is set up as follows. Section 2.2 explains the Dutch pension system and the ways it changed between January 2008 and December 2014. Section 2.3 provides the theoretical underpinning for our empirical analysis. In section 2.4 we present the data used for the analysis, followed by a description of our empirical strategy in section 2.5. Section 2.6 contains the results and section 2.7 concludes.

2.2 Pension reforms and the crisis

The Dutch pension system consists of four pillars: (1) public pension, (2) mandatory occupational pensions, (3) voluntary private pension products such as life annuities, and (4) all other (voluntary) assets such as private savings and housing wealth. In this section we describe these pillars and their developments between January 2008 and December 2014 (the months in which the survey data were collected). In the calculations of projected pension annuities we take these developments into account.

[4]Note that, even though an appropriate decrease in retirement expenditure goals does result in better pension savings adequacy relative to those goals, it still implies that the individual endures a welfare loss. This means that there is less need to worry about individuals adapting their plans appropriately to their new situation. It however does not imply that retirement incomes can decrease without any costs to the individual.

Public pension 2.2.1

The first pillar consists of a flat rate public old age pension, financed through a pay-as-you-go system. For every year that individuals live in the Netherlands, they build up rights to 2% of the full public pension. Individuals who lived in the Netherlands during all 50 years before the statutory retirement age receive a full public pension (50% of the minimum wage for individuals living in a couple and 70% of the minimum wage for singles). For retirees with less than full public pension rights as a consequence of living abroad, and insufficient other resources, the first pillar is topped up with social assistance to guarantee a social minimum.

In 2008 the public pension eligibility age was 65. In 2012 an amendment passed that stipulated a stepwise increase of the public pension eligibility age to 67 in 2023, after which it would be linked to life expectancy. In 2014 legislation was proposed to speed up the increase such that the public pension eligibility age will reach 67 in 2021. If individuals work longer, they will also build up more pension wealth as a consequence of this act.[5] Since there was a lot of media attention for the increase in the statutory retirement age, in our calculations we take the accelerated increase of the retirement age into account (which became an Act of Parliament in 2015).

Occupational pensions 2.2.2

The Dutch save massively for their retirement via occupational pensions. 90% of all employees in the Netherlands have a mandatory pension scheme with their employer (Bovenberg and Meijdam 2001) and for many households pension savings are by far their largest financial assets. About 1344 billion euros is accumulated in Dutch pension funds (end 2017), i.e. on average nearly 175,000 euros per household. Employees cannot choose to which pension fund they want to contribute, but are mostly assigned to a sector-specific fund. Changing pension funds would thus often require to change to a job in a different industry.

[5] Mastrobuoni (2009) and Staubli and Zweimüller (2013) indicate that an in increase in the statutory retirement age is likely to result in an increase in the actual retirement age.

Most occupational pension schemes are defined-benefit plans and for many years people did not worry about their pension. Pension funds had large reserves and participants were not aware of the fact that indexation was conditional on the financial situation of their fund. This changed dramatically in 2008. Whereas in 2007 the average funding ratio was 144% and only 7 pension funds had a reserve deficit, by the end of 2008 300 pension funds had a reserve deficit and the average funding ratio[6] was 96% (source: Dutch Central Bank). Most pension entitlements were no longer indexed for inflation and some entitlements were even cut in nominal terms. For example, large funds for metal electro, metal technologies, and tooth technologies had to cut nominal pensions in 2013 with 5.2%, 6.3% and 7.0%, respectively. The biggest pension fund in the Netherlands (ABP), covering about 2.8 million individuals, has not been able to index pension entitlements and pension benefits since 2010 and on top of that had to cut pensions by 0.5% in 2013. In total the forgone indexation between 2008 and 2014 amounts to 9.93% (source: website ABP).

There are vast differences in funding ratio trends between funds: figure A1 in Appendix 2.A shows that the relative decline in funding ratio during 2008 is spread between 0 and -60%. Such variation is explained by (a) the pursued interest rate hedging policy, (b) the asset mix of the investment portfolio, (c) contributions to the fund, and (d) sensitivity to increased life expectancy (which is higher for funds with a relatively large proportion of young participants) (DNB 2014). Because of these different trends, households are confronted with different shocks to pension wealth. These shocks were unanticipated and are exogenous. Exogeneity is embedded in the system, because individual participants in Dutch pension funds have no influence on their contribution and investment strategy. Moreover, it is difficult to change funds, since funds often cover entire industries (for instance, there is one single fund for all government employees and

[6]The funding ratio is the main measure of financial health of pension funds. The legally required funding ratio in accordance with the European pension fund guidelines is 104.2% (IORP Directive, PbEG 2003/41/EG). Pension funds need to hand over a recovery plan to the Dutch Central Bank if their funding ratio is below 104.2% and need to cut pensions when their funding ratio is below 104.2% in five consecutive years. A fund is allowed to index pensions for price inflation when their funding ratio exceeds 130%. Between 110% and 130% partial indexation is possible (DNB 2016).

another one for all of dentistry). However, they were aware of the developments, since the 2008 Pension Act obliged all pension providers to provide a standardized yearly overview of current and projected entitlements (the Uniform Pension Overview, UPO). The fact that shocks to pension wealth were exogenous and salient makes them interesting to investigate.

Pension reforms took place in the aftermath of the financial crisis. In 2014 annual tax-favored pension accruals have been reduced from 2.25% to 2.15%, and it was decided to reduce them further to 1.875% in 2015. This means that the percentage by which pensions are built up each year is reduced. Moreover, as of 2014 the age that forms the basis for the accrued pension rights increased from 65 to 67. This means that occupational pensions will be less generous for future retirees.

Voluntary private savings 2.2.3

The third pillar plays a relatively minor role in the Netherlands. It is formed by voluntary individual pension products, such as life annuities. The self-employed and individuals with a gap in their pension entitlements are allowed to buy life annuities on fiscally attractive terms. Voluntary retirement savings in savings accounts, stocks and/or bonds are not very common in the Netherlands because of the fiscally attractive and high accumulation of wealth in occupational pension plans. For example, in 2014 the median household in the fifth decile owned only 8,300 euros of financial wealth (source: Statistics Netherlands). Such small amounts are probably precautionary savings rather than aimed for retirement. Finally, households may accumulate housing wealth (the fourth pillar). After a long period of steady increases, house prices have taken a hit between 2008 and 2014, decreasing by 20% on average.

Theory 2.3

How do we expect individuals' retirement expenditure goals to respond to shocks in pension entitlements and housing wealth? We start by explaining

the effect of a shock in pension entitlements on planned consumption during retirement (conditional on housing wealth). Consider the standard problem of a consumer who lives for many periods and chooses optimal consumption to maximize the expected value of a lifetime time-separable utility function. We assume that there are perfect capital markets and that the consumer has no bequest motive. The consumers' problem can be written as

$$\max \mathrm{E}_t \left[\sum_{\tau=t}^{L} \frac{u(c_\tau)}{(1+\rho)^{\tau-t}} \right] \tag{2.1}$$

$$\text{s.t.} \sum_{\tau=t}^{L} \frac{c_\tau}{(1+r)^{\tau-t}} = (1+r)A_{t-1} + \sum_{\tau=t}^{L} \frac{y_\tau}{(1+r)^{\tau-t}}, \tag{2.2}$$

with c_t real consumption in period t, L the final period in the life cycle, $u(\cdot)$ a utility function, ρ the rate of time preferences, r the real interest rate, A_t real net worth at the end of period t, and y_t real income in period t. The optimal solution to this maximization problem is $(c_t^*, c_{t+1}^*, ..., c_L^*)$.

When we assume quadratic preferences or CARA (to obtain a closed form solution) and assume $\rho = r$, the Euler equation becomes $c_t = \mathrm{E}_t[c_\tau], \tau = t+1, \ldots, L$. Substitution of this Euler equation into the expected lifetime budget constraint and re-arranging yields the following closed form solution for c_t:

$$c_t = \left(\sum_{\tau=t}^{L} (1+r)^{t-\tau} \right)^{-1} \left((1+r)A_{t-1} + \sum_{\tau=t}^{L} \frac{\mathrm{E}_t[y_\tau]}{(1+r)^{\tau-t}} \right), \tag{2.3}$$

which can be rewritten as

$$c_t = \left(\sum_{\tau=t+1}^{L} (1+r)^{t+1-\tau} \right)^{-1} \left((1+r)A_t + \sum_{\tau=t+1}^{L} \frac{\mathrm{E}_t[y_\tau]}{(1+r)^{\tau-(t+1)}} \right). \tag{2.4}$$

Using (2.4) the difference between planned retirement consumption (c_R) in the year t and in the year $t+s$ is given by

$$\begin{aligned} &\mathrm{E}_{t+s}[c_R] - \mathrm{E}_t[c_R] = \\ &\left(\sum_{\tau=t+1}^{L} (1+r)^{t+1-\tau}\right)^{-1} \left[\sum_{\tau=t+1}^{t+s} \frac{\mathrm{E}_{t+s}[y_\tau] - \mathrm{E}_t[y_\tau]}{(1+r)^{\tau-(t+1)}} + \right. \\ &\left. \sum_{\tau=t+s+1}^{R-1} \frac{\mathrm{E}_{t+s}[y_\tau] - \mathrm{E}_t[y_\tau]}{(1+r)^{\tau-(t+1)}} + \sum_{\tau=R}^{L} \frac{\mathrm{E}_{t+s}[y_\tau] - \mathrm{E}_t[y_\tau]}{(1+r)^{\tau-(t+1)}}\right], \end{aligned} \tag{2.5}$$

with R the period of retirement. The first part between the square brackets of this equation shows the difference in expected and observed income between t and $t+s$, the second part reflects adjustments in expected income until retirement, and the third part reflects the change in expected pension entitlements. In this paper t is January 2008, just before the crisis and $t+s$ is December 2014. Equation (2.5) shows us that a decline in pension entitlements due to the crisis will have a negative effect on planned consumption during retirement. Furthermore, this equation gives reason to expect that planned retirement consumption of older household is more sensitive to shocks in pension entitlements than that of young households, because pension entitlements constitute a larger part of future income which will be spread out over fewer years.

With regard to shocks in housing wealth it is more difficult to predict theoretically the effect on planned consumption during retirement (as is also explained by Campbell and Cocco (2007) and Attanatio et al. (2011)). On the one hand housing is an asset. Thus, increases in house prices lead to an increase in one's wealth and this may increase current as well as future consumption (the wealth effect). On the other hand, houses provide housing services. For homeowners who expect to live in their current house for a very long time, a higher house price has no *real* wealth effect. For young households, who plan to increase house size later in life, an increase in housing prices leads to an increase in the price of future additional housing services and this may affect current and future consumption negatively. Furthermore, houses can be used as collateral in a loan. An increase in house prices may lead to an increase

in consumption because it allows borrowing constrained homeowners to smooth consumption over the life cycle. In the presence of reverse mortgages, homeowners who expect to live in their current house for a very long time would be able to increase consumption. In practice, however, reverse mortgages are rare and housing prices are more likely to affect the next generation via bequests.

2.4 Data

In this study we match survey and administrative data at the individual level. Section 2.4.1 describes the survey data on retirement expenditure goals, and section 2.4.2 the administrative data on assets.

2.4.1 Survey data

Survey data are taken from the Longitudinal Internet Study in the Social Sciences (LISS panel), gathered by CentERdata.[7] This panel is recruited through address-based sampling (no self-selection), and households without a computer and/or internet connection receive an internet connection and computer for free. This roughly nationally representative household panel (Van der Laan 2009) receives online questionnaires on different topics each month. When respondents complete a questionnaire they receive a monthly incentive. A variety of data is available from studies conducted in the LISS panel.

As a proxy for consumption during retirement ($E_t[c_R^*]$), we use a question regarding retirement expenditure goals elicited from LISS-respondents both in a single-wave study in January 2008, constructed by Johannes Binswanger and Daniel Schunk (Binswanger et al. 2013), and in a single-wave study in December 2014, constructed by the authors.[8] In both studies the

[7] For more information we refer to http://www.lissdata.nl/lissdata/.

[8] The recession in the Netherlands, defined as a period of two quarters of negative GDP growth, started in the second quarter of 2008. The last period of recession was between the third quarter of 2012 and the second quarter of 2013. Appendix 2.B shows the development of consumer confidence between the two surveys.

question is placed at the beginning of the survey, after a couple of items regarding housing costs during retirement. The question is phrased as follows:

> *This question refers to the overall level of spending that applies to you [and your partner/spouse] during retirement. What is the minimal level of monthly spending that you want during retirement? Please think of all your expenditures, such as food, clothing, housing, insurance etc. Remember, please assume that prices of the things you spend your money on remain the same in the future as today (i.e., no inflation).*

We find that people provide reasonable answers to this question. As shown by De Bresser and Knoef (2015) and below, people provide decent answers compared to their current income level. Furthermore, non-retirees provide a similar distribution of answers as retirees (who know what it is to be retired).[9] Finally, we asked people whether they found it difficult to answer this question.[10] In our models we control for the fact that answers given by respondents who indicate they find it difficult to answer could be systematically higher or lower than others. De Bresser and Knoef (2015) found no evidence of systematically different answers from individuals who found it difficult or easy to answer the question.

It is important to understand how respondents interpreted the question. Therefore, in December 2014 at the end of the questionnaire (after other questions about health expectations, health care, and pension expectations) we asked respondents how satisfied they would be with a retirement income of X euros, where X is their self-reported minimal retirement expenditure goal from the beginning of the questionnaire. Most people report a satisfaction level 3 or a 4 on a scale from 1 to 7. In that same survey we also asked respondents about their preferred retirement expenditure goal (taking into account that there is a trade of between current and future expenditures). When asked to rate their preferred retirement income level on a scale from 1 to 7, most people report a 4 or a 5. Both minimal

[9] These descriptives can be found in Appendix 2.C.

[10] Appendix 2.C provides more details about how respondents rated the difficulty of the question.

and preferred retirement expenditure goals increase with income, and the difference between them has the same order of magnitude across income groups (the relative difference even declines a bit from 14% of current income for the lowest income quintile to 9% of current income for the highest income quintile). All of this suggests that respondents did not interpret the question as subsistence consumption, but rather as the amount of expenditure they would need to reach a neutral satisfaction level.

In 2008 it was safe to assume individuals did not take into account health care expenditures when reporting expenditure goals, since long term care costs were almost fully covered by the government and mandatory insurance at that time. By 2014 this was no longer the case, so we asked respondents whether they took health care costs into account in their answer. If so, they were subsequently asked what their minimal expenditures would be without these costs. We analyze minimal expenditures net of health care costs to safeguard comparability.

The survey was administered to household heads and their spouses as from the age of 25, with a reported net monthly household income higher than 800 euros (in this way students are excluded). In 2008 the survey was administered to a random half of the eligible panel members, in 2014 the full eligible sample was included. Descriptive statistics of socio-economic variables can be found in Appendix 2.D.

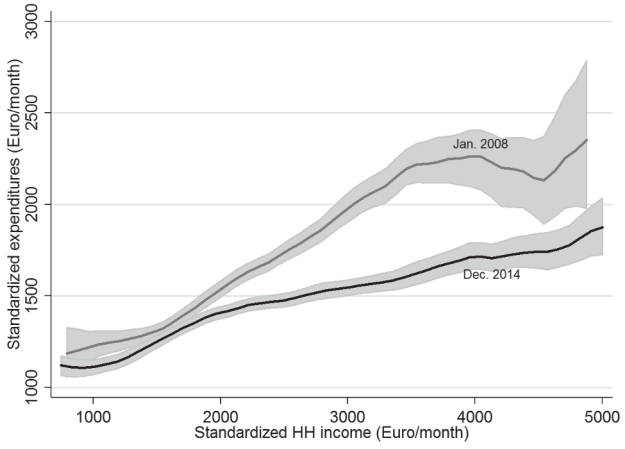
Table 2.1 presents descriptive statistics of self-reported retirement expenditure goals in 2008 and 2014 (both expressed in 2014 euros using the consumer price index). The median retirement expenditure goal dropped by 165 euros (10%), from 1625 euros/month in 2008 to 1460 euros/month in 2014. Both ends of the inter-quartile range (1218 and 2031 euros/month in 2008) also decreased by approximately 10%, indicating that retirement expenditure goals decreased across the distribution. Expenditure goals declined not only in absolute terms, but also relative to current income: replacement rates dropped from a median of 75% in 2008 to 63% in 2014. This can also be seen in figure 2.1, which shows how retirement expenditure goals are related to current income. Reported goals increase with

Table 2.1: Descriptive statistics retirement expenditure goals

		N	Mean	SD	p25	Mdn	p75
a. Retirement expenditure goals							
2008	Monthly expenditures [a]	1396	1744	733	1218	1625	2031
	Replacement rate (%)[b]	1396	76	28	57	75	91
2014	Monthly expenditures [a]	2755	1495	570	1095	1460	1825
	Replacement rate (%)[b]	2717	67	29	47	63	80
b. Changes in Retirement expenditure goals							
	Monthly expenditures	456	-267	640	-571	-227	79
	Replacement rate (%-points)	452	-11	30	-28	-11	5

[a] Retirement expenditure goals are standardized to a one-person household and expressed in 2014 euros.
[b] Replacement rate is defined as the retirement expenditure goal divided by current income.

Figure 2.1: Kernel regressions of retirement expenditure goals on household income (shaded areas are 95% confidence bands).



income in both years, but this relationship was flatter in 2014 compared to 2008.

The bottom panel of table 2.1 describes the differences between 2008 and 2014 for those individuals whom we observe twice. Due to panel attrition the number of individuals who are in the sample in both years is relatively low: we retain around 450 individuals or one third of the 2008 sample. Among those who do remain in the sample, most revised their retirement expenditure goal downwards with a median revision of -227 euros/month. The median revision in the replacement rates is -11%-points. However, there is a lot of variation: a quarter of the individuals reduced their minimum consumption level by at least 571 euros/month, while another quarter of individuals increased their retirement expenditure goal by 79 euros or more. Retirement expenditure goals are fairly strongly correlated across the years: the correlation coefficient is 0.55 for levels and 0.29 for replacement rates.

2.4.2 Administrative data

Administrative data are taken from the Complete Asset Data of the Netherlands 2008 and 2013 (CAD), the Public Pension Entitlements data 2008 and 2012 (PPE), the Public Pension Benefits data 2008 and 2012 (PUBLB), the Occupational Pension Entitlements data 2008 and 2012 (OPE), and the Private Pension Benefits data 2008 and 2013 (PRIVB), all gathered by Statistics Netherlands.

The CAD consists of all households in the Netherlands and contains data on savings accounts, stocks, securities, property, business wealth, and debt. Debt is categorized in mortgage and other debt. Although most of these data are derived from tax records, banks also provide information about bank accounts. Banks have to report accounts with a balance of 500 euros or more (or 15 euros in interest payments), which means that we only miss small amounts of money held in bank accounts.

PPE and OPE contain data on public and occupational pension entitlements for the entire Dutch population between the ages of 21 and 64. PUBLB and PRIVB contain data on public and private pension benefits

received by all retirees (based on tax records). Third pillar pensions (e.g. life annuities) are, unfortunately, only observed in administrative data once they are claimed, because they are subject to taxation only in the payout phase. Therefore, the LISS Assets Survey is used to supplement the administrative data of pre-retirees with survey data on third pillar pension entitlements. We use the administrative records from 2008 to match the survey answers provided in 2008. To match the survey answers provided in 2014 we use the most recent administrative data available and adjust for aggregate changes between the time of measurement and 2014.

Panel a. of table 2.2 summarizes the monthly annuities from pensions and wealth (more details about the wealth data can be found in Appendix 2.E). We use three definitions of after-tax pension annuities: (1) annuities based on public and private pensions, (2) annuities based on pensions plus private wealth other than real estate, and (3) annuities based on all wealth (including real estate). The assumptions used to annuitize wealth can be found in Appendix 2.F.[11] The median projected annuity based on public and occupational pensions declined by around 400 euros, or 20%, from 2146 to 1723 euros/month between 2008 and 2014.[12] We observe similar declines of 15-20% (300-400 euros) for the 25th and 75th percentiles of the distribution, respectively. The smaller absolute decline for the 25th percentile compared to the median and the 75th percentile can be explained by the fact that the flat rate public pension makes up a large share of entitlements for pension-poor households. This public pension tracks the minimum wage and has been adjusted for inflation during the period spanned by our sample (and, according to our assumptions, will be indexed for inflation in the future). Pension-rich households, on the other hand, rely more on occupational pensions, many of which have not been indexed fully for inflation or have even been cut in nominal terms.

[11]The 2008 figures differ slightly from the numbers in De Bresser and Knoef (2015). In that paper the 2008 figures were adjusted to reflect the situation at that time (2014) as closely as possible. In this chapter however, we aim to produce figures as close to the 2008 situation as possible.

[12]The descriptives in table 2.2 refer to the baseline scenario regarding future indexation of pensions, descriptives for other scenarios that are used for robustness checks are available on request.

Table 2.2: Descriptive statistics of assets, debt and annuities

		N	Mean	SD	p25	Mdn	p75
a. Annuities							
	Pensions	900	2170	729	1673	2146	2537
	Percentage of total	890	72	18	61	71	83
2008	Pensions + wealth	890	2401	959	1811	2271	2767
	Percentage of total	890	78	16	68	76	92
	Pensions + wealth + housing	890	3275	1650	2306	3104	3900
	Pensions	3646	1789	768	1352	1723	2135
	Percentage of total	3426	73	33	61	74	93
2014	Pensions + wealth	3429	2103	1447	1479	1890	2409
	Percentage of total	3429	81	20	70	80	100
	Pensions + wealth + housing	3429	2781	1947	1734	2469	3277
b. Changes in annuities between 2008 and 2014							
	Pensions	630	-355	502	-515	-284	-84
2008	Pensions + wealth	597	-298	809	-514	-256	-37
	Pensions + wealth + housing	597	-507	1441	-806	-449	-114
	Pensions	630	-13	19	-22	-12	-5
2014	Pensions + wealth	597	-10	26	-20	-11	-2
	Pensions + wealth + housing	597	-13	26	-24	-15	-4

Monthly standardized annuities in 2014 euros.

Taking non-housing wealth into account does not change the pattern, which suggests that accumulation of discretionary wealth did not compensate much of the decline in pensions across the annuity distribution. The last definition, based on all wealth components, shows the remarkable decline in the value of real estate. The median monthly annuity according to this definition declined by 635 euros (20%), from 3104 to 2469 euros/month. In relative terms the decline is more pronounced for the 25th percentile (572 euros or 25%) than for the 75th percentile (623 euros or 16%).

The bottom panel of table 2.2 describes the distribution of changes in annuities between 2008 and 2014 for those households that we observe twice and can be matched to administrative data in both waves. A similar picture emerges: the crisis and subsequent pension reforms substantially reduced the financial resources available during retirement. The median attainable pension (public plus private), dropped by around 20% due to reductions in real occupational pension entitlements. Furthermore, annuities based on all wealth declined by a similar percentage as a result of the decline in house prices.

Empirical strategy 2.5

Our empirical strategy follows two steps. First, we use the subsample of individuals whom we observe before and after the Great Recession to investigate the size and nature of co-movements between pension wealth, housing wealth and retirement expenditure goals (described in section 2.5.1). Second, we simulate to what extent co-movements between wealth and retirement expenditure goals mitigated adverse effects of the Great Recession on retirement savings adequacy (described in section 2.5.2).

2.5.1 Size and nature of co-movements between wealth and retirement expenditure goals

In the first step, we investigate the relationship between wealth shocks and retirement expenditure goals. Based on the equation (2.5), we regress changes in retirement expenditure goals on changes in wealth, controlling for common factors and demographic variables. More precisely, we estimate[13]

$$\Delta R_i = \beta_0 + \beta_1 \Delta PA_i + \beta_2 \Delta HA_i + \mathbf{\Delta x}_i \boldsymbol{\beta}_3 + \varepsilon_i \tag{2.6}$$

$$\Delta HA_i = \gamma_0 + \gamma_1 \Delta HP_i + \gamma_2 \Delta PA_i + \mathbf{\Delta x}_i \boldsymbol{\gamma}_3 + \nu_i \tag{2.7}$$

In (2.6) ΔR_i is the change in retirement expenditure goals between 2008 and 2014 for individual i, PA_i is the pension annuity, HA_i is the annuity from net housing wealth,[14] and ε_i an error term. $\mathbf{x}_i$ contains individual-level covariates such as income, education, marital status and primary activity.

In addition to estimation of equation (2.6) by OLS, we use 2SLS to disentangle exogenous variation in housing wealth from individual decisions (e.g. extra mortgage down payments). Similar to Angrisani et al. (2015) we instrument shocks in net housing wealth with shocks in house prices (HP_i in equation (2.7)). However, Angrisani et al. (2015) and most of the literature use regional variation in the development of house prices to identify the causal link between shocks in housing wealth and current spending, because reliable data on housing wealth at the household level are rare. Our administrative data do allow us to exploit shocks in house prices at the household level. In this way we can also exploit the idiosyncratic component of house price risk specific to each dwelling (e.g. variation across neighborhoods and types of buildings) to identify the causal effect of housing wealth shocks on changes in retirement expenditure goals (β_2).

[13] This framework is comparable to the framework used by others, such as Parker (1999), Johnson et al. (2006), Agarwal et al. (2007), Disney et al. (2010), and Christelis et al. (2015).

[14] Defined as the difference between the total annuity and the annuity from pensions and non-housing wealth.

Causal effects of changes in wealth are called 'pure' or direct wealth effects. Another possibility is that there are common macro-economic factors that affect both consumption and wealth. For example, future income prospects or general optimism or pessimism may influence both asset prices, house prices, and retirement expenditure goals. Distinguishing 'pure' wealth effects from common factors is important as they have a different impact on the development of retirement savings adequacy after a wealth shock. Pure wealth effects diminish the negative effect of a recession on retirement savings adequacy (measured by the difference between expenditure goals and resources). Common factors can also contribute to mitigate this negative effect on retirement savings adequacy, but to a lesser extent. By definition common factors affect all individuals regardless of the size of their change in wealth. Unlike 'pure' wealth effects, such aggregate adjustments of goals are not concentrated among those individuals who experience large shocks.

To identify the 'pure' effect of pension wealth, we exploit variation across households in pension wealth shocks (ΔPA) brought about by the Great Recession. As explained in section 2.2.2, pension contributions are mandatory in the Netherlands. Participants cannot choose their own pension fund, set their level of contributions, or influence the investment strategy. This implies that changes to pension wealth are plausibly exogenous and we can interpret β_1 as the 'pure' effect of pension wealth.

It could be argued that common macroeconomic factors, such as optimism, pessimism, and risk aversion, affect both asset prices (Campbell 1991) and retirement expenditure goals. In this way, macroeconomic factors could be a third factor influencing both pension wealth shocks and retirement expenditure goals. However, this would not impede us from identifying a 'pure' effect of pension wealth shocks, because the identification relies on variation between pension funds. This variation is caused by differences in the pursued interest rate hedging policy of the fund, the asset mix of the investment portfolio, contributions, and the average age of the participants in a fund. These factors cannot be influenced by individual households. Common macroeconomic factors such as general

pessimism are captured by β_0. β_0 also contains age effects, as age and period effects can not be distinguished in our model.

Finally, one could argue that during the crisis households may have observed reduced rates of return on retirement saving. This could lead to lower voluntary retirement saving, through a substitution effect, and hence cause lower retirement expenditure goals. However, as explained in section 2.2.3 retirement savings in voluntary private saving accounts are rather low in the Netherlands. Even if households would halve their private retirement savings, this would be inconsequential compared to the accumulated wealth in pension funds. Moreover, table 2.2 shows that although the perceived profitability of savings may have reduced, private savings on average increased between 2008 and 2014, probably because of increased precautionary savings.[15]

2.5.2 Co-movements and the development of retirement savings adequacy

The last part of this chapter analyzes to what extent co-movements between wealth and retirement expenditure goals mitigated the negative effect of the Great Recession on retirement savings adequacy. To this end, we compare the simulated preparedness based on a SUR model describing changes in wealth and retirement expenditure goals with a counterfactual scenario in which goals are kept constant at their 2008 level. In this way we isolate the impact of revisions in expenditure goals on the development of the adequacy of retirement resources.

We estimate SUR models to analyze how wealth and retirement expenditure goals of different socio-economic groups changed during the crisis. In these models we utilize data on all individuals (also those whom we observe only in 2008 or 2014). Separate equations for annuities and expenditures in 2008 and 2014 allow the relationships between goals and resources on the one hand and individual and household characteristics on the other to be different in 2014 compared to 2008. Hence, socio-economic

[15] Table 2.2a shows that the average annuity from private savings (excluding housing wealth) increased 83 euros ((2103-1789)-(2401-2170)) and table 2.2b shows that for those individuals whom we observe twice private savings increased 57 euros (-298-(-355)).

groups are allowed to be affected differently by the recession (or, alternatively, the composition of subgroups may have changed). Moreover, we allow the error terms of the equations for expenditure goals and annuities to be correlated between individuals in a given household and across the waves in which the household participates.

The model consists of six equations, three for 2008 and three for 2014:

$$M_i^t = \mathbf{x}_{m,i}^{t\prime}\boldsymbol{\beta}_m^t + u_{m,i}^t \tag{2.8}$$

$$N_i^t = \mathbf{x}_{n,i}^{t\prime}\boldsymbol{\beta}_n^t + u_{n,i}^t \tag{2.9}$$

$$W_i^t = \mathbf{x}_{w,i}^{t\prime}\boldsymbol{\beta}_w^t + u_{w,i}^t \tag{2.10}$$

where M_i^t is the log of the retirement expenditure goal reported by the man in household i in wave $t \in \{2008, 2014\}$ and N_i^t is the log of the retirement expenditure goal reported by the woman in household i and wave t. For singles, only one of the equations for minimal expenditures is relevant for each year (depending on gender). W_i^t is log annuitized household wealth, and $\mathbf{x}$ is a vector containing individual and household characteristics. We assume that the error terms follow a 6-variate normal distribution with mean zero and covariance matrix Σ and estimate the SUR model by maximum likelihood (see Roodman 2011, for details on the CMP command that we used to estimate the model in Stata). Differences between the estimated coefficients for 2008 and 2014 reveal how the crisis (and the subsequent reforms) affected retirement goals and wealth for different socio-economic groups.

To assess the effect of co-movements between wealth and retirement expenditure goals on retirement savings adequacy, we use the SUR estimates to simulate preparedness in 2008, in 2014 and for the counterfactual scenario with annuities at their 2014 level and retirement expenditure goals at their 2008 level. We simulate goals and annuities for all individuals in the sample, regardless of whether they are actually observed in the data (to safeguard representativeness for the Dutch population). Since the dependent variables are missing at random conditional on covariates, the model estimates allow us to simulate preparedness in a way that is

representative for the Dutch population.[16] Simulations are based on an expanded sample in which we replicate each observation 50 times (replicated observations have the same values of covariates but different error terms). From this expanded sample we calculate descriptive statistics of the distribution of the difference between annuities and retirement expenditure goals. Confidence intervals are obtained by means of parametric bootstrap consisting of 500 draws of parameter vectors from their estimated asymptotic distribution. We control for perceived question difficulty by setting the difficulty of imagining how much one would need to spend in retirement to the lowest value.

2.6 Results

2.6.1 Results on size and nature of co-movements

Table 2.3 presents estimation results for the model described in equations (2.6) and (2.7). The baseline estimates reported in column (1) show that a 1 euro drop in the pension annuity reduced retirement expenditure goals with 33 cents on average. So, one third of the drop in pension wealth is compensated by lower retirement expenditure goals (the remainder could be compensated by working longer, saving more or reducing bequests). The coefficient on real estate is 0.06 and not statistically significant. The constant, which captures aggregate common factors like pessimistic future income prospects, is large though insignificant.

To establish that results are not driven by outliers, we rerun the model after winsorizing changes in both goals and annuities. The results in column (2) show that the effect of the change in pension annuities on expenditure goals becomes smaller, but remains economically and statisti-

[16] The sub-sample for which we observe both wealth and retirement expenditure goals is not representative for the population. That is caused by substantial non-response to the expenditure question and incomplete linkage with administrative data for both years in our sample. De Bresser and Knoef (2015) show that non-response and failure to match administrative records are correlated with observed characteristics that are related to goals and resources, such as income. However, they also show that selection into the sample is exogenous once we condition on those observed characteristics.

Table 2.3: Shocks to annuities and changes in expenditure goals

	Dependent variable: Δ retirement expenditure goal			
	(1) OLS	(2) OLS	(3) 2SLS	(4) 2SLS
Δ Pension (β_1)	0.332**	0.229**	0.332**	0.0672
	(0.141)	(0.0983)	(0.134)	(0.0988)
Δ Real estate (β_2)	0.0591	0.152	0.0599	-0.0398
	(0.108)	(0.120)	(0.117)	(0.0327)
Constant (β_0)	-103.3	-132.3***	-103.1	-0.153***
	(71.4)	(44.9)	(68.0)	(0.0278)
Wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, n-1)	–	–	116.1***	21.4***
Endogeneity Δ real estate F(1, n-1)	–	–	2.07e-04	5.77**
n (number HHs)	282	282	282	272
N (total obs.)	307	307	307	296

[a] Δ Annuities and Δ expenditures are winsorized at p5 and p95. Winsorizing at p1-p99, or p2.5-p97.5 leads to similar results.
[b] This column regresses Δ log (goals) on Δ log (annuities).
The models also control for the individual-level covariates listed in Appendix 2.D (with the exception of gender, age, education and degree of urbanization, since those variables display little or no variation within individuals). Annuities and expenditures are standardized to a one-person household. Standard errors clustered at the household level, in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

cally significant at 0.23. Moreover, the size and especially the precision of the constant increases, providing stronger evidence for the role of common factors. On average expenditure goals declined by 132 euros in 2014 relative to 2008, keeping pension and real estate annuities constant. The 2SLS estimates in column (3) are virtually identical to the OLS results in column (1), which is confirmed by failure to reject the null of the endogeneity test for the annuity from real estate. Finally, when we express both annuities and goals in logs rather than levels, we find no evidence for an effect of either annuity on the retirement expenditure goals (column (4)). However, we do find a significant overall reduction in average goals of 15%.

The differences between models in logs and levels may indicate that our results are mainly driven by individuals that experienced substantial wealth shocks. Though the estimates do not depend on those in the lower or upper five percent of the distributions of changes in annuities and goals, it appears that only larger reductions in wealth lead to downward revisions in retirement expenditure goals. This could be explained by bounded

rationality, mental accounting or inattention underlying the 'magnitude hypothesis'.[17] This hypothesis states that individuals do smooth large income shocks, but that they will not bother to adjust optimally to small income changes.

Panels a. and b. of table 2.4 report estimates for subsamples based on age. All models in levels, columns (1)-(3), tell a similar story: expenditure goals of people younger than 50 are affected by changes in the annuity from real estate, while the goals of older individuals are influenced more strongly by pension annuities. For the younger group the estimates based on winsorized data show that a decrease of 1 euro in the expected monthly annuity from real estate reduced expenditure goals significantly with 37 cents. A similar decline in the expected annuity from pensions reduced goals insignificantly with 2 cents. For older individuals the pattern is reversed: the coefficient on the real estate annuity is 0.03 (insignificant), while the coefficient on pensions is 0.34 (significant).[18] The 2SLS model in column (3) does not indicate endogeneity for the annuity from real estate in either sample. Panel c. shows that the differences between coefficients for the two samples are marginally significant for non-winsorized data, but only the difference in the effect of pensions remains significant once we winsorize. Furthermore, the effect of common factors, estimated on winsorized data in table 2.3, seems to be driven primarily by older individuals for whom expenditure goals declined by 172 euros on average (conditional on wealth shocks). Finally, column (4) shows that these results are not robust to taking logs of annuities and goals (elasticities). The models in logs do not provide sufficient evidence to conclude that either annuity affects expenditure goals in either subsample. As described above, this may be explained by the magnitude hypothesis. We do find significant overall declines for both samples: goals were reduced by 10% for the young and by 21% for the older age group.

[17] Evidence supporting the magnitude hypothesis can be found in Browning and Collado (2001), Hsieh (2003), Coulibaly and Li (2006) and Scholnick (2013).

[18] Similar conclusions can be drawn from the quantile models in Appendix 2.G. Following Christelis et al. (2015), Appendix 2.H shows that these results are largely confirmed in models that only control for household composition. Hence, they are not driven by the potential endogeneity of some of our control variables.

Table 2.4: Shocks to annuities and changes in expenditure goals - heterogeneity by age group

	Dependent variable: Δ retirement expenditure goal			
	(1) OLS	(2) OLS	(3) 2SLS	(4) 2SLS
a. Age 25-49[c]				
Δ Pension (β_1)	0.0625	0.0161	0.0618	-0.0133
	(0.114)	(0.151)	(0.115)	(0.166)
Δ Real estate (β_2)	0.229***	0.371**	0.219**	-0.0304
	(0.0640)	(0.181)	(0.0875)	(0.0343)
Constant (β_0)	-86.2	-81.7	-87.9	-0.100**
	(63.9)	(75.7)	(66.2)	(0.0437)
Wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, 117)	–	–	34.2***	670.1***
Endogeneity Δ real estate F(1, 117)	–	–	0.020	9.2***
n (number HHs)	118	118	118	118
N (total obs.)	129	129	129	129
b. Age 50+[c]				
Δ Pension (β_1)	0.419**	0.341***	0.419**	0.0741
	(0.182)	(0.115)	(0.182)	(0.125)
Δ Real estate (β_2)	-0.0455	0.0257	-0.0286	-0.111
	(0.144)	(0.141)	(0.156)	(0.150)
Constant (β_0)	-142.5	-171.5***	-137.9	-0.205***
	(105.4)	(57.5)	(104.6)	(0.0592)
Wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, n-1)	–	–	134.1***	7.2***
Endogeneity Δ real estate F(1, n-1)	–	–	0.047	1.30
n (number HHs)	168	168	168	158
N (total obs.)	178	178	178	167
c. Difference between ages 25-49 and 50+				
(H_0: equal coefficients; statistics follow $\chi^2(1)$ distribution)				
Δ Pension (β_1)	2.75*	2.95*	–	–
Δ Real estate (β_2)	3.03*	2.28	–	–
Constant (β_0)	0.21	0.91	–	–

[a] Δ Annuities and Δ expenditures are winsorized at p5 and p95. Winsorizing at p1-p99, or p2.5-p97.5 leads to similar results.
[b] This column regresses $\Delta \log(\text{goals})$ on $\Delta \log(\text{annuities})$.
[c] OLS models on age sub-samples are estimated jointly.
The models also control for the individual-level covariates listed in Appendix 2.D (with the exception of gender, age, education and degree of urbanization, since those variables display little or no variation within individuals). Annuities and expenditures are standardized to a one-person household. Standard errors clustered at the household level, in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2.5: Shocks to annuities and changes in expenditure goals – heterogeneity by income

	Dependent variable: Δ retirement expenditure goal			
	(1) OLS	(2) OLS	(3) 2SLS	(4) 2SLS
a. Low net household income in 2008[c]				
Δ pension	-0.0192	-0.00951	-0.0200	0.00246
	(0.0902)	(0.134)	(0.0894)	(0.135)
Δ real estate	0.0330	0.152	0.112	0.0228
	(0.115)	(0.190)	(0.230)	(0.0292)
Constant	-99.6*	-100.1*	-86.0	-0.1010***
	(51.8)	(52.6)	(60.4)	(0.0366)
wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, n-1)	–	–	41.7***	15.5***
Endogeneity Δ real estate F(1, n-1)	–	–	0.18	0.24
n (number HHs)	137	137	137	134
N (total obs.)	149	149	149	146
b. High net household income in 2008[c]				
Δ pension	0.484**	0.314**	0.487**	0.130
	(0.200)	(0.133)	(0.199)	(0.135)
Δ real estate	0.0273	0.0518	0.0498	0.134
	(0.127)	(0.140)	(0.134)	(0.0922)
Constant	-109.7	-204.5**	-102.0	-0.1360***
	(13.9)	(83.0)	(137.7)	(0.0506)
wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, n-1)	–	–	86.2***	110.8***
Endogeneity Δ real estate F(1, n-1)	–	–	0.14	1.84
n (number HHs)	145	145	145	138
N (total obs.)	158	158	158	150
c. Difference between low and high income groups				
(H_0: equal coefficients; statistics follow $\chi^2(1)$ distribution)				
Δ pension	5.27**	2.95*	–	–
Δ real estate	0.00	0.18	–	–
Constant	0.00	1.13	–	–

[a] Δ annuities and Δ expenditures are winsorized at p5 and p95. Winsorizing at p1-p99, or p2.5-p97.5 leads to similar results.

[b] This column regresses $\Delta \log(\text{goals})$ on $\Delta \log(\text{annuities})$.

[c] OLS models on income sub-samples are estimated jointly. Cutoff between low and high income group is chosen to include about half of the respondents in each group. The models also control for the individual-level covariates listed in Appendix 2.D (with the exception of gender, age, education and degree of urbanization, since those variables display little or no variation within individuals). Annuities and expenditures are standardized to a one-person household. Standard errors clustered at the household level, in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

The data point towards heterogeneous effects of shocks to different components of wealth. Moreover, sensitivities of goals to wealth components vary across age groups. This might reflect age variation in mental accounts: the young may be more likely to see real estate as a means to finance retirement, while older individuals may see their house as a bequest more often. There is some suggestive evidence for this, since 51% of the older respondents indicate they are not willing to move house in order to free resources in retirement, compared to 34% of the younger group. For those who want to live in their current house as long as possible, a higher house price has no real wealth effect. Alternatively, housing may be more salient to the young while pensions are more salient to older people. Yet another interpretation is that different age groups interpret shocks differently, with younger individuals more likely to see shocks to pension entitlements as transitory.

Table 2.5 shows that the results are mainly driven by households with a relatively high income level. Thus, although the marginal propensity to consume out of shocks is found to be larger for households with a low amount of resources (McCarthy 1995, Dynan et al. 2004, and Johnson et al. 2006), we find that low-income households adjust their long run consumption less after a wealth shock. This could be due to low income households having fewer possibilities to adjust their retirement expenditure goals downward, as they have relatively more essential spending. In the long run low income households may prefer to retire later, rather than to lower their retirement expenditure goals (this is in line with results on planned retirement dates studied by Lindeboom and Montizaan 2018).

Unfortunately, there is little overlap between the samples for 2008 and 2014 and this reduces our sample size. Though the clean, individual-level measurement of wealth shocks from administrative data and the innovative outcome variable are contributions to the literature, we should view the results with caution.

Figure 2.2: Simulated preparedness for retirement: fraction that cannot afford retirement expenditure goals (spikes are 90% CIs)

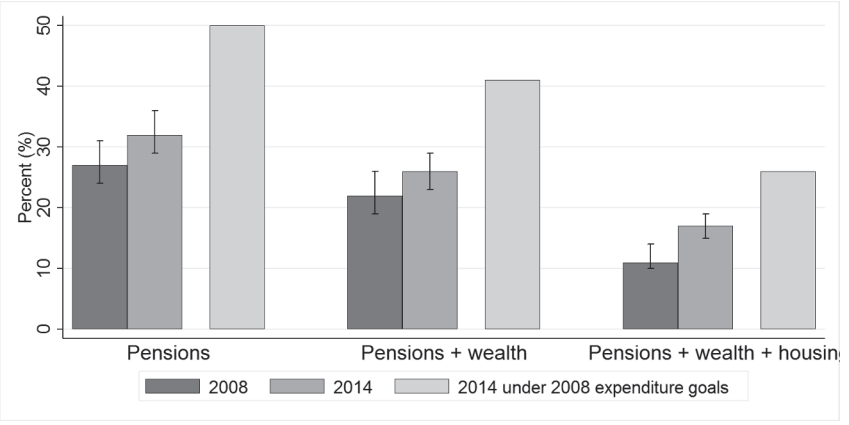


Table 2.6: Aggregate simulation results: differences between annuities and expenditure goals

		Pensions	Pensions + wealth	Pensions + wealth + housing
2008	Median goal (2014 euro)	1565 (1494; 1648)	1560 (1492; 1645)	1561 (1491; 1645)
	Median annuity (2014 euro)	1989 (1964; 2013)	2146 (2119; 2179)	2795 (2758; 2838)
	Median difference (%)	24 (19; 29)	32 (27; 37)	57 (52; 61)
2014	Median goal (2014 euro)	1371 (1310; 1437)	1375 (1313; 1442)	1376 (1315; 1444)
	Median annuity (2014 euro)	1656 (1644; 1670)	1846 (1829; 1866)	2314 (2290; 2338)
	Median difference (%)	20 (15; 25)	31 (26; 35)	53 (48; 58)

90% confidence intervals in parentheses. CIs are obtained by parametric bootstrap over the asymptotic distribution of the ML estimator (500 iterations). In each iteration we replicate the sample 50 times.
Simulations are corrected for over-representation of homeowners in the LISS panel. Understanding of items measuring consumption goals is controlled for by setting it to the highest level.

2.6.2 Simulation results on retirement savings adequacy

To investigate the extent to which co-movements between wealth and retirement expenditure goals tempered the negative effect of the crisis on retirement savings adequacy, we simulate retirement savings adequacy using the SUR estimates presented in Appendix 2.I. Figure 2.2 and table 2.6 summarize the simulation results (a more detailed description can be found in Appendix 2.J). We find that between 2008 and 2014 the fraction of individuals who do not accumulate a sufficiently generous pension entitlement to afford their self-reported retirement expenditure goal increased from 27% to 32%. Furthermore, the median difference between pension annuities and retirement expenditure goals decreased from 24% in 2008 to 20% in 2014. Hence, based on pensions alone the aggregate preparedness for retirement of the Dutch population declined only slightly during the period of the financial crisis and the subsequent recession. A similar picture of modest decline in preparedness emerges if we include discretionary wealth and/or housing wealth: the fraction for whom the annuity will fall short of their consumption goal increased by a similar amount and the median excess annuity declined by less than 5%-points. In particular, while 11% of the population was predicted to fall short of their retirement expenditure goal in 2008 even if they would draw down housing wealth, this fraction had risen to 17% by 2014.

In order to separate changes in goals and resources we simulate the fraction that would have failed to meet their expenditure goals had the relationship between goals and covariates remained the same in 2014 as it was in 2008 (so that goals are fixed for a given level of covariates). In this counterfactual scenario the fraction with insufficient resources to afford their retirement expenditure goals would have almost doubled from 27% to 50% (if we only take pensions into account). Adjusting goals reduced the fraction of insufficiently prepared by 18%-points. Based on all wealth components, co-movements between wealth and retirement expenditure goals mitigated the fraction falling short from about a quarter to 17%. So, the results show that co-movements between wealth and retirement expenditure goals mitigated the decline in retirement savings adequacy considerably.

Results are very similar if we do not control for question difficulty. In that case expenditure goals are slightly lower in both years so that the median difference and the fraction that falls short respectively increase and decrease with 3%-points across the board. Hence, our simulations are not driven by the adjustment of expenditure goals for question difficulty. Moreover, robustness checks with different indexation scenarios for occupational pensions in 2014 indicate that annuities are robust with regard to reasonable variation in the assumptions under which they are computed. Robustness checks of the simulations are available on request.

2.7 Conclusion

This chapter investigates co-movements between wealth and retirement expenditure goals using variation brought about by the Great Recession. These co-movements have important implications for retirement savings adequacy, and become increasingly important as the generosity of public pensions declines and people depend more on financial markets and housing wealth. We quantify co-movements and separate 'pure' wealth effects from common factors that influence both wealth and retirement expenditure goals. Furthermore, we examine how adjustments to expenditure goals mitigated the negative effect of the Great Recession on retirement savings adequacy, defined by the difference between individual retirement expenditure goals and annuitized wealth.

The setting of the Netherlands during the aftermath of the crisis is particularly interesting for this study, because it constituted an exogenous shock to a system that enrolls individuals into mandatory public and occupational pension schemes. Participants cannot choose their own pension fund, their contribution level, and their investment strategy. Hence, variation across funds in shocks to pension wealth, the most important source of income in retirement, is exogenous to workers. Moreover, house prices decreased by 20% on average between 2008 and 2013, eating into the most important category of discretionary wealth. This context of large and exogenous changes to wealth provides a unique opportunity to study the updating of expenditure goals.

For this study we match individual level administrative data on pension wealth, real estate and other forms of wealth with survey data on expenditure goals in retirement. Goals and resources are observed in 2008 and 2014. The combination of administrative data and surveys before and during the Great Recession is unique. However, since a limited number of individuals can be observed twice, some caution is needed when drawing conclusions.

The results show that between January 2008 and December 2014 both 'pure' wealth effects and common factors played a role in co-movements between wealth and retirement expenditure goals. At the level of the individual, we find suggestive evidence for heterogeneous effects of shocks to pensions and real estate wealth. Shocks to pensions exert the stronger effect overall, with a reduction in goals of 23-33 cents on average for a 1 euro decrease in the pension annuity. Moreover, the relative importance of shocks in wealth components varies with age: individuals younger than 50 adjusted goals more strongly after a shock to housing wealth, while the goals of older people were most affected by shocks to pensions. One possible explanation is that mental accounts change as people age. We do observe that the young are more likely to report a willingness to move and use their home to finance retirement if necessary than older individuals. Interestingly, while in the short run consumption of low income households is found to be more sensitive to wealth shocks (they have a relatively high marginal propensity to consume), we find that they adjust their retirement expenditure goals less after a wealth shock. Since low income households have relatively high essential spending, in the long run they may prefer to work more or retire later instead of adjusting their retirement expenditure goals downward. The fact that all effects of annuities disappear in log-log specifications suggests that only substantial changes to wealth induce updates of spending targets. We believe that these results warrant further attention.

Comparison of the two cross-sectional waves shows that in case people would not have adjusted their goals, the percentage falling short with respect to their own retirement expenditure goals would have risen from 11% in 2008 to 26% in 2014 if we take all wealth components into account.

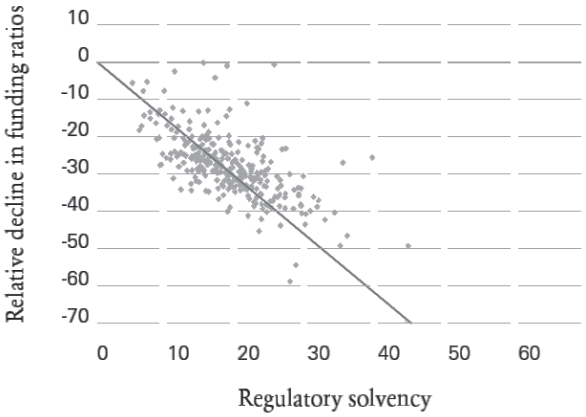
Instead, people adjusted their goals downwards and the fraction who was ill-prepared increased only to 17% (based on all wealth components). The results underline the importance of co-movements between wealth and retirement expenditure goals, and that a static benchmark for the assessment of savings sufficiency not only misses cross-sectional differences in preferences, but also cannot capture adjustments to a changing environment.

Funding ratios 2.A

The y-axis of figure 2.3 shows the relative decline in funding ratios of Dutch pension funds during 2008. The figure shows that there are vast differences across funds in the relative decline in funding ratios. In the first quarter of 2008 only 7 funds had a funding ratio below 105%, 256 funds had a funding ratio between 105% and 130% and 166 funds had a funding ratio above 130%. In the first quarter of 2009 the number of funds with a funding ratio below 105% increased to 314, 65 funds had a funding ratio between 105 – 130%, and only 20 funds had a funding ratio above 130%.

Pension funds with a low funding ratio were forced to draw up recovery plans in early 2009 in order to bring their funding ratios back to the required levels within five years. These plans ended in late 2013. DNB (2014) reports that funding ratios recovered primarily as a results of rising equity prices, but as interest rates fell further and life expectancy rose, the recovery remained relatively limited. All in all, about 25% of the original decline in funding ratios since the credit crisis was recovered at the end of 2013 (with vast differences between individual funds).

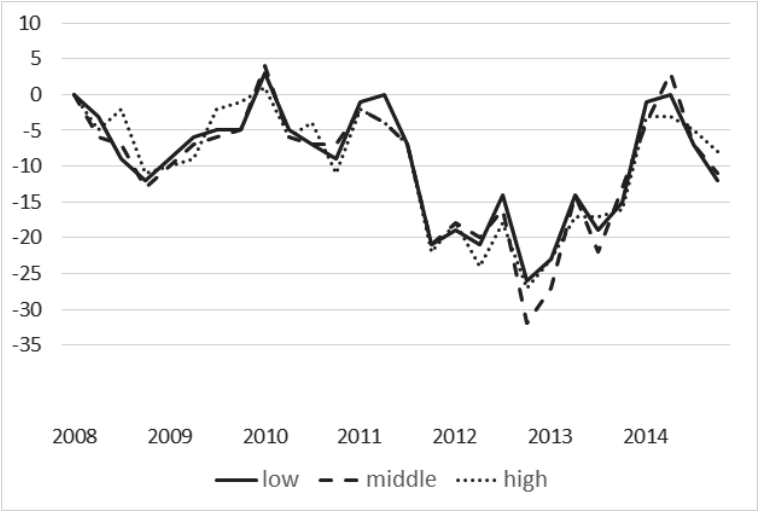
Figure 2.3: Relationship between regulatory solvency and relative decline in funding ratios during 2008, source: DNB (2009)



2.B Development consumer confidence

Figure 2.4 shows the development of a subquestion of consumer confidence, namely people's confidence in their financial situation in the next 12 months. The vertical axes shows the balance between positive and negative answers, normalized to 0 in the first quarter of 2008 for low, middle and high education levels. While the levels of confidence are higher for high education groups than for low education groups, the development is almost the same in both groups.

Figure 2.4: Development of people's confidence in their financial situation in the next 12 months, by education level



2.C Thinking about retirement and difficulty of the questions

Respondents find questions on expenditure goals during retirement challenging. This appendix first compares the distribution of retirement expenditure goals between retirees and non-retirees. Secondly, we provide descriptive statistics on the extent to which respondents have thought about retirement and how they evaluated the difficulty of the questions.

Comparison retirees and non-retirees

Table 2.7 shows that retirees reported higher expenditure goals than non-retirees across the distribution, especially in 2008. The mean and the first and third quartiles were 100-200 euros higher among retirees. Such differences cannot be explained by current incomes, as illustrated by replacement rates that were also around 10pp higher among retirees. However, the differences in levels are modest compared to the standard deviations in excess of 700 euros for both sub-samples. Differences were

Table 2.7: Descriptive statistics of minimum expenditures during retirement

		N	Mean	SD	p25	Mdn	p75
a. Non-retired							
2008	Min. monthly expenditures[a]	1142	1716	721	1218	1625	2031
	Min. replacement rate (%)[b]	1142	74	28	56	73	88
2014	Min. monthly expenditures[a]	1918	1471	567	1095	1460	1825
	Min. replacement rate (%)[b]	1891	67	31	46	63	81
b. Retired							
2008	Min. monthly expenditures[a]	254	1871	772	1335	1625	2226
	Min. replacement rate (%)[b]	254	83	26	67	81	98
2014	Min. monthly expenditures[a]	837	1549	576	1168	1460	1825
	Min. replacement rate (%)[b]	826	65	24	49	63	77

[a] Monthly retirement expenditure goals are standardized to a one-person household and denoted in 2014 euros.
[b] Replacement rate := monthly expenditure goal/current income

smaller in 2014: less than 100 euros or 5pp in replacement rates. Hence, though we do find that retirees had more ambitious goals than those not yet retired, the order of magnitude was the same for both groups. Furthermore, the variation within groups far exceeds that between groups.

Difficulty of the questions

Table 2.8 summarizes items that are related to perceived difficulty of the questions. These questions allow us to investigate whether those who do not understand the questionnaire give systematically different answers. When asked whether individuals find the question difficult to answer, in 2014 more individuals said they fully agree to the statement than in 2008. This holds especially for individuals under 54.

Table 2.9 summarizes the retirement expenditure goals by level of question difficulty. Individuals who find the question more difficult on average report lower retirement expenditure goals. However, when retirement expenditure goals are measured relative to current household income this is no longer the case, suggesting that question difficulty correlates with current household income.

Table 2.8: Descriptives of self-reported question difficulty

		Mean	Age 25-39	Age 40-54	Age 55+
I find it very difficult to imagine how much money I would want to have during retirement.					
2008	Fully disagree	0.07	0.04	0.06	0.15
	Somewhat disagree	0.09	0.07	0.10	0.11
	Somewhat agree	0.42	0.43	0.42	0.41
	Fully agree	0.41	0.47	0.41	0.33
	N	1610	502	728	380
2014	Fully disagree	0.05	0.02	0.03	0.11
	Somewhat disagree	0.09	0.06	0.08	0.12
	Somewhat agree	0.38	0.31	0.39	0.41
	Fully agree	0.48	0.60	0.51	0.36
	N	3272	851	1257	1164

Table 2.9: Descriptives of retirement expenditure goals by level of question difficulty

		N	Mean	Median	SD
I find it very difficult to imagine goals...					
a. Consumption goals: levels					
	...fully disagree	109	1913	1669	913
	...somewhat disagree	133	1918	1787	797
2008	...somewhat agree	530	1695	1625	677
	...fully agree	399	1639	1625	674
	...difficult to imagine: missing	266	1904	1669	856
	...fully disagree	139	1633	1460	735
	...somewhat disagree	212	1449	1430	518
2014	...somewhat agree	806	1482	1460	572
	...fully agree	804	1434	1400	532
	...difficult to imagine: missing	865	1546	1460	577
b. Consumption goals: replacement rates (in %, relative to current household income)					
	...fully disagree	109	74	75	22
	...somewhat disagree	133	80	77	37
2008	...somewhat agree	530	74	71	28
	...fully agree	399	71	71	27
	...difficult to imagine: missing	266	83	82	26
	...fully disagree	137	67	63	26
	...somewhat disagree	208	64	58	30
2014	...somewhat agree	790	67	63	31
	...fully agree	789	69	66	31
	...difficult to imagine: missing	852	65	63	24

2.D Descriptives socio-economic variables

Table 2.10: Descriptives of socio-economic variables: individual-level variables

	2008		2014	
	Mean	SD	Mean	SD
Single	0.16	0.37	0.29	0.45
Female	0.52	0.5	0.53	0.5
Age	49	13	53	15
HH head	0.59	0.49	0.64	0.48
Any kids	0.48	0.5	0.37	0.48
Number of kids	0.92	1.11	0.72	1.06
Homeowner	0.77	0.42	0.73	0.45
Education:				
- Primary	0.09	0.29	0.07	0.25
- Intermediate secondary	0.26	0.44	0.22	0.42
- Higher secondary	0.08	0.27	0.08	0.28
- Intermediate vocational	0.25	0.43	0.26	0.44
- Higher vocational	0.23	0.42	0.26	0.44
- University	0.08	0.28	0.12	0.32
Primary activity:				
- Salary worker	0.58	0.49	0.5	0.5
- Self-employed	0.08	0.28	0.07	0.25
- Family business	0.02	0.13	0.01	0.11
- ZZP	0.07	0.25	0.06	0.23
- HH work	0.12	0.32	0.08	0.27
- Retired	0.15	0.36	0.24	0.43
- Disabled	0.03	0.17	0.04	0.2
- Other	0.04	0.2	0.08	0.27
Marital status:				
- Married	0.71	0.45	0.59	0.49
- Separated/divorced	0.08	0.28	0.11	0.32
- Widowed	0.03	0.17	0.06	0.24
- Never married	0.18	0.38	0.23	0.42
Urbanization:				
- Extremely urban	0.12	0.33	0.15	0.36
- Very urban	0.27	0.44	0.26	0.44
- Moderately urban	0.22	0.42	0.23	0.42
- Slightly urban	0.23	0.42	0.21	0.41
- Not urban	0.15	0.36	0.14	0.35
Net personal income	2025	7812	1796	4580
Net household income	3528	6920	3052	4682
N	2308		5623	

Table 2.11: Descriptives of socio-economic variables: household-level variables

	2008		2014	
	Mean	SD	Mean	SD
Single	0.2	0.4	0.38	0.48
Female × single	0.12	0.32	0.21	0.41
Age HH head	50	13	53	16
Any kids	0.47	0.5	0.34	0.48
Number of kids	0.88	1.1	0.66	1.03
Homeowner	0.76	0.43	0.69	0.46
Education:				
- Primary	0.04	0.2	0.05	0.21
- Intermediate secondary	0.18	0.39	0.16	0.36
- Higher secondary	0.08	0.27	0.07	0.25
- Intermediate vocational	0.27	0.44	0.26	0.44
- Higher vocational	0.31	0.46	0.31	0.46
- University	0.12	0.33	0.16	0.37
Primary activity				
- 1 salary worker	0.7	0.46	0.6	0.49
- all salary workers	0.45	0.5	0.41	0.49
- 1 self-employed	0.13	0.34	0.11	0.31
- all self-employed	0.04	0.19	0.03	0.18
- 1 family business	0.02	0.16	0.02	0.13
- all family business	0.01	0.09	0.01	0.07
- 1 zzp	0.11	0.31	0.1	0.3
- all zzp	0.02	0.15	0.02	0.15
- 1 retired	0.2	0.4	0.28	0.45
- all retired	0.1	0.3	0.19	0.4
- 1 disabled	0.06	0.23	0.06	0.24
- all disabled	0.01	0.1	0.02	0.15
Marital status:				
- Married	0.68	0.47	0.52	0.5
- Separated/divorced	0.09	0.29	0.14	0.35
- Widowed	0.04	0.19	0.07	0.25
- Never married	0.19	0.39	0.27	0.44
Urbanization:				
- Extremely urban	0.13	0.33	0.17	0.38
- Very urban	0.27	0.44	0.26	0.44
- Moderately urban	0.22	0.42	0.22	0.42
- Slightly urban	0.23	0.42	0.2	0.4
- Not urban	0.15	0.36	0.14	0.35
Net HH income	3529	7604	2987	5423
N	1894		4098	

2.E Descriptive statistics assets and debts

Table 2.12 presents descriptive statistics of various categories of assets and debt. The most important types of assets in both years are saving accounts and owner-occupied real estate. On average saving accounts made up 27% of total assets in 2008 with a median value of 19.5 thousand euros. Residential real estate made up close to two thirds of the 2008 assets portfolio on average and the median value was 246 thousand euros. By 2014 the median value of residential real estate declined to 170 thousand euros and the average share in the assets portfolio declined to 58%. Consequently, the relative importance of saving accounts increased to 36% of the portfolio, despite a decrease in median savings to 14.2 thousand euros. Each of the other asset classes make up less than 5% of the asset portfolio in both years. As for debt, mortgage debt is by far the most important among the two types of debt that we observe: it accounts for 95% of total debt on average in both years. The median mortgage debt declined from 88 to 80 thousand euros between 2008 and 2014.

Table 2.12: Descriptive statistics of assets and debts

		% portfolio[a]	Mean	SD	p25	Mdn	p75
2008	Saving account	27	41.4	59.4	6.3	19.5	47.2
	Risky assets	4	24.3	133.0	0.0	0.0	5.9
	Residential real estate	65	247.2	222.8	104.7	246.0	335.6
	Non-residential real estate	3	17.0	82.7	0.0	0.0	0.0
	Business	1	2.4	33.7	0.0	0.0	0.0
	Other assets	0	2.0	20.9	0.0	0.0	0.0
	Mortgage debt	95	116.5	126.3	0.0	88.3	195.6
	Other debt	5	5.2	28.1	0.0	0.0	0.0
	N	890					
2014	Saving account	36	41.5	76.4	3.5	14.2	45.4
	Risky assets	3	22.9	169.7	0.0	0.0	0.1
	Residential real estate	58	168.5	149.0	0.0	170.2	242.7
	Non-residential real estate	2	16.7	86.3	0.0	0.0	0.0
	Business	1	4.4	63.6	0.0	0.0	0.0
	Other assets	0	4.4	80.0	0.0	0.0	0.0
	Mortgage debt	95	111.3	128.7	0.0	80.0	187.9
	Other debt	5	7.2	61.8	0.0	0.0	0.0
	N	3,429					

[a] Mean share of category in HH portfolio conditional on having non-negative total asset/debt.
Assets and debt in thousands of 2014 euros.

2.F Assumptions underlying the annuities

An annuity value is an estimated monthly income from pensions, savings, and housing at the date of retirement. In order to construct such annuities we need to make assumptions about the future. The future looked different in 2008 and 2014, so that in some cases the assumptions differ between those years. The scenario for the future from the perspective of 2014 was set up in correspondence with specialists at the Dutch Ministry of Social Affairs and Employment, the Ministry of the Interior and Kingdom Relations, the Ministry of Finance and the Netherlands Bureau for Economic Policy Analysis (CPB). In this section we explain the assumptions underlying the annuity values. Moreover, we describe how we updated the private pension data to include policy changes introduced in 2013.

Life course

The level of a public pension depends on the number of years someone lived in the Netherlands between the ages of 25 and 67, and on one's marital status during retirement. We observe the number of years individuals lived outside the Netherlands up to 2012 and assume that they will not leave the Netherlands from this moment onwards. Moreover, we assume that marital status stays the same. That is, we take into account marital status in our models, but we do not model future divorces, marriages, or widowhood. Lastly, we assume that individuals stay in the same job until they reach the statutory retirement age. That is, individuals who are unemployed remain unemployed and individuals who are employed will not become unemployed, will have constant wages, and will not retire early.

Statutory retirement age and private pension target age

In 2008 the age at which one could claim public pensions was 65, which was also the target age for defined benefit calculations in private pension plans. At that point there was no indication that this would change in the

future (Goudswaard 2011). We thus assume a retirement age of 65 when calculating the 2008 annuities.

The situation was completely different in 2014. In 2012 a law had been passed ensuring an increase in the statutory retirement age[19] and in 2014 an amendment was proposed that accelerated the process. When calculating 2014 annuities we assume the situation as in the amendment: a stepwise increase of the statutory retirement age to 67 in 2021, after which it will raise in accordance with life expectancy. The target age for the defined benefit calculations in private pension plans was set at 67 for the part of the claim built up after 2012. The increase of the target age went hand in hand with a lowering of the maximum of tax advanced yearly accrual rates. We assume that in the future any further increases in the target age will be accompanied with lower accrual rates, such that the pension level remains roughly unchanged.

Inflation and indexation

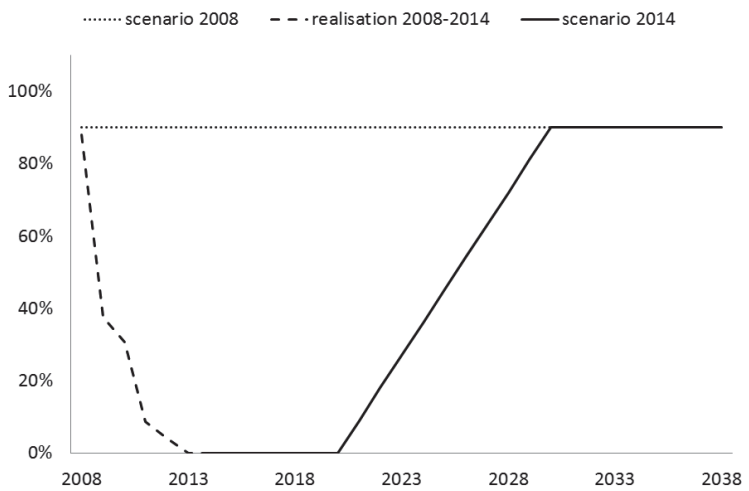
We assume an inflation of 2% each year. Both for the 2008 and the 2014 annuities, we assume the level of public pension benefits to be fully adjusted for inflation.

In 2008, 90% of occupational pension wealth was adjusted for inflation. During that time the financial position of private pension funds seemed perfectly in order, and in January 2008 people were optimistic about their future pensions. For the calculation of 2008 annuities we assume the situation remains unchanged and all pension entitlements are adjusted for inflation by 90%.

In the past years, however, occupational pension wealth has rarely been adjusted for inflation, so the value of pension wealth has declined in real terms. For the 2014 annuities we assume that pension funds will not adjust pension entitlements for inflation until 2020, after which indexation will rise gradually to 90% in 2030 and the years after. The 2008 and 2014 expected indexation patterns and the realizations for the years 2008-2014 are shown in Figure 2.5.

[19] Wet verhoging AOW- en pensioenrichtleeftijd

Figure 2.5: Indexation scenario's and realisations, after Knoef et al. (2016b)



Development of private savings and housing wealth

We take into account the current level of private savings and assume a real yearly interest rate of 1% per year. Private savings are annuitized at the moment of retirement given the most recent mortality tables of the CBS and a real interest rate of 1%. The annuitization procedure is explained in detail in Knoef et al. (2016a).

We assume that real housing prices increase with 1% a year. For individuals with positive net housing wealth we assume that the net imputed rent (1%) is put in a savings account where it receives an annual interest of 1%. For individuals who have a mortgage we assume mortgage payments are made. As of 2013 only individuals holding a mortgage contract with a pay off scheme of at most thirty years can benefit from fiscal benefits. We therefore assume individuals born before 1968 will pay off 25% of the remaining mortgage debt, individuals born between 1968 and 1978 will pay 50%, and individuals born after 1878 will pay 75%. Housing wealth is annuitized at the moment of retirement, given the most

recent mortality tables of the CBS and a real interest rate of 1%, similar to private savings.

The third-pillar pensions (voluntary individual pension products) are not shown in the administrative data, since they are not subject to taxation until they are paid out. However, the LISS survey does provide information on wealth accumulated in these products. For individuals who are self-employed and have a third-pillar pension product according to the survey, we assume that they will contribute 1.875% of their gross wage until retirement, in line with the contributions of salary workers to their occupational pension plans.

Updating 2012 occupational pension data

The latest administrative data available on occupational pension entitlements dates from 2012. Between 2012 and the end of 2014 several policy changes have taken place that will affect pension entitlements. Furthermore, most pension funds have not been able to correct the DB entitlements for inflation, and some even cut entitlements.

The entitlement data consist of two elements: (1) the accrued rights; (2) the rights to be accrued assuming income remains unchanged. First, we correct the accrued rights for the absence of inflation adjustment between 2012 and 2014. Second, we decrease the accrued rights by an amount equal to the cuts made in the respondent's pension fund.[20] The administrative data contain information on the amount of pension rights, but not on the name of the pension fund. Therefore, we provided the survey respondents in 2014 a list with the biggest pension funds in the Netherlands and asked them indicate at which of those they had entitlements. Third, maximum pension contributions declined from 2.25% to 2.15% in 2014, and further to 1.875% in 2015. The total relative decline is 17%. We assume that the actual build up percentages decrease to the same extent for all pension funds, hence we decrease the rights to be accrued until retirement by 17%

[20]Five major pension funds needed to apply cuts to accrued rights, that is ABP (0.5% in 2013), PME (5.1% in 2013), PMT (6.3% in 2013), Tandarts(specialisten) (3.2% in 2012 and 2.2% in 2013), Tandtechniek (7.0% in 2013, 2.0% in 2014).

for all individuals. Finally, the target age used to calculate the DB income changed from 65 to 67 in 2013. Accrued pension rights are therefore adjusted to the new statutory retirement age, at a rate of 7.5% per year. Pension rights that are to be accumulated in the years until retirement are extrapolated to the new statutory retirement age.

Quantile models of changes in expenditure goals 2.G

Table 2.13: Quantiles of shocks to annuities and changes in expenditure goals – heterogeneity by age

	Dependent variable: Δ retirement expenditure goal (2014 Euros)				
	p30	p40	p50	p60	p70
a. Complete sample					
Δ pension	0.162	0.206**	0.287**	0.263**	0.254**
	(-0.042;0.467)	(0.026;0.475)	(0.053;0.465)	(0.068;0.491)	(0.095;0.536)
Δ real estate	0.126	0.052	0.008	0.059	0.104
	(-0.154;0.357)	(-0.114;0.306)	(-0.115;0.294)	(-0.095;0.267)	(-0.210;0.231)
Sample quantiles	-469	-324	-206	-125	18
N (total obs.)			307		
b. Age 25-49					
Δ pension	-0.016	0.123	0.194	0.322	0.168
	(-0.303;0.597)	(-0.180;0.599)	(-0.117;0.588)	(-0.059;0.581)	(-0.058;0.574)
Δ real estate	0.235	0.199	0.150	0.146**	0.170**
	(-0.250;0.487)	(-0.102;0.501)	(-0.090;0.471)	(0.008;0.491)	(0.016;0.594)
Sample quantiles	-430	-311	-206	-82	120
N (total obs.)			129		
c. Age 50+					
Δ pension	0.284	0.272**	0.262**	0.272**	0.298**
	(-0.011;0.601)	(0.043;0.544)	(0.031;0.528)	(0.049;0.547)	(0.056;0.685)
Δ real estate	0.154	0.031	-0.010	0.001	-0.185
	(-0.212;0.400)	(-0.253;0.360)	(-0.267;0.324)	(-0.343;0.282)	(-0.431;0.199)
Sample quantiles	-494	-326	-204	-139	-18
N (total obs.)			178		

The models also control for the individual-level covariates listed in Appendix 2.D (with the exception of gender, age, education and degree of urbanization, since those variables display little or no variation within individuals). Annuities and expenditures are standardized to a one-person household. Bootstrapped 95% confidence intervals in parentheses; ** significant at 5%.

Table 2.14: Quantiles of shocks to annuities and changes in expenditure goals – heterogeneity by income

	Dependent variable: Δ retirement expenditure goal (2014 Euros)				
	p30	p40	p50	p60	p70
a. Low net household income in 2008					
Δ pension	0.074	0.121	0.079	0.074	0.123
	(-0.258;0.335)	(-0.212;0.333)	(-0.169;0.370)	(-0.152;0.392)	(-0.205;0.395)
Δ real estate	0.029	-0.024	0.041	-0.014	0.074
	(-0.473;0.537)	(-0.442;0.472)	(-0.368;0.416)	(-0.456;0.335)	(-0.407;0.400)
Sample quantiles	-351	-261	-165	-81	83
N (total obs.)			149		
b. High net household income in 2008					
Δ pension	0.341	0.234	0.334	0.381**	0.471**
	(-0.071;0.627)	(-0.080;0.686)	(-0.006;0.680)	(0.031;0.662)	(0.037;0.763)
Δ real estate	0.119	0.088	0.078	0.158	0.188
	(-0.320;0.358)	(-0.275;0.325)	(-0.272;0.294)	(-0.280;0.296)	(-0.251;0.271)
Sample quantiles	-603	-393	-209	-165	-18
N (total obs.)			158		

The models also control for the individual-level covariates listed in Appendix 2.D (with the exception of gender, age, education and degree of urbanization, since those variables display little or no variation within individuals). Annuities and expenditures are standardized to a one-person household. Bootstrapped 95% confidence intervals in parentheses; ** significant at 5%.

2.H Models of changes in expenditure goals that only control for family composition

Table 2.15: Shocks to annuities and changes in expenditure goals

	Dependent variable: Δ retirement expenditure goal			
	(1) OLS	(2) OLS	(3) 2SLS	(4) 2SLS
Δ Pension (β_1)	0.329**	0.235**	0.329***	0.101
	(0.129)	(0.0930)	(0.128)	(0.0969)
Δ Real estate (β_2)	0.0618	0.139	0.0836	0.00449
	(0.105)	(0.114)	(0.111)	(0.0149)
Constant (β_0)	-87.4	-111.2***	-83.1	-0.127***
	(63.9)	(41.4)	(62.1)	(0.0272)
Wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, n-1)	–	–	104.9***	20.0***
Endogeneity Δ real estate F(1, n-1)	–	–	0.17	1.72
n (number HHs)	282	282	282	272
N (total obs.)	307	307	307	296

[a] Δ Annuities and Δ expenditures are winsorized at p5 and p95. Winsorizing at p1-p99, or p2.5-p97.5 leads to similar results.
[b] This column regresses Δ log (goals) on Δ log (annuities).
The models also control for living with a partner and the number of children in the household. Annuities and expenditures are standardized to a one-person household. Standard errors clustered at the household level, in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2.16: Shocks to annuities and changes in expenditure goals – heterogeneity by age

	Dependent variable: Δ retirement expenditure goal			
	(1) OLS	(2) OLS	(3) 2SLS	(4) 2SLS
a. Age 25-49[c]				
Δ Pension (β_1)	0.144	0.140	0.141	0.102
	(0.123)	(0.142)	(0.124)	(0.166)
Δ Real estate (β_2)	0.154*	0.182	0.113	-0.00580
	(0.0832)	(0.191)	(0.104)	(0.0138)
Constant (β_0)	-110.6*	-104.7	-115.8*	-0.106**
	(66.0)	(68.0)	(66.0)	(0.0441)
Wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, 117)	–	–	34.2***	1174.2***
Endogeneity Δ real estate F(1, 117)	–	–	0.35	3.02*
n (number HHs)	118	118	118	118
N (total obs.)	129	129	129	129
b. Age 50+[c]				
Δ Pension (β_1)	0.400**	0.304**	0.398**	0.102
	(0.170)	(0.119)	(0.170)	(0.114)
Δ Real estate (β_2)	-0.00845	0.0857	0.0581	0.0529
	(0.155)	(0.139)	(0.157)	(0.106)
Constant (β_0)	-115.8	-134.3***	-99.0	-0.138***
	(91.1)	(51.9)	(89.9)	(0.0427)
Wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, n-1)	–	–	94.2***	5.49**
Endogeneity Δ real estate F(1, n-1)	–	–	0.81	0.00
n (number HHs)	168	168	168	158
N (total obs.)	178	178	178	167
c. Difference between ages 25-49 and 50+				
(H_0: equal coefficients; statistics follow $\chi^2(1)$ distribution)				
Δ Pension (β_1)	1.49	0.78	–	–
Δ Real estate (β_2)	0.85	0.17	–	–
Constant (β_0)	0.00	0.12	–	–

[a] Δ Annuities and Δ expenditures are winsorized at p5 and p95. Winsorizing at p1-p99, or p2.5-p97.5 leads to similar results.
[b] This column regresses $\Delta \log(\text{goals})$ on $\Delta \log(\text{annuities})$.
[c] OLS models on age sub-samples are estimated jointly.
The models also control for living with a partner and the number of children in the household. Annuities and expenditures are standardized to a one-person household. Standard errors clustered at the household level, in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2.17: Shocks to annuities and changes in expenditure goals– heterogeneity by income

	Dependent variable: Δ retirement expenditure goal			
	(1) OLS	(2) OLS	(3) 2SLS	(4) 2SLS
a. Low net household income in 2008[c]				
Δ pension	0.0496	0.0674	0.0509	0.123
	(0.0840)	(0.125)	(0.0837)	(0.129)
Δ real estate	0.0153	0.0779	-0.0109	-0.00161
	(0.0980)	(0.186)	(0.169)	(0.0172)
Constant	-125.2***	-121.2***	-128.3***	-0.108***
	(43.8)	(45.7)	(44.9)	(0.0344)
wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, n-1)	–	–	39.9***	21.8***
Endogeneity Δ real estate F(1, n-1)	–	–	0.059	1.86
n (number HHs)	137	137	137	134
N (total obs.)	149	149	149	146
b. High net household income in 2008[c]				
Δ pension	0.434**	0.297**	0.435**	0.107
	(0.184)	(0.137)	(0.184)	(0.133)
Δ real estate	0.0729	0.118	0.0904	0.124
	(0.133)	(0.145)	(0.139)	(0.0949)
Constant	-61.9	-120.1	-56.9	-0.104**
	(119.6)	(79.6)	(120.5)	(0.0498)
wealth expressed as	annuity	winsorized annuity[a]	annuity	log(annuity)[b]
First stage F(1, n-1)	–	–	90.9***	120.4***
Endogeneity Δ real estate F(1, n-1)	–	–	0.080	0.91
n (number HHs)	145	145	145	138
N (total obs.)	158	158	158	150
c. Difference between low and high income groups				
(H_0: equal coefficients; statistics follow χ^2 (1) distribution)				
Δ pension	3.61*	1.53	–	–
Δ real estate	0.12	0.03	–	–
Constant	0.25	0.00	–	–

[a] Δ annuities and Δ expenditures are winsorized at p5 and p95. Winsorizing at p1-p99, or p2.5-p97.5 leads to similar results.

[b] This column regresses Δ log (goals) on Δ log (annuities).

[c] OLS models on income sub-samples are estimated jointly. Cutoff between low and high income group is chosen to include about half of the respondents in each group. The models also control for living with a partner and the number of children in the household. Annuities and expenditures are standardized to a one-person household. Standard errors clustered at the household level, in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

2.I Estimation results SUR model

Retirement expenditures equations

Tables 2.18 and 2.19 present estimation results of the expenditure equations (2.8) and (2.9). The coefficients for 2008 show that homeowners, highly educated men and women, self-employed men, and separated/divorced men had relatively high retirement expenditure goals. Widowers reported 19% lower retirement expenditure goals while widows required 17% higher expenditures relative to married couples. Furthermore, household income plays a significant role in explaining retirement expenditure goals, with an elasticity of 0.48 for both men and women.[21]

We observe interesting changes between the coefficients of 2008 and 2014. Homeowners reported 6-9% higher retirement expenditure goals than renters in 2008, but that difference disappeared by 2014 (in line with the decline in house prices). The income elasticity of the retirement expenditure goals dropped from 0.48 to 0.35, and highly educated women reduced their retirement expenditure goals. Finally, self-employed men, who had relatively high retirement expenditure goals in 2008, did not have these relatively high goals anymore in 2014.

[21] The estimates for the retirement expenditure goals for 2008 are mostly similar to those documented by De Bresser and Knoef (2015). The largest difference is a stronger relationship between retirement goals reported by men and household income: our estimates imply that a 10% increase in the income of the husband increases his expected annuity by 4.8%, compared with 3.3% according to De Bresser and Knoef (2015). Moreover, this correlation is similar for the income of his wife, so that household income is an important covariate of expenditure goals of both men and women regardless of who brings it in. Though the differences in average reported retirement expenditure goals between education groups are smaller than in the earlier paper, they remain large and highly statistically significant with university graduates reporting 27-28% higher goals than those with no education beyond primary school.

Table 2.18: Joint models of annuities and retirement expenditures – expenditure equations – men

		2008		2014 − 2008	
	Partner	-0.023	(0.0425)	-0.117**	(0.0509)
	Age/10	-0.011	(0.0143)	0.016	(0.0177)
	HH head	0.008	(0.0470)	0.037	(0.0599)
	Any Children	-0.063	(0.0518)	-0.019	(0.0665)
	Number Children	0.009	(0.0233)	0.018	(0.0301)
	Homeowner	0.059*	(0.0309)	-0.055	(0.0376)
	log pers. Income	0.010	(0.0167)	-0.039*	(0.0216)
	Log HH income	0.482***	(0.0372)	-0.130***	(0.0454)
	Has simPC	-0.013	(0.0637)	-0.065	(0.0757)
Education[a]	Inter. secondary	0.026	(0.0469)	0.012	(0.0618)
	Higher secondary	0.140**	(0.0600)	-0.039	(0.0752)
	Inter. vocational	0.113**	(0.0463)	-0.053	(0.0612)
	Higher vocational	0.132***	(0.0464)	-0.005	(0.0612)
	University	0.277***	(0.0539)	-0.094	(0.0696)
Labor market status[a]	Family business	-0.079	(0.1006)	-0.024	(0.1361)
	Self-employed	0.147***	(0.0448)	-0.131**	(0.0570)
	Home maker	0.153	(0.1535)	0.135	(0.2097)
	Retired	0.145	(0.1493)	-0.112	(0.2256)
	Disabled	0.046	(0.0728)	-0.027	(0.0901)
	Other primary act.	0.063	(0.0742)	-0.069	(0.0856)
Marital status[a]	Separated/divorced	0.105**	(0.0519)	-0.087	(0.0617)
	Widow	-0.186**	(0.0939)	0.153	(0.1034)
	Never married	-0.001	(0.0391)	0.010	(0.0489)
Thought about retirement[a b]	Thought some	-0.052	(0.0513)	0.057	(0.0736)
	Thought a little	-0.031	(0.0516)	0.008	(0.0734)
	Hardly thought	-0.031	(0.0622)	0.019	(0.0839)
	No answer	-0.233	(0.2089)	0.011	(0.2946)
Urbanization[a]	Extremely urban	0.082*	(0.0423)	-0.028	(0.0510)
	Very urban	0.068**	(0.0323)	-0.047	(0.0392)
	Slightly urban	0.054*	(0.0324)	-0.038	(0.0393)
	Not urban	0.018	(0.0382)	-0.042	(0.0464)
Difficult to imagine spending[a b]	Somewhat disagree	0.069	(0.5469)	-0.165**	(0.0753)
	Somewhat agree	0.003	(0.0473)	-0.110*	(0.0666)
	Totally agree	-0.065	(0.0501)	-0.025	(0.0697)
	No answer	0.235	(0.1449)	-0.263	(0.1909)
	Constant	3.426***	(0.2980)	1.260***	(0.3630)
	Sigma epsilon	0.309***	(0.0081)		
	Log likelihood	-1310.22			
	N	4,521			

[a] The reference categories are *primary education*; *salary worker*; *married*; *thought a lot about retirement*; *moderately urban*; and *totally disagree*.

[b] The full questions read respectively "How much have you though about retirement?" and "I find it difficult to imagine how much I need to spend in retirement."

Dependent variables are logs of monthly retirement expenditure goals. Expenditures standardized to a one-person household; equations reported from models of annuity excluding housing wealth but including other savings. Standard errors in parentheses. *significant at 10%; **significant at 5%; ***significant at 1%.

Table 2.19: Joint models of annuities and retirement expenditures – expenditure equations – women

		2008		2014 − 2008	
	Partner	-0.041	(0.0528)	-0.131**	(0.0653)
	Age/10	0.050***	(0.0152)	-0.032*	(0.0180)
	HH head	-0.009	(0.0447)	-0.020	(0.0550)
	Any Children	-0.052	(0.0479)	0.072	(0.0620)
	Number Children	-0.001	(0.0219)	-0.003	(0.0291)
	Homeowner	0.091***	(0.0303)	-0.096***	(0.0369)
	log pers. Income	-0.003	(0.0067)	0.000	(0.0082)
	Log HH income	0.478***	(0.0374)	-0.131***	(0.0444)
	Has simPC	-0.056	(0.0615)	0.017	(0.0703)
Education[a]	Inter. secondary	0.045	(0.0464)	-0.071	(0.0582)
	Higher secondary	0.173***	(0.0585)	-0.150**	(0.0704)
	Inter. vocational	0.160***	(0.0505)	-0.153**	(0.0626)
	Higher vocational	0.181***	(0.0503)	-0.132**	(0.0622)
	University	0.275***	(0.0654)	-0.162**	(0.0790)
Labor market status[a]	Family business	0.106	(0.0964)	-0.051	(0.1305)
	Self-employed	-0.036	(0.0542)	0.007	(0.0687)
	Home maker	-0.026	(0.0418)	0.004	(0.0530)
	Retired	-0.009	(0.1861)	0.101	(0.2477)
	Disabled	0.028	(0.0727)	-0.023	(0.0828)
	Other primary act.	-0.029	(0.0607)	0.041	(0.0693)
Marital status[a]	Separated/divorced	0.057	(0.0498)	-0.072	(0.0613)
	Widow	0.167**	(0.0786)	-0.240***	(0.0889)
	Never married	0.079*	(0.0418)	-0.097*	(0.051)
Thought about retirement[a b]	Thought some	0.051	(0.0623)	-0.065	(0.0788)
	Thought a little	0.034	(0.0598)	-0.074	(0.0753)
	Hardly thought	0.037	(0.0657)	-0.029	(0.0818)
	No answer	0.128	(0.2887)	-0.201	(0.3481)
Urbanization[a]	Extremely urban	-0.066	(0.0443)	0.111**	(0.0518)
	Very urban	0.018	(0.0339)	0.036	(0.0404)
	Slightly urban	0.014	(0.0351)	0.018	(0.0421)
	Not urban	-0.041	(0.0400)	0.009	(0.0475)
Difficult to imagine spending[a b]	Somewhat disagree	-0.044	(0.0614)	0.016	(0.0779)
	Somewhat agree	-0.071	(0.0493)	0.082	(0.0633)
	Totally agree	-0.083*	(0.0505)	0.050	(0.0643)
	No answer	-0.152	(0.2163)	0.171	(0.2392)
	Constant	3.321***	(0.3140)	1.245***	(0.3713)
	Sigma epsilon	0.312***	(0.0086)		
	Log likelihood	-1310.22			
	N	4,521			

[a] The reference categories are *primary education*; *salary worker*; *married*; *thought a lot about retirement*; *moderately urban*; and *totally disagree*.

[b] The full questions read respectively "How much have you though about retirement?" and "I find it difficult to imagine how much I need to spend in retirement."

Dependent variables are logs of monthly retirement expenditure goals. Expenditures standardized to a one-person household; equations reported from models of annuity excluding housing wealth but including other savings. Standard errors in parentheses. *significant at 10%; **significant at 5%; ***significant at 1%.

Annuity equations

Tables 2.20 and 2.21 present estimation results of the annuity equation (2.10), both for annuities from public and occupational pensions and for annuities from total wealth (including real estate). The estimates for 2008 show that annuities from pensions were relatively high for homeowners, for households with highly educated heads, and for households that contain at least one salary worker. On the other hand, those annuities were relatively low on average for single females, for households with a family business, and for households with self-employed or a disabled household member. Furthermore, we estimate the elasticity of annuities with respect to net household income at 0.3. Taking into account wealth outside pensions changes some patterns: single men, but not women, now do better than couples and home-ownership plays a much more prominent role.[22]

Comparing the estimated coefficients for 2008 and 2014 in tables 2.20 and 2.21 we find interesting differences. Strikingly, the age gradient of pension annuities switched from negative to positive. While the average annuity from pensions in 2008 *de*creased with 1.5 percent for a 10 year increase in age, in 2014 this was associated with a 2.0 percent *in*crease in the average annuity. Moreover, the income elasticity of pension annuities decreased from 0.3 to 0.2 (in table I1 we saw that this was mirrored by a lower income elasticity of retirement expenditure goals) and the gaps between households with and without wage workers and with and without self-employed adults narrowed somewhat (as self-employed men

[22]The estimates for 2008 are mostly similar to those reported in De Bresser and Knoef (2015). The only exceptions are the estimated coefficients on household income and on the education dummies. Our estimates of the elasticity of the annuities with respect to net household income are around 0.3, while De Bresser and Knoef (2015) report smaller estimates around 0.1. This difference stems from the use of another survey variable for household income: the variable we use has been augmented with imputations and responses to unfolding bracket questions, while the earlier paper used a less streamlined income measure. This choice for a different income variable also reduces the differences in annuities between university graduates and the lowest education group from 33-45% to 24-27%, which confirms the interpretation that the large differences reported in that paper partly reflect measurement error in income (De Bresser and Knoef 2015). All other estimates for the annuity equations in 2008 are qualitatively and quantitatively similar to those reported in the earlier paper.

also reduced their retirement expenditure goals, their relative position compared to wage workers improved).

All these changes can be explained by the worsened situation of occupational pensions, which are relatively more important for high income earners. As a result, pension cuts affect high earners disproportionately and this flattens the association between income and annuities. The relative positions of old and young individuals, and of wage workers and the self-employed are aligned by the same mechanism (since occupational pensions typically play a minor role for the self-employed). Though the changes we observe can plausibly be attributed to changing circumstances, a change in the composition of socio-economic groups may also play a role.

We find broadly similar patterns when we take into account all private wealth. While in 2008 the annuity based on all wealth increased by 1.8% on average for a 10 year increase in age, in 2014 the corresponding figure was 4.1%. Similarly, the relationship between income and annuities flattened and the gap between households with and without salary workers closed. Unsurprisingly, the role of homeownership changed between 2008 and 2014 once we take into account housing wealth. The importance of housing in the household portfolio decreased as a result of lower house prices: the difference between the average annuity of homeowners compared to renters was 49% in 2008 and 45% in 2014.

Table 2.20: Joint models of annuities and retirement expenditure goals – annuity equations – pensions.

		2008		2014 - 2008	
	Single	-0.005	(0.0328)	0.044	(0.0353)
	Female × single	-0.079**	(0.0354)	0.038	(0.0372)
	Age HH head/10	-0.015*	(0.0085)	0.035***	(0.0094)
	Any kids	-0.104***	(0.0270)	0.077**	(0.0316)
	Number children	0.026**	(0.0115)	-0.030**	(0.0135)
	Homeowner	0.089***	(0.0180)	0.025	(0.0195)
	log HH income	0.304***	(0.0215)	-0.096***	(0.0238)
Education[a]	Inter. secondary	0.010	(0.0348)	0.038	(0.0394)
	Higher secondary	0.031	(0.0399)	0.023	(0.0451)
	Inter. vocational	0.075**	(0.0347)	0.015	(0.0391)
	Higher vocational	0.167***	(0.0349)	0.023	(0.0394)
	University	0.241***	(0.0399)	-0.036	(0.0446)
Labor market status	1 salary worker	0.119***	(0.0273)	-0.051*	(0.0308)
	All salary workers	0.056***	(0.0197)	0.023	(0.0235)
	1 family business	-0.053	(0.0612)	-0.047	(0.0707)
	All family business	-0.218**	(0.0980)	0.106	(0.1109)
	1 self employed	-0.147***	(0.0295)	0.071**	(0.0340)
	All self employed	-0.208***	(0.0555)	0.084	(0.0638)
	1 retired	0.037	(0.0329)	-0.036	(0.0373)
	All retired	0.051	(0.0329)	0.009	(0.0360)
	1 disabled	-0.076**	(0.0315)	0.030	(0.0369)
	All disabled	0.137*	(0.0793)	-0.110	(0.0861)
Marital status[a]	Separated/divorced	-0.022	(0.0445)	0.009	(0.0481)
	Female × sep/div	-0.063	(0.0512)	0.016	(0.0559)
	Widow	-0.021	(0.0455)	0.033	(0.0476)
	Never married	-0.051**	(0.0241)	0.048*	(0.0264)
Urbanization[a]	Extremely urban	-0.005	(0.0244)	0.053**	(0.0260)
	Very urban	0.023	(0.0253)	0.035	(0.0269)
	Slightly urban	0.013	(0.0253)	0.050*	(0.0270)
	Not urban	-0.01	(0.0281)	0.062**	(0.0301)
	Constant	5.123***	(0.1595)	0.273	(0.1776)
	Sigma epsilon	0.229***	(0.0056)		
	Log likelihood	-1310.33			
	N	4,521			

[a] The reference categories are *primary education*; *married*; and *moderately urban*.
Dependent variables are logs of monthly annuities. Annuities standardized to a one-person household. Standard errors in parentheses. *significant at 10%; **significant at 5%; ***significant at 1%.

Table 2.21: Joint models of annuities and retirement expenditure goals – annuity equations – pensions + wealth + housing.

		2008		2014 - 2008	
	Single	0.169***	(0.0378)	-0.003	(0.0414)
	Female × single	-0.143***	(0.0404)	0.068	(0.0428)
	Age HH head/10	0.018*	(0.0098)	0.023**	(0.0111)
	Any kids	-0.069**	(0.0309)	0.026	(0.0380)
	Number children	0.024*	(0.0132)	0.002	(0.0163)
	Homeowner	0.493***	(0.0208)	-0.040*	(0.0231)
	log HH income	0.333***	(0.0250)	-0.068**	(0.0283)
Education[a]	Inter. secondary	0.030	(0.0397)	0.015	(0.0465)
	Higher secondary	0.087*	(0.0458)	-0.040	(0.0537)
	Inter. vocational	0.074*	(0.0396)	0.027	(0.0461)
	Higher vocational	0.204***	(0.0399)	-0.015	(0.0464)
	University	0.271***	(0.0460)	0.021	(0.0527)
Labor market status	1 salary worker	0.058*	(0.0314)	-0.046	(0.0367)
	All salary workers	0.014	(0.0226)	0.051*	(0.0285)
	1 family business	0.098	(0.0742)	0.098	(0.0890)
	All family business	-0.208*	(0.1146)	-0.019	(0.1374)
	1 self employed	-0.154***	(0.0337)	0.186***	(0.0411)
	All self employed	-0.090	(0.0646)	0.000	(0.0782)
	1 retired	0.028	(0.0374)	-0.031	(0.0442)
	All retired	0.016	(0.0374)	0.049	(0.0419)
	1 disabled	-0.095**	(0.0369)	0.043	(0.0455)
	All disabled	0.091	(0.0901)	-0.117	(0.1011)
Marital status[a]	Separated/divorced	-0.026	(0.0514)	-0.055	(0.0570)
	Female × sep/div	-0.087	(0.0589)	0.102	(0.0664)
	Widow	-0.024	(0.0514)	0.129**	(0.0545)
	Never married	-0.003	(0.0282)	0.016	(0.0317)
Urbanization[a]	Extremely urban	0.057**	(0.0281)	-0.016	(0.0306)
	Very urban	0.083***	(0.0290)	0.012	(0.0315)
	Slightly urban	0.101***	(0.0291)	-0.002	(0.0318)
	Not urban	0.091***	(0.0323)	0.022	(0.0353)
	Constant	4.724***	(0.1852)	0.161	(0.2115)
	Sigma epsilon	0.273***	(0.0066)		
	Log likelihood	-2404.34			
	N	4,420			

[a] The reference categories are *primary education*; *married*; and *moderately urban*.
Dependent variables are logs of monthly annuities. Annuities standardized to a one-person household. Standard errors in parentheses. *significant at 10%; **significant at 5%; ***significant at 1%.

Error correlations

Table 2.22 reports the estimated correlations between the error terms for all equations of the SUR model. We find that the cross-sectional correlations between annuities and retirement expenditure goals are positive and significant in both years (0.17-0.22). Hence, individuals in households that can look forward to generous annuities conditional on their demographic characteristics, are also more ambitious regarding their retirement expenditure goals. The cross-sectional correlations between expenditure goals of partners within couples are even stronger, around 0.44-0.50, suggesting some agreement between partners on the retirement expenditure goal they should meet.[23]

As for correlations between the years we find that conditional on background characteristics annuities are relatively persistent, even in times of economic turbulence. The estimated correlations between the errors of the annuity equations in 2008 and 2014 are 0.56 and 0.65 for annuities based on pensions and on all wealth, respectively. Retirement expenditure goals are autocorrelated as well, but less strongly with estimated correlations around 0.36.

[23]For the revision of retirement expenditure goals between 2008 and 2014 we also find some agreement between partners, with a correlation of 0.5 conditional on observed characteristics.

Table 2.22: Error correlations

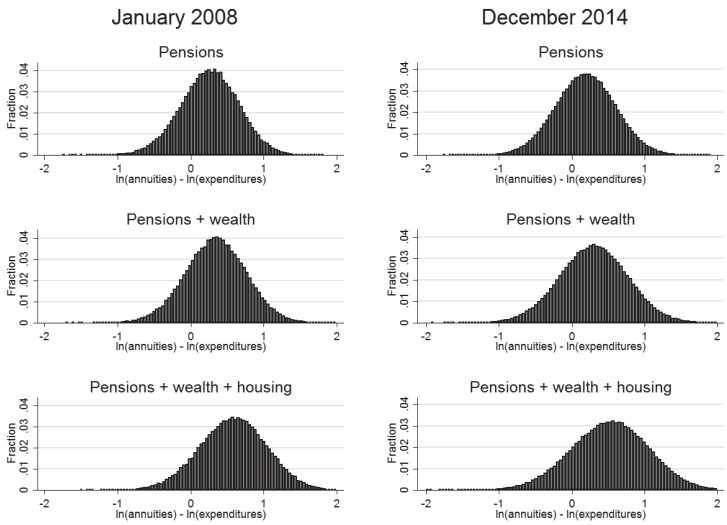
	Annuity 2008	Min exp. men 2008	Min exp. women 2008	Annuity 2014	Min exp. men 2014	Min exp. women 2014
a. Annuities from pensions						
Annuity 2008	1					
Min exp. men 2008	0.22***	1				
Min exp. women 2008	0.21***	0.44***	1			
Annuity 2014	0.56***	0.06	0.07	1		
Min exp. men 2014	0.14***	0.36***	0.03	0.15***	1	
Min exp. women 2014	0.07	0.22***	0.38***	0.15***	0.49***	1
b. Annuities from pensions and all wealth						
Annuity 2008	1					
Min exp. men 2008	0.17***	1				
Min exp. women 2008	0.19***	0.44***	1			
Annuity 2014	0.65***	0.09	0.09	1		
Min exp. men 2014	0.05	0.36***	0.06	0.12***	1	
Min exp. women 2014	0.08	0.24***	0.39***	0.12***	0.50***	1

**significant at 5%; ***significant at 1%

2.J Distribution of differences between goals and annuities

Figure 2.6 shows the simulated differences between retirement expenditure goals and annuities (both in logs and at the level of the individual). The differences subtract expenditure goals from annuities, so a positive difference means that the predicted annuity is sufficient to afford one's retirement expenditure goal and a negative difference implies insufficient funds. The graphs in the left column correspond to 2008 and those on the right to 2014, while different rows vary the scope of wealth from which annuities are computed. Comparing the columns, one notices that the locations of the distributions did not change much between 2008 and 2014. However, the spread increased slightly: the Great Recession increased inequality in retirement preparedness.

Figure 2.6: Simulated differences between annuities and expenditure goals



3 Health and Consumption Preferences - Estimating the Health State Dependence of Utility using Equivalence Scales

Abstract

This chapter estimates health state dependence of utility in Europe. For identification we introduce a new method using insights from the research domain of living standards. We estimate how much extra (or less) income is needed to maintain the same level of financial wellbeing after a health shock, and we derive a simple relation between this estimate and the health state dependence parameter. The results show positive health state dependence. This is not driven by medical expenditures, and is robust across alternative specifications and health measures. Interestingly, for cognitive limitations we find negative health state dependence, presumably resulting from a decreased ability to plan.

A working paper version of this chapter is published as Kools and Knoef (2017) and is currently under review. The chapter is co-authored by Marike Knoef. The authors thank the Network for Studies on Pensions, Aging and Retirement (Netspar) for financial support. Furthermore we thank Rob Alessie, Clementine Garrouste, Arie Kapteyn, Tabea Bucher-Koenen, Irene Ferrari, Luigi Pistaferri, Arthur van Soest, Joachim Winter, and participants of the Dutch Economist Day 2014, ESPE conference 2015, International Panel Data Conference 2015, Pension Workshop Paris 2015, Winter School on Inequality and Welfare 2016, Quantitative Society for Pensions and Saving Workshop 2016, EEA-ESEM 2016 and seminars at CPB, Netspar (2017) and the Munich Center for the Economics of Aging (2017).

3.1 Introduction

Assumptions about the degree and sign of health state dependence of the utility function, i.e. the change in the marginal utility of consumption with health status, have large implications for the optimal design of social security and long term care systems (Viscusi and Evans 1990, and Finkelstein et al. 2013). Health state dependence of utility influences the optimal level of life-cycle savings and health insurance. Health state dependence can also serve as an explanation for observed spending phenomena, such as the decreasing consumption path in old age (Börsch-Supan and Stahl 1991, and Domeij and Johannesson 2006).

Theoretically, health state dependence of utility could be positive just as well as negative. Some goods are valued more in bad health (so called complements to good health) and will raise marginal utility of consumption when ill, whereas other goods are less valuable in bad health (the substitutes to good health), thereby decreasing the marginal utility of consumption when ill. Examples of the first category are market services for physically demanding housework, like doing laundry, gardening, housecleaning, and cooking. Consumption of leisure activities is often placed under the second category (e.g. traveling may become less enjoyable in bad health). However, if individuals do not lose interest in leisure activities, but the activities become more costly due to the extra help or comfort required (e.g. travel assistance rather than solo traveling), those leisure activities actually fall under the first category. Whether marginal utility of consumption increases in bad health (positive health state dependence) or decreases (negative health state dependence), depends on the importance of both the complements and the substitutes and the importance attached to keeping up the pre-sickness lifestyle and activities when ill.[1] The size and sign of health state dependence may thus be determined by factors like socio-economic status and cultural background. Without

[1]For clarification, health state dependence is about the utility of nonmedical consumption. Changes in utility following a tightened budget constraint due to decreased income or increased medical expenditures are not captured by the concept of health state dependence.

empirical grounding, it is impossible to make assumptions on health state dependence.

Unfortunately, empirical work on the effects of health on consumption preferences provide ambiguous results. Research, mostly based on US data, shows evidence in favor of negative health state dependence of utility (Finkelstein et al. 2013), in favor of positive health state dependence (Lillard and Weiss 1997), and against the existence of health state dependence in either direction (De Nardi et al. 2010). The variation in outcomes may be attributed to the different methods used (Finkelstein et al. 2009). However, there may also be important heterogeneities in the effect, such that the choice of sample and health measure may be the source of the variety in results.

This chapter estimates health state dependence of utility in Europe using the Survey of Health, Aging and Retirement in Europe (SHARE). In order to do so, we build upon insights from the research domain of living standards and income adequacy (Pradhan and Ravallion 2000). We derive a 'health equivalence scale' and show that health state dependence is a transformation of this parameter.[2]

The results indicate positive health state dependence in Europe. We show that the findings are not driven by medical expenditures. Furthermore, the results are robust for different (physical) health measures and functional form assumptions. Among the robustness checks we find one interesting anomaly: cognitive limitations lead to negative health state dependence. When cognitive health declines, the willingness to undertake (leisure) activities may decline and this may lower expenditures. On the other hand, with physical health problems leisure activities may become more expensive because of the extra help required.

The contribution of this chapter to the literature is threefold. First, we introduce a new simple method for estimating the health state dependence of utility using questions widely available in survey data. Second, by

[2]We define the health equivalence scale as the relative change in income needed to maintain the same standard of living after a health shock. This equivalence scale is named after the common 'household equivalence scale', which measures the relative change in income needed to maintain the same standard of living with additional household members.

analyzing different measures of physical and cognitive health, we provide insight into the mechanisms underlying health state dependence. Third, to our knowledge we are the first to estimate health state dependence of utility for Europe. Differences in consumption patterns between US and Europe (Banks et al. 2015) may give rise to different sizes and even signs of health state dependence.

The rest of this chapter is set up as follows. Section 3.2 discusses the theoretical and empirical model underlying the analysis. Section 3.3 describes the data used, followed by the results of the empirical analysis in section 3.4. Finally, section 3.5 concludes.

3.2 Method

Many different methods have been developed to estimate health state dependence of the utility function, all with their own benefits and flaws. The contradictory results in the empirical work on the relationship between health and the marginal utility of consumption can in large part be attributed to differences between these methods. Finkelstein et al. (2009) distinguish two classes of methods to investigate health-state dependence. The first class exploits individuals' revealed demand for reallocating resources across health states. If there is some form of health state dependence of utility and individuals are forward-looking, they can be expected to already reallocate resources across health states before they fall sick, so that more can be consumed when marginal utility is highest. One could for example investigate health insurance demand or compare consumption paths across individuals who vary in their predicted probability of entering poor health (Lillard and Weiss 1997, and Butrica et al. 2009).

The second class of methods focuses on observed utility changes. By comparing within-individual utility changes associated with a health shock for poor and rich individuals, one can identify the change in the marginal utility of consumption due to a health shock. This can be done by using a direct proxy for utility, such as happiness (Finkelstein et al. 2013). Another way is to ask individuals how much money would be required to compensate them for hypothetical exposure to specific health

risks, and examine how these self-reported compensating differentials vary with income (Viscusi and Evans 1990, Evans and Viscusi 1991, and Sloan et al. 1998).

The method we propose builds upon the second class of methods. Similar to the second class of models, we make use of within individual comparisons associated with a health shock. However, rather than comparing overall utility changes at different income levels, we analyze average individual changes in financial wellbeing. In this way the method is less sensitive to bias stemming from unobserved characteristics correlated with income, that influence the effect of a health shock on overall utility. The intuition behind our method is as follows. Suppose an individual is asked: "are you able to make ends meet, yes or no?" and answers affirmative. In that case, his financial means must be above a personally set benchmark level. No suppose this individual falls ill, his financial means remain the same, but he now answers no to the posed question. Then his personally set benchmark level must have changed. We argue that, in case medical expenditures are covered by insurance, this change can only come from a shift in the marginal utility of consumption and thus the size of the average change in individual benchmark levels is sufficient to identify health state dependence of utility.

Section 3.2.1 explains the theoretical framework used to analyze the health state dependence parameter. Here, we also show the relation between the health state dependence parameter and the health equivalence scale. Section 3.2.2 explains how to estimate the health equivalence scale using data on financial wellbeing. Finally, the assumptions underlying our approach are laid out in section 3.2.3, together with the possible threats to identification.

Theoretical framework 3.2.1

Our theoretical model is strongly related to that of Finkelstein et al. (2013), who provide a thorough description of the theoretical framework under

which one can study health state dependence of the utility function.[3] Consider a retired individual and let S denote this individual's health status. For expositional purposes we define health to be binary: one is either healthy ($S = 0$) or sick ($S = 1$). An individual lives two periods. In the first period the individual is healthy. In the second period a negative health shock arises with probability p. The health shock itself is unanticipated, but the individual is aware of his chances to fall ill.

We assume that retired individuals derive utility $U(C(S), S)$ from consumption (C) and health (S). Health thus has a direct effect on overall utility (in general people do not like to be ill), but can also have an indirect effect on utility through consumption (which may be positive or negative). This is in accordance with Viscusi and Evans (1990) and Evans and Viscusi (1991), who explain that an adverse health shock may not only reduce utility, but can also alter the structure of the utility function (i.e., it may change the marginal utility of consumption). We are interested in this last part: how does health affect the marginal utility of consumption.

Consider the following standard intertemporal utility maximization problem of an individual

$$\begin{aligned} \max \quad & U = \frac{1}{1-\gamma} C_1^{1-\gamma} + \frac{1}{1+\delta} \left(-\phi_0 S + (1+\phi_1 S) \frac{1}{1-\gamma} C_2^{1-\gamma} \right) \\ \text{s.t.} \quad & Y = C_1 + \frac{1}{1+r} C_2 \end{aligned} \tag{3.1}$$

where C_t is consumption in period t, Y is lifetime income, γ denotes the coefficient of relative risk aversion, δ the discount rate and r the real interest rate. This chapter aims to estimate ϕ_1, the health state dependence parameter. Sickness decreases second period utility with ϕ_0 and multiplies

[3] For sake of simplicity we use a standard intertemporal utility function. Finkelstein et al. 2013 adopt a more general model. Comparing (3.6) and (3.7) in this chapter with equations (9) and (10) in Finkelstein et al. 2013, we find that this leads to the same indirect utility functions in case $b = 1$ and except for ϕ_0, which ends up in β_4 of equation (15) in Finkelstein et al. 2013. When using the more general model, w also includes a parameter for the elasticity of intertemporal substitution. In addition, Finkelstein et al. generalize the model by including health insurance, which covers a fraction b of second-period health expenditures. In the analysis, they select respondents with full health insurance. In the European countries under consideration in this study, most of the medical expenditures are covered by health insurances. Therefore, we assume $b = 1$. In section 3.4.2 we test whether this assumption is justified.

the marginal utility of second period consumption by a factor $(1+\phi_1)$. We assume $\gamma \geq 0$ and $\phi_1 \geq -1$. By rewriting the budget constraint, we obtain

$$C_1 = Y - \frac{C_2}{1+r}. \tag{3.2}$$

Health in period two is a random variable with probability of sickness p. Combining (3.1) and (3.2) gives us expected utility

$$\begin{aligned} E[U] = &\frac{1}{1-\gamma}\left(Y - \frac{C_2}{1+r}\right)^{1-\gamma} \\ &+ \frac{1}{1+\delta}\left(-\phi_0 p + \frac{1}{1-\gamma}(1+\phi_1 p)C_2^{1-\gamma}\right) \end{aligned} \tag{3.3}$$

which is maximized for

$$C_1^* = Y - \frac{C_2^*}{1+r} \tag{3.4}$$

where

$$C_2^* = \frac{((1+\phi_1 p)(1+r)/(1+\delta))^{1/\gamma}}{1+((1+\phi_1 p)(1+r)/(1+\delta))^{1/\gamma}/(1+r)} Y \equiv wY. \tag{3.5}$$

w is a proportionality factor, which is a function of the probability of sickness, the real interest rate, the discount rate, the coefficient of relative risk aversion, and the health state parameter ϕ_1. In the remainder of this chapter we denote Y to be permanent income. Since permanent income is a fraction of lifetime income, this only changes the definition of w. From (3.5) it follows that indirect second period utility in the healthy and sick state are as follows

$$V(Y, S=0) = \frac{1}{1-\gamma}(wY)^{1-\gamma}, \tag{3.6}$$

$$V(Y, S=1) = -\phi_0 + (1+\phi_1)\frac{1}{1-\gamma}(wY)^{1-\gamma}. \tag{3.7}$$

These two equations are equivalent to equations (9) and (10) in Finkelstein et al. (2013)[4] and show that ϕ_1 can be identified separately from δ in equation (3.1). The idea is that in period 1, before individuals know their future health status, individuals choose how much of permanent income Y to consume in the first period and how much to save for the second period. In the second period health is realized, and individuals experience utility $V(Y, S = 0)$ or $V(Y, S = 1)$, dependent on being in the healthy or the sick state.

We define μ to be the proportionality factor indicating how much extra (or less) income is needed in the sick state to be *financially* as well of as in the healthy state and give it the name 'health equivalence scale' after the more common household equivalence scales, which illustrate how needs change when household size increases. The value of μ is such that

$$V(\mu Y, S = 1) + \phi_0 = V(Y, S = 0). \tag{3.8}$$

That is, μ equates the indirect utility derived from income in the sick state and the healthy state and *does not* capture the direct effect of health on indirect utility. This is in correspondence with the construction of household equivalence scales, which do not take into account the utility derived from having a spouse or children, but only aim to capture economies of scale in a household.

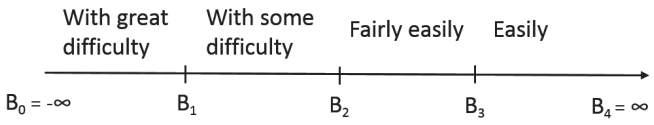
Combining equations (3.6), (3.7), and (3.8) we find that

$$\mu = (1 + \phi_1)^{-\frac{1}{1-\gamma}}. \tag{3.9}$$

The aim of our empirical analysis is to obtain an unbiased estimate of ϕ_1. If we find an unbiased estimate of μ, we can retrieve $\hat{\phi}_1$ using (3.9). To find an estimate of μ we rely on information about financial wellbeing, as we will explain in the following section.

[4] In case $b = 1$ and except for ϕ_0, as explained in footnote 3.

Figure 3.1: Benchmark levels making ends meet



Empirical model 3.2.2

Time periods in the empirical analysis can be thought of as repeated observations of an individual in period 2 of the theoretical model. To estimate the health equivalence scale μ we follow the reasoning of Pradhan and Ravallion (2000). For each individual i in period t we observe answers to a question on financial wellbeing (z_{it}) on a qualitative scale. Answers range from $z_{it} = 1$ to $z_{it} = K$, with higher values corresponding to higher levels of wellbeing.[5]

Financial wellbeing (making ends meet) depends on income and individual specific benchmark levels. Individual i reports financial wellbeing level k if his permanent income Y_i is at or above a certain benchmark B_{k-1}, but below B_k (see figure 3.1 for the situation where $K = 4$), with

$$\begin{aligned} \ln B_{k,it} &= \alpha_k + \rho \ln Y_i + \beta S_{it} + X_{it}\eta + \nu_i + \varepsilon_{it} \quad \text{for } k = 1, \ldots, K-1, \\ B_{0,it} &= -\infty, \text{ and } B_{K,it} = \infty, \end{aligned} \tag{3.10}$$

where S_{it} and Y_i are health status and permanent income, respectively. X_{it} is a vector of time constant and time varying variables of individual i in period t, ν_i an individual specific effect, and ε_{it} errors which we assume to be distributed as standard normal with mean zero and variance one independent of ν_i. Note that the individual benchmark levels $B_{k,it}$ depend on income Y_i. Just as Pradhan and Ravallion (2000), we follow the literature and assume a log-linear specification for the benchmark levels.

[5] Specifically, we observe whether individuals can make ends meet with great difficulty ($z_{it} = 1$), with some difficulty ($z_{it} = 2$), fairly easily ($z_{it} = 3$), or easily ($z_{it} = 4$).

The probability of observing outcome k is given by,

$$\begin{aligned}
Prob(z_{it} = k) &= Prob(B_{k-1,it} \leq Y_i < B_{kit}), && (3.11)\\
&= Prob(\ln B_{k-1it} \leq \ln Y_i < \ln B_{kit}), && (3.12)\\
&= Prob(\alpha_{k-1} + \rho \ln Y_i + \beta S_{it} + X_{it}\eta + \nu_i + \varepsilon_{it} && (3.13)\\
&\quad \leq \ln Y_i < \alpha_k + \rho \ln Y_i + \beta S_{it} + X_{it}\eta + \nu_i + \varepsilon_{it}),\\
&= Prob(-\alpha_k + (1-\rho) \ln Y_i - \beta S_{it} - X_{it}\eta - \nu_i && (3.14)\\
&\quad \leq \varepsilon_{it} < -\alpha_{k-1} + (1-\rho) \ln Y_i - \beta S_{it} - X_{it}\eta - \nu_i),\\
&= \Phi(-\alpha_{k-1} + (1-\rho) \ln Y_i - \beta S_{it} - X_{it}\eta - \nu_i) && (3.15)\\
&\quad - \Phi(-\alpha_k + (1-\rho) \ln Y_i - \beta S_{it} - X_{it}\eta - \nu_i),\\
&\quad k = 1, \ldots, K,
\end{aligned}$$

such that we estimate the following random effects ordered probit model

$$z^*_{it} = \theta \ln Y_i + \beta S_{it} + X_{it}\eta + \nu_i + \varepsilon_{it}, \tag{3.16}$$

where $\theta = -(1-\rho)$ and where observed ordinal responses z_{it} are generated from a latent continuous response such that

$$z_{it} = \begin{cases} 1 & \text{if } -\alpha_1 < z^*_{it}, \\ 2 & \text{if } -\alpha_2 < z^*_{it} \leq -\alpha_1, \\ \vdots & \\ K & \text{if } z^*_{it} \leq -\alpha_{K-1}. \end{cases} \tag{3.17}$$

To investigate within individual changes in health (and other time varying variables), we follow Mundlak (1978) and parameterize the individual specific effect as a linear function of the average time-varying explanatory variables over time, plus a random individual specific effect that is assumed to be independent of the explanatory variables,

$$\nu_i = \zeta_0 \bar{S}_i + \bar{X}_i \zeta_1 + \xi_i, \tag{3.18}$$

where $\bar{S}_i$ and $\bar{X}_i$ are the individual means of S_{it} and X_{it} respectively (the Mundlak terms) and ξ_i i.i.d. normal distributed with mean zero and

variance σ^2_ξ.[6] Equation (3.16) can now be rewritten as

$$z^*_{it} = \theta \ln Y_i + \beta S_{it} + X_{it}\eta + \zeta_0 \bar{S}_i + \bar{X}_i \zeta_1 + \xi_i + \varepsilon_{it}. \tag{3.19}$$

μ is defined as the proportionality factor indicating how much extra (or less) income is needed in the sick state to be financially as well off as in the healthy state. From (3.19) it follows that the extra income needed in the sick state to reach the same level of financial wellbeing as in the healthy state is

$$\mu = \exp\left(\frac{\beta}{-\theta}\right), \tag{3.20}$$

combining equations (3.9) and (3.20) shows that $\hat{\phi}_1$ can be consistently estimated by

$$\hat{\phi}_1 = \hat{\mu}^{\gamma-1} - 1 \tag{3.21}$$

$$= \left(\exp\left(\frac{\hat{\beta}}{-\hat{\theta}}\right)\right)^{\gamma-1} - 1, \tag{3.22}$$

where we need to fill in an appropriate value for the risk aversion parameter γ.

Underlying assumptions and threats to identification 3.2.3

To summarize, we identify health state dependence of utility from the effect of a health shock on within individual financial wellbeing. Compared

[6] Permanent income is constant across time. θ is thus identified by variation between individuals and we assume that there are no unobserved characteristics that influence both permanent income and financial wellbeing. If any, however, we would expect a positive correlation between permanent income and the individual unobserved effect. For example, someone has an expensive hobby, therefore he works relatively much and he receives a relatively high permanent income, but he is also demanding and therefore he is more inclined to struggle to make ends meet. In this case the estimated θ is higher than the true θ and this would bias our estimated health state dependence parameter towards zero, whether or not the true state dependence were positive or negative (so, the sign of the health state dependence parameter would not change because of this possible bias).

to other methods this has the advantage that it is sufficient to analyze average compensating differentials within individuals, rather than comparing compensating differentials across poor and rich individuals. In this way the method is less sensitive to bias stemming from unobserved characteristics correlated with income, that influence the effect of a health shock on overall utility. As an example one could think of the distance between individuals and their social network. High income people often live further away from their social network (e.g. a carpenter can find a job near to his family rather easily, whereas the university professor will have more difficulties in finding a job close to his family and will generally move greater distances during his life, which may also lead to friends being more geographically spread, Kalmijn 2006). This may lead high income individuals to suffer relatively more from a negative health shock, when they become physically less able to visit their social network because of the larger distances (e.g. they do not have the energy anymore to bridge large distances). In the second class of models, described in section 3.2, this may bias health state dependence. When using financial wellbeing one does not have to unravel the effect of health on consumption preferences from the effect of health on other aspects of life. This may lead to more precise estimates.

A number of assumptions, however, are still needed when using this approach. First, in the above model we assume that wealth in the sick state is predetermined. Our sample is therefore limited to retired individuals of age 65 and over, such that health shocks do not have a first order effect on income (as in Finkelstein et al. 2013). Threats, however, can occur because of changes in mortality risk, changes in out-of-pocket (OOP) medical expenditures, or when health changes are anticipated. If a health shock increases the (perceived) mortality risk, wealth per remaining year increases and in that way financial wellbeing may increase even if the marginal utility of consumption does not change, leading to a negative bias in the estimates of health state dependence[7]. We cannot control for this in our model, but we compared the results from the full sample to the results from a sample limited to those with more than 75% of wealth

[7]This bias is weakened under the existence of a bequest motive.

annuitized. The results are highly similar, which is as expected since wealth is relatively small compared to income (appendix A). OOP medical expenditures may bias us towards finding positive state dependence. We check for this in section 3.4.2 and find that OOP medical expenditures do not drive our results. In case health changes are anticipated, people who know that they will become sick save more (less) than they otherwise would have in case of positive (negative) health state dependence. Then, the actual health shock will not result in a lower or higher financial wellbeing, biasing our estimate of health state dependence towards zero (but the sign remains correct). Finally, health shocks may decrease home production and/or increase informal care, which can be considered as income in broad terms. In this chapter we consider a narrow income definition, such that the health state dependence parameter increases in case more domestic services and repairs need to be bought in bad health.[8]

Second, a disadvantage stemming from the use of subjective data is that of differential item functioning: different people interpret scales in different ways. An optimistic individual may be more inclined to use the top ends of the scale, whereas one who sees the glass half empty will give answers towards the bottom end of the scale (Ferrer-i Carbonell and Frijters 2004). We take differential item functioning into account by investigating the effect of within individual health transitions on the ability to make ends meet. However, answering styles are not necessarily fixed over time. Health shocks, for example, may change an individual's answering style with regard to financial wellbeing. Our data shows that individuals with limitations report negative feelings more often (such as sadness, self-blame, and irritability), which may influence answering styles. When we add negative and positive feelings as control variables in the model the conclusions do not change (Appendix A.3). However, by including positive and negative feelings in the model we introduce simultaneity bias, such that the cure could be worse than the disease.

[8]This is common in the literature. For example, Finkelstein et al. (2009) mention that the marginal utility of consumption could increase with deteriorating health, as prepared meals or assistance with self-care may be substitutes for good health. Furthermore, Banks et al. (2015), among others, classify domestic services and repairs as 'housing related' expenditures and not as medical spending.

Third, socioeconomic status and other third factors can make it difficult to establish causal relationships. Socioeconomic status (a combination of factors like education, income, and family background) may influence both health and financial wellbeing. As far as it concerns income and education, we control for these variables in the model. As far as it concerns factors such as family background, this is captured by following individuals over time and by analyzing variation in health within individuals. Similarly, if risk aversion is correlated with permanent income or the likelihood of a negative health shock this is controlled for by following individuals over time.

Fourth, people who struggle to make ends meet may experience more stress and this may affect their health negatively (reverse causality). The effect of stress on health is, however, probably more substantial in the long run than in the short run. Long term stress may lead, amongst other things, to a higher probability of cardiovascular diseases and thereby also to more problems with activities of daily living in old age. These long term differences between people, that probably arose already before retirement, are captured by the individual specific effects. Reverse causality may also arise when people who struggle to make ends meet are unable to pay for customary medical interventions. This may lead to a higher probability to encounter health problems. In the European countries under consideration customary medical interventions are generally paid for by health insurance or the government and we do not expect this reverse causality to have a big influence on our results. Finally, a problem may arise when health shocks lead to administrative help and making ends meet improves through this pathway. This would lead to a negative bias in our estimate of ϕ_1.[9]

Fifth, attrition may be systematically related to health. However, there is no reason to assume that the probability of leaving the panel after a health shock depends on the (change in) ability to make ends meet. We thus believe that the results will not be heavily affected by attrition bias.

[9]This is biasing against our finding of positive state-dependent utility, such that the true health state dependence parameter would even be more positive.

Data 3.3

To estimate the parameters of the model we use the Survey of Health, Aging, and Retirement in Europe (SHARE). SHARE is a multidisciplinary database on health, socioeconomic status and social and family networks of individuals aged 50 and over. Data are collected in 2004/2005, 2006/2007, 2008/2009, 2011/2012 and 2013 in several European countries. In 2004/2005 eleven European countries and Israel contributed data to the SHARE project. Over the years, eight more countries started to participate. Data were collected by face-to-face computer-aided personal interviews (CAPI), plus a self-completion drop-off part with questions that require more privacy.

The CAPI questionnaires of waves 1, 2, 4, and 5 are divided into eighteen modules of which the order remained roughly unchanged over the waves.[10] The interview is split in two parts, in between which some physical measurements are taken. The health module is in the first half of the interview, together with a module about demographics, and followed by a module on employment and pensions. Questions about informal care, income, wealth, and consumption take place in the second half of the questionnaire.

It is well known that questions asked earlier in a survey may influence how people respond to later questions. Fortunately, questions about health and about 'making ends meet' (in the consumption module) are in different parts of the questionnaire.

The remainder of this section describes the data used for the baseline regressions. We refer to appendix 3.A for more details and a description of the data used for the additional analyses.

Sample selection 3.3.1

For all households we select the 'household respondent', who answers the question on 'making ends meet'. Furthermore, our interest is in

[10] Wave 3 is a special wave which focuses on people's life histories and collects retrospective information.

Table 3.1: Summary statistics

	mean	sd. within	sd. between	min	max	N
Age	75.85	2.93	5.83	65	102	25827
Male	0.36	0	0.48	0	1	25827
Partner in household	0.41	0.19	0.45	0	1	25827
Highly educated	0.15	0	0.36	0	1	25827

Note: A respondent is considered highly educated with an ISCED level of five or higher.

individuals of age 65 and older, for whom the spouse (if present) is 65 or older, and for whom annual household income from a job or from self-employment is less than 2000 euro. In this way we drop those households for whom a health shock may lead to a substantial loss of income, due to a job loss or early retirement.

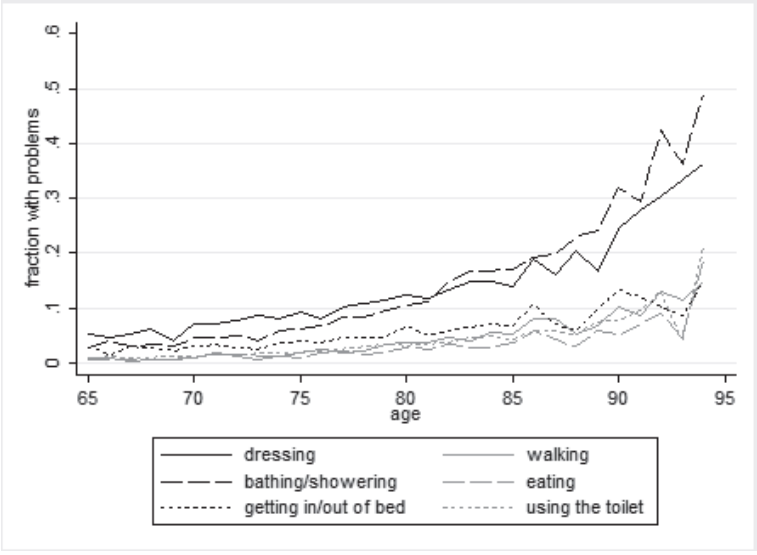
We select singles and 'household respondents' with a spouse and do not consider households with more than two people. SHARE only samples individuals living independently (i.e. not in a nursing home). However, if individuals make a transition into a nursing home over the course of the survey they remain in the sample. We drop individuals who are permanently living in a nursing home and we only consider individuals for whom data on two or more waves are available, so that we are able to measure transitions in health status. Because of this constraint we are left with fifteen countries, namely: Austria, Belgium, Switzerland, Germany, Denmark, Spain, France, Greece, Italy, the Netherlands, Sweden, Czech Republic, Poland, Estonia, and Slovenia.[11] Finally, we drop individuals in the top and bottom percentile of the income distribution, for each wave and country separately. The resulting data set contains 25,827 observations on 10,943 individuals. Table 3.1 shows descriptive statistics.

3.3.2 Health

SHARE includes many different measures of health, ranging from self-perceived health status to reported limitations and major health conditions,

[11] Israel is excluded because their surveys were conducted at different points in time than the rest and differ slightly from those of the other countries.

Figure 3.2: Prevalence of limitations in activities of daily living across age



health care usage, and physical performance measures. In this chapter our main health measure is limitations in activities of daily living (ADL), which are encountered by many in old age. Figure 3.2 plots the prevalence of the six types of limitations in ADL against age. The prevalence of limitations increases with age and the most common problems are those with dressing, bathing, and showering. We define individuals to be 'limited' when they have one or more problems with ADL. 19.5% of the individuals in our sample experience a health shock, measured by the presence of ADL limitations.

The use of ADL limitations has several advantages. First of all, by taking a measure of physical limitations, we focus on that aspect of health that is assumed to be a mediating factor between health and consumption. When considering a measure like the number of chronic diseases, the range of illnesses (from asthma to cancer) could make it hard to differentiate the possible implications of these illnesses to everyday life, hindering interpretation of the estimation results. Moreover, ADL limitations are relatively objective, in the sense that the domains of functionality are

narrowly focused and the interviewer can partly validate the answer by observing the respondent.[12] This is important since the variable we aim to explain (making ends meet) has a subjective component. Would the measure of limitations also be subjective (for example self-assessed health status), then correlated errors could bias the results.[13] We also conduct analyses using measures of IADL, chronic disease count, and cognitive functioning. More information on these health measures can be found in Appendix 3.A.1.

3.3.3 Income and assets

SHARE contains data on income, assets, and housing wealth. We construct an aggregated measure of household income by computing the sum of net household income and 5 percent of net financial assets (following Finkelstein et al. 2013).[14] In this way we account for the fact that elderly households may be spending down their accumulated financial savings. As a measure of permanent income we take the average of income over the different waves for each individual. All amounts are equivalized to a one person household using the OECD equivalence scale[15] and ppp-adjusted to 2004 German price levels.

For our selected sample average net household income equals 17,999 euro per year. Net household income is right skewed, with a median of 13,080 euro (substantially lower than the average). The average value of net financial assets equals 31,620 euro, with a median of 7914 euro. Permanent income is on average 19,225 euro, with a median of 15,092 euro. We control for the presence of positive net housing wealth and we add wave dummies to take into account possible effects of the Great Recession.

[12] Mete (2005) calls ADL limitations a quasi-objective health measure by using scare quotes ('objective').

[13] Responses to self-assessed health questions are not that robust. Using SHARE data, Lumsdaine and Exterkate (2013) show that self-assessed health suffers from question order and framing effects, which depend on observable characteristics such as health.

[14] We only use the income and wealth information from the observations included in the sample. If for example an individual is present in all waves, but only retired in wave 4, only the information from wave 4 and 5 is used to construct the permanent income measure.

[15] This means that income is divided by 1.5 for two person households.

Appendix 3.A.2 provides detailed information on income and assets per country.

Like all major household surveys, SHARE suffers from item non-response. In this chapter we make use of the imputations provided by SHARE for total net household income, net financial assets and net housing wealth. A multiple imputation technique has been used, which means that we have five different complete data sets that differ from one another only with respect to the imputed values. To capture uncertainty due to the imputation of missing values we perform the regressions on each dataset separately and then combine the results from all five datasets using the imputation method explained by Christelis (2011).[16]

Making ends meet 3.3.4

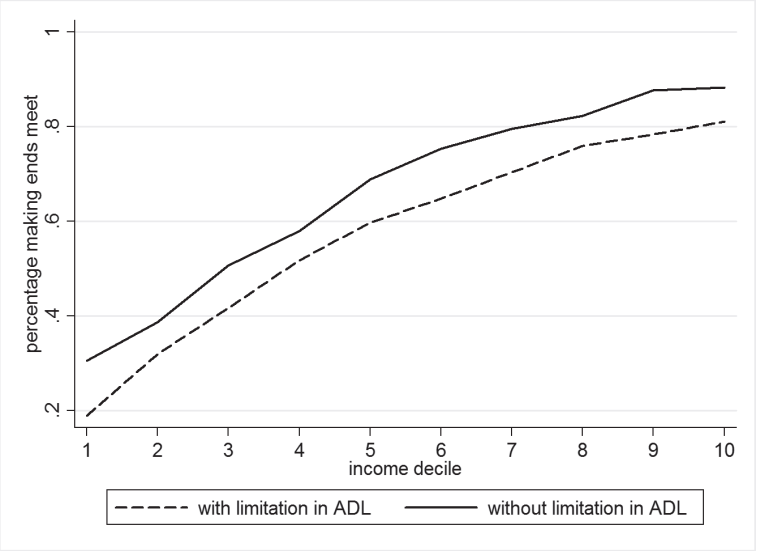
To measure financial wellbeing, respondents were asked the following question:

> Thinking of your household's total monthly income, would you say that your household is able to make ends meet...

Respondents can answer by choosing either one of the categories (1) with great difficulty, (2) with some difficulty, (3) fairly easily, or (4) easily. Figure 3.3 shows the fraction of individuals without difficulties to make ends meet (i.e. they answered fairly easily, or easily) across income percentiles for two groups: (1) those without any ADL limitations and (2) those with one or more ADL limitations. As expected, the ability to make ends meet increases with income. Moreover, conditional on income, individuals without physical limitations struggle less to make ends meet.

[16]The averages in this section are based on the first set of imputations.

Figure 3.3: 'Making ends meet' across health and income



3.4 Results

3.4.1 Baseline

Column (1) of table 3.2 presents the baseline estimation results. The first coefficient in panel A shows that after a health shock individuals report a lower ability to make ends meet. More income is required after a health shock for individuals to reach the same level of financial wellbeing as before. This is reflected by the health equivalence scale reported in panel B, which has a point estimate of 1.133. This implies that individuals on average need 13.3% more income after a health shock to be financially as well off as before. To calculate the corresponding health state dependence parameter we assume the risk aversion parameter γ to be 3, which is a reasonable value obtained by previous studies (e.g. Skinner 1985, and Palumbo 1999). We find positive health state dependence: the marginal utility of consumption is higher in bad health than in good health, with a proportionality factor of $\widehat{\phi}_1 = 0.284$.

Table 3.2: Baseline results

	(1) baseline	(2) excl. wave 4	(3) excl. wave 4 + OOP med. exp.
A. Estimation results			
limitation	-0.110***	-0.0813	-0.0719
	(0.0353)	(0.0564)	(0.0561)
$\ln(Y)$	0.880***	0.834***	0.844***
	(0.0323)	(0.0476)	(0.0479)
HH OOP med. exp.			-3.75e-05*
			(2.10e-05)
age	-0.0482**	0.00848	0.00786
	(0.0217)	(0.0312)	(0.0313)
has partner in household	0.281***	0.213***	0.216***
	(0.0596)	(0.0810)	(0.0811)
positive housing wealth	-0.0186	-0.0598	-0.0590
	(0.0467)	(0.0799)	(0.0800)
male	0.169***	0.169***	0.169***
	(0.0264)	(0.0410)	(0.0410)
high education	0.328***	0.331***	0.330***
	(0.0371)	(0.0597)	(0.0596)
cut-off point 1	8.552***	8.142***	8.261***
	(0.372)	(0.548)	(0.551)
cut-off point 2	10.14***	9.714***	9.834***
	(0.375)	(0.554)	(0.556)
cut-off point 3	11.63***	11.22***	11.34***
	(0.379)	(0.559)	(0.561)
σ^2_ξ	0.890***	0.895***	0.893***
	(0.0327)	(0.0509)	(0.0507)
Observations	25,827	11,000	11,000
Number of individuals	10,943	4,882	4,882
b. Health state dependence			
unit change in limitations			
Health equivalence scale ($\widehat{\mu}$)	1.133	1.102	1.089
	(1.044, 1.223)	(0.956, 1.249)	(0.947, 1.231)
Health state dependence ($\widehat{\phi}_1$)	0.284	0.215	0.186
	(0.081,0.487)	(-0.108, 0.538)	(-0.124, 0.495)
sd change in limitations			
Health equivalence scale ($\widehat{\mu}$)	1.029	1.022	1.019
	(1.010, 1.047)	(0.992, 1.053)	(0.989, 1.049)
Health state dependence ($\widehat{\phi}_1$)	0.058	0.045	0.039
	(0.020,0.096)	(-0.018, 0.107)	(-0.022, 0.010)

Standard errors clustered at the individual level. Income and OOP medical expenditures in 2004 German euros and equivalized to a one-person household using the OECD equivalence scale. Mundlak terms and dummies for countries, waves, and wave participation are included. $\widehat{\mu}$ and $\widehat{\phi}_1$ are constructed according to equations (3.20) and (3.22), choosing $\gamma = 3$. Panel A: standard errors in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$. Panel B: 95% confidence intervals between parentheses.

In order to be able to compare the results for different kinds of health limitations, we also provide the estimates for a standard deviation change in limitations, rather than a unit change. In this way we can take into account that the scale as well as the severity of limitations differ. After a standard deviation increase in limitations, one needs 2.9% more income to be financially as well off as before, with a health state dependence parameter of 0.058.[17]

3.4.2 Medical costs

For the baseline estimates we assumed that health insurance coverage in Europe is universal and that the insured in Europe have negligible OOP medical expenditures, so that we do not need to worry about private expenditures on health care. However, if individuals do bear medical costs in (some) European countries and are not insured for that, the positive health state dependence of utility that we find could be driven by increased medical expenditures, rather than an increase in the marginal utility of non-medical consumption. OECD (2015) shows that especially in the north of Europe health care is mainly financed by the government and social security. Private OOP medical expenses are low. However, other sources raise concern when it comes to universal coverage and OOP medical expenses (Cylus and Papanicolas 2015, and Scheil-Adlung and Bonan 2012). Therefore, we perform two tests for medical expenditures.

First, we include survey data on OOP medical expenditures in our regression. Unfortunately, data on OOP medical expenditures in SHARE are erratic.[18] The non-response rate is high and as from wave 4 the question components were changed. More importantly, in wave 4 the questions were

[17] The health state dependence parameter is sensitive to the risk aversion parameter γ, however, it does not change the sign of health state dependence. When we assume a risk aversion parameter of 2, health state dependence for a standard deviation change in health is 0.029. When we assume a risk aversion parameter of 4, health state dependence is 0.123.

[18] Therefore, we did not include them in the baseline regression.

temporarily moved to the drop-off questionnaire[19], so that imputations are not available for wave 4. If we want to include OOP medical expenditures we thus need to drop wave 4, leading to a large decline in the number of observations. To facilitate comparisons, we first re-estimate the baseline model on the new (smaller) sample. The results are shown in column (2) of table 3.2. The estimated health state dependence of utility is slightly smaller for this subsample, namely $\widehat{\phi}_1 = 0.215$ and insignificant, probably due to the lack of observations. Next, we add household OOP medical expenditures to the regression, for which the results are shown in column (3) of table 3.2. The point estimate for health state dependence hardly changes, from 0.215 to 0.186. We thus conclude that the presence of OOP medical expenses may lead to a slight overestimation of the health state dependence parameter, but does not drive our results.

Second, we exploit variation in institutions between countries. The data cover fifteen European countries that can be roughly divided into three groups based on their health care system. In the northern European countries the government is mainly responsible for organizing care. In the south/eastern countries, on the contrary, the family of the individual bears the main responsibility for providing care. In central countries, the responsibility for care is shared between the government and the family.[20] Medical costs are likely to be higher in countries where the government plays a small role in health care. Therefore, we expect individuals in central and south/eastern European countries to encounter more medical costs than individuals in northern European countries. In table 3.3 we interact all variables in the baseline model with the north, central and south/east European country groups. In correspondence with the test for OOP medical expenses, we find a slightly higher health state dependence parameter for central and southern/eastern European countries (0.272 and

[19]The drop-off questionnaire is a traditional paper questionnaire separately from the CAPI questionnaire. This questionnaire has a higher non-response than the regular CAPI questionnaire.

[20]Because of the number of observations we group countries together. Verbeek-Oudijk et al. (2014) classify northern countries: Denmark, Sweden, the Netherlands; central countries: Austria, Belgium, France, Germany; and south/eastern countries: Czech Republic, Estonia, Hungary, Italy, Poland, Portugal, Slovenia, Spain, Switzerland. We classified Greece to the south/east group.

Table 3.3: Test for coverage of health care costs

	North	Central	South/East
a. Estimation results			
limitation	-0.0837	-0.106**	-0.113*
	(0.0810)	(0.0525)	(0.0589)
$\ln(Y)$	0.901***	0.876***	0.913***
	(0.0873)	(0.0458)	(0.0533)
age	-0.0662**	-0.0216	-0.0726***
	(0.0264)	(0.0240)	(0.0252)
partner in household	0.483***	0.160	0.270***
	(0.110)	(0.104)	(0.0983)
positive housing wealth	0.266**	-0.0635	-0.0608
	(0.124)	(0.0711)	(0.0706)
male	0.0895	0.258***	0.112***
	(0.0602)	(0.0407)	(0.0433)
high education	0.187**	0.297***	0.485***
	(0.0802)	(0.0515)	(0.0718)
Observations		25,827	
Number of individuals		10,943	
b. Health state dependence			
unit change in limitations			
Health equivalence scale ($\widehat{\mu}$)	1.097	1.128	1.131
	(0.903, 1.292)	(0.995, 1.261)	(0.988, 1.275)
Health state dependence ($\widehat{\phi}_1$)	0.204	0.272	0.280
	(-0.223, 0.632)	(-0.028, 0.573)	(-0.046, 0.606)
sd change in limitations			
Health equivalence scale ($\widehat{\mu}$)	1.020	1.029	1.029
	(0.981, 1.060)	(1.000, 1.058)	(0.999, 1.060)
Health state dependence ($\widehat{\phi}_1$)	0.041	0.059	0.059
	(-0.039, 0.122)	(-0.001, 0.118)	(-0.004, 0.122)

Standard errors are clustered at the individual level. Income is in 2004 German euros and equivalized to a one-person household using the OECD equivalence scale. Mundlak terms and dummies for countries, waves, and wave participation are included. $\widehat{\mu}$ and $\widehat{\phi}_1$ are constructed according to equations (3.20) and (3.22), choosing $\gamma = 3$. Panel A: standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Panel B: 95% confidence intervals between parentheses.

0.280 compared to 0.204). However, health state dependence of utility in the central and south/east European countries is not significantly different from health state dependence in north European countries, suggesting that (OOP) medical expenses are not driving our results.

Alternative health measures 3.4.3

In order to improve our understanding of the mechanisms underlying health state dependence of utility, we re-estimate our model using other health measures. First, we examine problems with instrumental activities of daily living (IADL) and a measure of mobility problems. IADL and mobility problems are composite measures for a range of problems that can be encountered in daily life (just as ADL problems). Examples of IADL are shopping for groceries, taking medications and preparing a hot meal. Examples of mobility tasks are walking 100 meters, climbing stairs, and pulling/pushing large objects. IADL and mobility problems are of a milder nature than the ADL limitations, are observed at a higher frequency, and occur already earlier in life.

Columns (1) and (2) of table 3.4 show that IADL and mobility problems result in slightly larger estimates of health state dependence than ADL problems (0.377 and 0.366, compared to 0.284). This may indicate that, apart from increased preferences for assistance in housework, increased preference for assistance in leisure activities causes the positive effect (e.g. special transport and adapted leisure activities to keep up old ways of living). When physical limitations are mild, one might try hard to maintain the old lifestyle, but if physical limitations become too severe, one rather gives up on some of these activities. When we include the three types of limitations in one model simultaneously, the estimates are not significantly different.

Column (3) of table 3.4 shows that for each additional chronic disease, individuals need about 5.9% more income to be just as well off as before, corresponding to a parameter of health state dependence equal to 0.121. So, also here we measure positive health state dependence. The estimates are highly comparable, with a health state dependence parameter of 0.058

Table 3.4: Results for alternative health measures

	(1) limitations in IADL	(2) limitations in mobility	(3) # chronic diseases	(4) cognitive dysfunctioning
a. Estimation results				
limitation	-0.138***	-0.135***	-0.0501***	0.0807***
	(0.0294)	(0.0283)	(0.0194)	(0.0287)
$\ln(Y)$	0.866***	0.868***	0.878***	0.886***
	(0.0319)	(0.0323)	(0.0316)	(0.0327)
Observations	25,827	25,824	25,797	25,334
Number of id	10,943	10,942	10,928	10,736
b. Health state dependence				
unit change in limitations				
Health equivalence scale ($\widehat{\mu}$)	1.173	1.169	1.059	0.913
	(1.094, 1.252)	(1.093, 1.244)	(1.013, 1.105)	(0.855, 0.971)
Health state dependence ($\widehat{\phi}_1$)	0.377	0.366	0.121	-0.166
	(0.191, 0.562)	(0.190, 0.542)	(0.024, 0.218)	(-0.272, -0.060)
sd change in limitations				
Health equivalence scale ($\widehat{\mu}$)	1.044	1.044	1.030	0.976
	(1.025, 1.063)	(1.026, 1.063)	(1.007, 1.052)	(0.959, 0.992)
Health state dependence ($\widehat{\phi}_1$)	0.090	0.091	0.060	-0.048
	(0.050, 0.129)	(0.052, 0.130)	(0.013, 0.107)	(-0.081, -0.015)

Standard errors are clustered at the individual level. Income is in 2004 German euros and equivalized to a one-person household using the OECD equivalence scale. Age, dummies for the presence of a partner in the household, positive housing wealth, gender, education, countries, waves, and wave participation are included in addition to Mundlak terms. $\widehat{\mu}$ and $\widehat{\phi}_1$ are constructed according to equations (3.20) and (3.22), choosing $\gamma = 3$. Panel A: standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Panel B: 95% confidence intervals between parentheses.

for a standard deviation in ADL limitations and a parameter of 0.060 for a standard deviation in the number of chronic diseases. This strengthens our baseline results.

In addition to measures of physical health, column (4) of table 3.4 shows the results for cognitive functioning as a measure of health status. As individuals age, their ability to take initiative, to plan, and to organize activities decreases. This is, amongst others, probably due to a decline in the functioning of the frontal lobe, which is vulnerable to the effects of aging (this has been found both in behavioral and MRI studies, Craik and Grady 2002). The frontal lobe is also heavily demanded in memory tasks. In this study we use a test on long term word recall to construct a measure of cognitive functioning.[21] Whereas we find a positive effect of physical limitations on the marginal utility of consumption, we find that cognitive limitations have a negative effect on the marginal utility of consumption. This suggests that, in contrast to individuals with physical limitations, individuals that experience a cognitive decline probably might not be willing to invest in adapted activities, as their motivation and will to undertake any activities declines.[22]

Robustness checks 3.4.4

The functional form of a model may have an impact on the size of the coefficients. Following the method of Pradhan and Ravallion (2000) we estimate the coefficients using a random effects ordered probit model. However, one could also think about a linear specification. Riedl and Geishecker (2014) compare linear and nonlinear ordered response estimators and find that in general the choice seems to have little effect on the size of ratios of estimated coefficient. Column (2) of table 3.5 shows the results for the linear specification. The estimated health state depen-

[21] See appendix 3.A for more details on the measure of cognitive functioning.

[22] One could argue that after a negative shock in cognitive health, administrative help may increase, and this could be the reason that making ends meet improves. However, only 6.8% of the individuals with a negative shock in cognitive health start to receive administrative help between two waves, which is not significantly different from those without a negative shock in cognitive health (on average 8.3% of them start to receive administrative help from one wave to the other).

Table 3.5: Specification checks

	(1)	(2)	(3)	(4)	(5)	(6)
		specification:		exclude individuals with		
	baseline	linear	country x year FE	< 75% wealth annuitized	low perm. income (Q1)	high perm. income (Q4)
a. Estimation results						
limitation	-0.110***	-0.0603***	-0.112***	-0.115***	-0.146***	-0.105**
	(0.0353)	(0.0196)	(0.0354)	(0.0372)	(0.0454)	(0.0419)
$\ln(Y)$	0.880***	0.468***	0.899***	1.130***	0.636***	1.214***
	(0.0323)	(0.0161)	(0.0331)	(0.0390)	(0.0469)	(0.0575)
age	-0.0482**	-0.0262**	-0.0393*	-0.0233	-0.0556**	-0.0251
	(0.0217)	(0.0116)	(0.0237)	(0.0221)	(0.0246)	(0.0239)
Observations	25,827	25,827	25,827	23,946	17,441	16,727
Number of id	10,943	10,943	10,943	10,224	7,412	7,123
b. Health parameter estimates						
unit change in limitations						
Health equivalence scale ($\widehat{\mu}$)	1.133	1.138	1.132	1.107	1.258	1.09
	(1.044, 1.223)	(1.044, 1.231)	(1.044, 1.220)	(1.035, 1.179)	(1.078, 1.438)	(1.016, 1.164)
Health state dependence ($\widehat{\phi}_1$)	0.284	0.294	0.281	0.226	0.582	0.188
	(0.081,0.487)	(0.081, 0.507)	(0.082, 0.480)	(0.067, 0.385)	(0.129, 1.035)	(0.026, 0.35)
sd change in limitations						
Health equivalence scale ($\widehat{\mu}$)	1.029	1.03	1.029	1.023	1.051	1.02
	(1.010, 1.047)	(1.001, 1.049)	(1.010, 1.047)	(1.008, 1.038)	(1.018, 1.083)	(1.004, 1.037)
Health state dependence ($\widehat{\phi}_1$)	0.058	0.06	0.058	0.047	0.104	0.041
	(0.020,0.096)	(0.021, 0.100)	(0.021, 0.095)	(0.016, 0.077)	(0.036, 0.172)	(0.008, 0.075)

Standard errors are clustered at the individual level. Income is in 2004 German euros and equivalized to a one-person household using the OECD equivalence scale. Age, dummies for the presence of a partner in the household, positive housing wealth, gender, education, countries, waves, and wave participation are included in addition to Mundlak terms. $\widehat{\mu}$ and $\widehat{\phi}_1$ are constructed according to equations (3.20) and (3.22), choosing $\gamma = 3$. Panel A: standard errors in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$. Panel B: 95% confidence interval between parentheses.

dence parameter is very similar (0.294 compared to 0.284). Also, including country-by-year fixed effects rather than country effects and year effects separately (to account for possible country-specific shocks), does not affect our results (0.281 compared to 0.284), as shown in column (3).

An important assumption underlying our model is that permanent income is not affected by health. This does not need to be the case since a health shock can affect life expectancy and thereby the amount of wealth that can be consumed each remaining year. When we rerun our model on the subsample of households with 75% of wealth annuitized the results hardly change (column (4) of table 3.5), so that our assumption seems justified.

Financial wellbeing is measured on a four-point scale. The fact that the scale is finite may lead to a bias in our estimates. The baseline estimates show that, on average, individuals rate their financial wellbeing lower after a health shock than before. Now imagine a healthy individual who is very dissatisfied with his or her financial situation. When this individual experiences a health shock, (s)he might experience an even lower financial wellbeing than before, but cannot express this in the answer to the survey question, because he was already at the lowest category of financial wellbeing. This would bias our estimate of the health state dependence of utility downward.

As a robustness check we therefore repeat the baseline regression, while excluding those in the lowest or highest income quartile for each country within the selected sample. As shown in figure 3.3, healthy individuals with low incomes often provide answers on the bottom end of the scale. Therefore, we would expect to find a larger decline in financial wellbeing after a health shock and thus a higher estimate of health state dependence when excluding the lowest income quartile. Indeed, as shown in column (5) of table 3.5, the estimated parameter is larger than in the baseline (0.582 compared to 0.284). The higher estimate can also be the result of heterogeneous effects for poor and rich individuals. In any case, the positive health state dependence found in our baseline specification remains convincing. This also holds when we exclude the highest income quartile for each country (column (6) of table 3.5).

In a recent paper, Finkelstein et al. (2013) find negative health state dependence using a measure of chronic disease count on a sample of elderly US citizens. For a comparable sample to ours (age ≥ 65, not in the labor force, and with health insurance) and under the assumption of $\gamma = 3$, they find a proportionality factor of -0.142 for a one standard deviation change in limitations.[23] To ensure that these differences in estimates do not stem from methodological differences, we re-estimate the model of Finkelstein et al. (2013) on our dataset.[24] Our hypothesis is that as long as there are no unobserved characteristics that are correlated with income and have an effect on the change in utility after a health shock, the methods should provide roughly the same results.

Finkelstein et al. (2013) use happiness as a proxy for utility. Unfortunately, this variable is not available in SHARE. Instead, we use general life satisfaction. General life satisfaction and happiness are no exchangeable concepts (as pointed out by Kahneman and Deaton 2010), but we believe it suffices for this context. We apply a linear random effect specification with Mundlak terms to deal with possible problems of endogeneity. Because the answering scale of the life satisfaction question was different in wave 1 than in the other waves, we drop wave 1 for this part of the analysis. Column (1) of table 3.6 shows the results of the baseline regression excluding the observations of wave 1. The estimate is almost equal to the estimate including wave 1, and has only a slightly larger 95% confidence interval. Column (2) shows the estimates based on the same sample, but using overall life satisfaction and the method of Finkelstein et al. (2013). We find the same sign of health state dependence, but the coefficient is somewhat larger ($\hat{\phi}_1 = 0.438$ instead of 0.285) and not significant. Lastly, we re-estimate column (2) using the same health measure as Finkelstein et al. (2013), namely the number of chronic diseases. This measure has been constructed in correspondence with the baseline health measure used in Finkelstein et al. (2013). As shown in column (3) of table 3.6 we find an

[23] See table 3, column 2b of Finkelstein et al. (2013).

[24] The alternative, to apply our method to the HRS data, unfortunately will not work, since individuals in this dataset are likely to have out-of-pocket health expenditures (also when they are insured), which are not recorded in the data.

Table 3.6: Robustness check: method

	(1) Baseline excl. wave 1	(2) Method Finkelstein et al. (2013)	(3)
	at least one ADL	at least one ADL	# chronic diseases
a. Estimation results			
limitation	-0.114***	-0.295***	-0.114***
	(0.0400)	(0.0513)	(0.0270)
$\ln(Y)$	0.911***		
	(0.0367)		
$Y^{1-\gamma}$		-0.0801***	-0.0773***
		(0.0195)	(0.0281)
limitation$*(Y^{1-\gamma})$		-0.0351	-0.0316*
		(0.0391)	(0.0179)
age	-0.0727**	-0.0123	-0.00427
	(0.0307)	(0.0371)	(0.0376)
partner in household	0.344***	0.257***	0.249***
	(0.0711)	(0.0786)	(0.0785)
positive housing wealth	0.0195	0.0222	0.0213
	(0.0526)	(0.0616)	(0.0617)
male	0.201***	0.00546	0.00868
	(0.0292)	(0.0316)	(0.0321)
high education	0.306***	0.244***	0.261***
	(0.0402)	(0.0378)	(0.0379)
Observations	20,025	20,508	20,490
Number of individuals	9,022	9,222	9,213
b. Health state dependence			
unit change in limitations			
Health state dependence ($\widehat{\phi}_1$)	0.285	0.438	0.409
	(0.063, 0.507)	(-0.540, 1.417)	(-0.102, 0.919)
sd change in limitations			
Health state dependence ($\widehat{\phi}_1$)	0.057	0.098	0.192
	(0.017, 0.098)	(-0.121, 0.316)	(-0.048, 0.431)

Dependent variable: general life satisfaction. Standard errors are clustered at the individual level. Income is in 2004 German euros and equivalized to a one-person household using the OECD equivalence scale. For sake of readability of the coefficients in column (2) and (3) Y is divided by 10,000. Mundlak terms and dummies for countries, waves, and wave participation are included. For the calculation of health state dependence in column (1) and the construction of the income measure in columns (2) and (3) we assume $\gamma = 3$. Panel A: standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Panel B: 95% confidence intervals between parentheses.

estimate of 0.192 (insignificant at a 5% level) for a one standard deviation change in limitations.

This analysis shows that the contradictory signs of the health state dependence of utility found between the US and Europe seem to be due to heterogeneity in the population, as opposed to methodological differences. This may be due to cultural differences, institutional differences or differences in consumption patterns. Banks et al. (2015) compare life-cycle consumption patterns for the US and the UK.[25] They show that budget shares on recreation are higher in the UK than in the US. If people in Europe find recreation a more essential consumption good than people in the US, they may need more money to keep doing these activities in bad health (e.g. more help and assistance during a holiday). Transportation costs, on the other hand, are relatively high in the US compared to the UK (maybe because of the longer distances in the US). The question arises whether Americans still incur these costs when they become sick. If not, this may explain (part of the) negative health state dependence in the US. Future research on consumption data is needed to provide clear answers.

3.5 Conclusion

This chapter estimates health state dependence of utility in Europe. We develop an approach that uses insights from the domain of living standards and income adequacy (Pradhan and Ravallion 2000). We derive a simple relationship between the health state dependence parameter within the optimal life cycle framework and a so called 'health equivalence scale', which we define as the relative change in income needed to maintain the same level of financial wellbeing after a health shock. This allows us to identify health state dependence of utility using a measure of financial wellbeing. Compared to other observed utility approaches a benefit of this method is that it does not require to compare individual level changes at different income levels to overcome the obstacle that health shocks affect both the level and shape of the utility function. Since making ends meet is

[25] The UK is not included in SHARE, but this comparison may give us some clue about differences between the US and Europe.

a common question in questionnaires, this method can easily be used by researchers all over the world.

We implement the approach using panel data from the Survey of Health, Aging and Retirement in Europe. Our baseline estimates show positive health state dependence of the utility function, with a proportionality factor equal to 0.284 in the presence of ADL limitations. We show that our results are not driven by medical expenditures, and that they are robust against alternative physical health measures and specifications. Interestingly, for cognitive health limitations we find negative health state dependence. When cognitive health declines, people's ability to plan, organize, and take initiative becomes worse. These developments seem to lower the marginal utility of consumption.

To compare the results for the US and Europe, we also apply the approach of Finkelstein et al. (2013) on the SHARE data. Because of data limitations the comparison is not completely clean, but again, we find positive health state dependence. The sign thus remains the same, however, the coefficient is somewhat larger and not significantly different from zero. Remarkably, for the European countries in our sample we find positive health state dependence repeatedly, whereas Finkelstein et al. (2013) find negative health state dependence in the US for the same age group (retirees). The results suggest that this is due to heterogeneity in the population as opposed to methodological differences. Differences in consumption patterns, such as budget shares on leisure and transportation may explain the contradictory signs in the US and Europe for health state dependence of the utility function. Further research on consumption patterns is needed to explain this difference. Further research could also aim at exploring heterogeneities in health state dependence of utility across individuals with different characteristics.

The health state dependence parameter is important for many economic questions such as the optimal savings rate and the optimal level of health insurance. Positive health state dependence implies that both the optimal savings rate and the optimal level of health insurance increase, relative to the situation where health state dependence is not taken into account. However, in old age cognitive health also declines and this lowers the

marginal utility of consumption. As far as health limitations occur early in the life-cycle when people are in good cognitive health, extra money may be desirable to be able to keep doing (leisure) activities. On the other hand, in old age, where health limitations coincide with a decline in cognitive health, our results show that extra money or a high insurance for health costs is less necessary.

Additional descriptive statistics 3.A

This appendix provides details about the data and a description of the data used for the additional analyses.

Health 3.A.1

This chapter uses five different health measures. In the baseline analysis we use data on limitations in activities of daily living, described in section 3.3. In the additional analyses we use data on instrumental activities of daily living, physical mobility, the number of chronic diseases and cognitive functioning. This section provides a description of these variables.

IADL and mobility

Both limitations in instrumental activities of daily living (IADL) and limitations in physical mobility are closely related to limitations in activities of daily living (ADL). IADL is scored with a list of seven activities: (1) using a map to get around, (2) preparing a hot meal, (3) shopping for groceries, (4) making telephone calls, (5) taking medications, (6) working around the house or garden, and (7) managing money. These type of limitations are slightly milder than limitations with ADL, therefore are more prevalent: 26% of observations have at least one limitation in IADL, see table 3.7. The most common limitations are limitations with house and garden work, grocery shopping and using a map. After the age of 80 also problems with managing money and preparing a hot meal start to become more prevalent, see figure 3.4.

Table 3.7: Summary statistics for health variables

	mean	sd. within	sd. between	min	max	N
limitation in ADL	0.16	0.29	0.23	0	1	25827
# chronic disease	1.59	0.65	1.03	0	7	25797
limitation in IADL	0.26	0.34	0.28	0	1	25827
limitation in mobility	0.65	0.35	0.32	0	1	25824
limitation in cognitive ability	0.75	0.35	0.26	0	1	25334

Figure 3.4: Limitations in instrumental activities of daily living across age

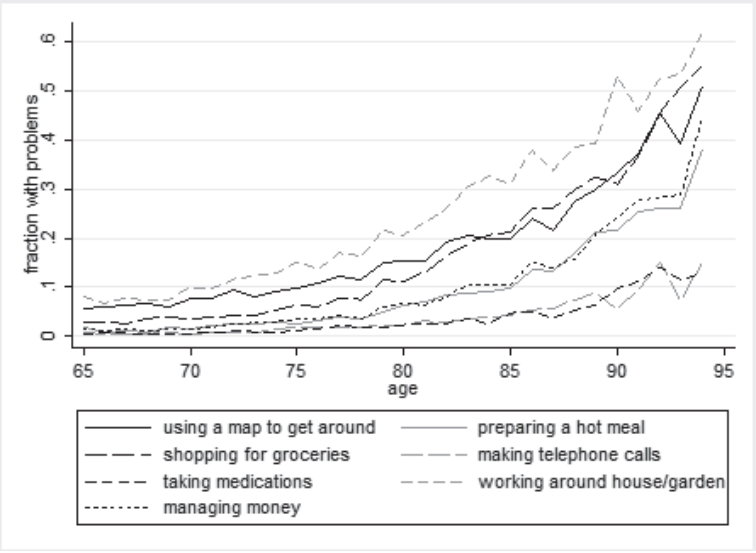
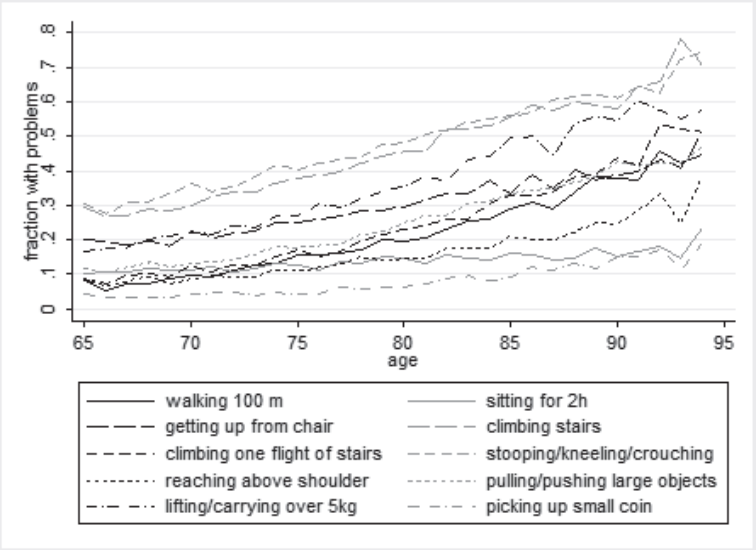


Figure 3.5: Limitations in physical mobility across age



Physical mobility is scored with a list of ten activities: (1) walking 100m, (2) sitting for 2 hours, (3) getting up from a chair, (4) climbing stairs, (5) climbing one flight of stairs, (6) stooping, kneeling or crouching, (7) reaching above one's shoulder, (8) pulling or pushing large objects, (9) lifting or carrying over 5kg, and (10) picking up a small coin. These limitations are even milder than those listed for IADL, not only do more individuals experience these type of limitations (65% of all observations, see table 3.7), also they start to appear earlier in life. At early ages individuals often report problems with stooping, kneeling or crouching or climbing stairs, followed by sitting for 2 hours and getting up from a chair, see figure 3.5.

As the listed limitations in physical mobility and IADL are slightly milder than those in ADL, often individuals who report to have limitations in ADL, also report to have limitations in one of the other categories. Figure 3.6 shows the prevalence of the different limitations across age.

Chronic diseases

The prevalence of chronic diseases is an objective measure of health which is often used in the literature. For sake of comparability, we construct this variable in correspondence with Finkelstein et al. (2013). SHARE contains a list of 18 chronic diseases (three of which are not included in the first wave), for which the respondents should indicate whether a doctor has ever told that them they had or that they are currently being treated for or bothered by that condition. We include (1) heart attack/heart problems, (2) high blood pressure/hypertension, (3) stroke/cerebral vascular disease, (4) diabetes/high blood sugar, (5) chronic lung disease (e.g. chronic bronchitis or emphysema), (6) arthritis, including osteoarthritis or rheumatism, (7) cancer or malignant tumor, including leukemia or lymphoma, but excluding minor skin cancers. Contrary to ADL limitations, chronic diseases are considered permanent. For each chronic disease, prevalence is modeled as an absorbing state. 82% of the person-year observations have at least 1 chronic disease. High blood pressure and arthritis are the most common diseases, see figure 3.7.

Figure 3.6: Different type of limitations across age

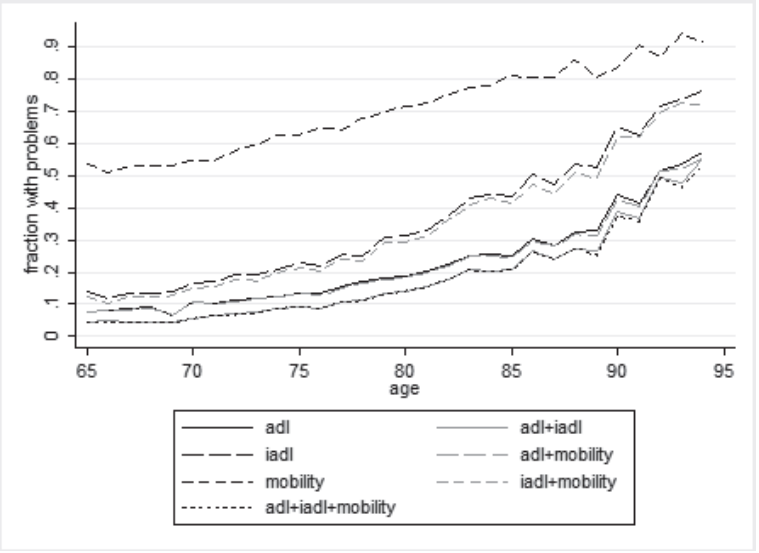
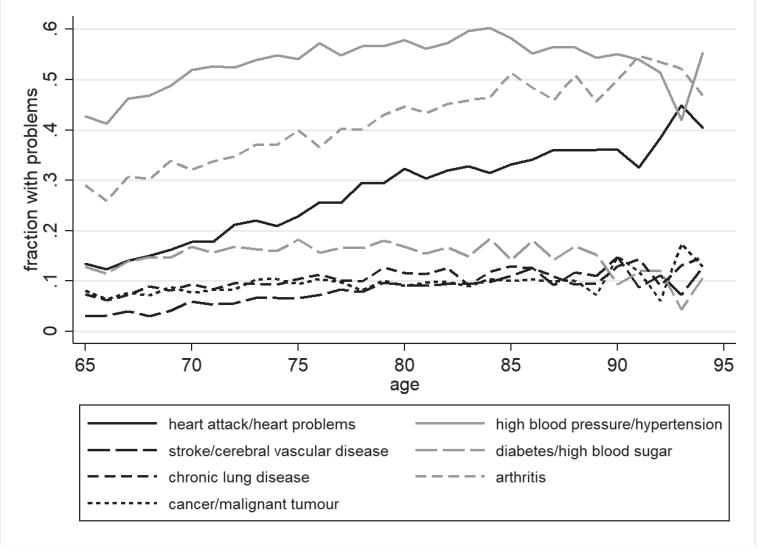


Figure 3.7: Chronic diseases across age



Cognitive functioning

Cognitive functions encompass aspects like attention, memory, perception, language, and decision making (Glisky 2007). As individuals age cognitive functioning may decline, such that the ability to take initiative, plan, and organize activities decreases and individuals experience more fear.

Our measure of cognitive functioning is based on a word recall question. The interviewer reads a list of 10 words to the survey respondent. After the interviewer is finished he asks the respondent to recall aloud as many words as possible, in any order. The number of words that the respondent remembers is stored and seen as the ability of short term recall. After this task the respondent is given a verbal fluency task and a series of computational tasks. Then he is asked again to remember aloud the words provided in the beginning of the series of tasks, in any order. He gets one minute to complete the task. The amount of words that the respondent remembers at this point is seen as the ability of long term recall. We use this second variable to measure cognitive functioning: the more words one can remember, the higher the cognitive functioning. We dichotomize the variables, in correspondence with the other health measures, by defining someone to experience cognitive limitations if the number of words recalled is less than average in our sample (5/10 words recalled). According to this cut off 75% of individuals has a limitation in cognitive functioning, see table 3.7.

Income and assets 3.A.2

SHARE contains detailed questionnaires on income and assets. Different income components are elicited stepwise, so that we can be sure transfers from private and public pension programs are earmarked as income.

Over the years there are some slight changes in the wording of the questionnaire. Most importantly, in the first wave the respondents are asked to provide gross amounts of income, whereas from wave two on net amounts are elicited. To correct for this we translated the gross amounts to net amounts using information from the OECD on each country's average

Table 3.8: Income split up by component and country

	permanent income		household income			financial wealth			housing			
				conditional on ownership			conditional on ownership			conditional on ownership		
	mean	median	owner	mean	median	owner	mean	median	owner	mean	median	N
Austria	18648	16183	1.00	18108	15562	0.92	13938	4231	0.46	164692	132314	2415
Germany	17899	16205	1.00	16990	15065	0.92	22548	9078	0.54	146961	125253	1259
Sweden	25376	23371	1.00	23715	18625	0.91	37430	20317	0.68	94539	64884	2005
Netherlands	24796	20786	1.00	23420	17802	0.96	34800	11185	0.49	189272	173505	1790
Spain	10651	9052	0.99	10366	8766	0.87	9472	2110	0.93	149524	101360	1519
Italy	12912	10771	0.98	12640	10373	0.76	14658	5901	0.80	138570	110357	1622
France	21227	18023	1.00	19196	15741	0.95	43641	10416	0.73	191325	143973	3100
Denmark	18930	16108	1.00	17645	13115	0.89	32815	13731	0.61	106058	76290	1596
Greece	9683	8550	0.98	9635	8169	0.42	11137	5859	0.84	90396	77316	954
Switzerland	34902	29106	1.00	30749	24983	0.95	89015	37597	0.49	317244	210254	1529
Belgium	29053	19547	1.00	26864	15139	0.95	49482	15455	0.75	162389	144667	3069
Czechia	9305	8988	0.98	9101	8887	0.66	10270	5473	0.69	83186	66327	1763
Poland	6223	5549	1.00	6115	5463	0.66	4123	3621	0.71	48101	32429	260
Slovenia	19022	12246	0.99	19029	10301	0.86	4489	502	0.84	116696	92260	804
Estonia	7217	6651	1.00	6558	5829	0.90	15029	2092	0.81	335775	92674	2142
Total	19225	15092	0.99	17999	13080	0.87	31625	7914	0.68	167822	114943	25827

Based on the first set of imputations

tax rate per decile for singles and couples separately.[26] Table 3.8 provides detailed information on income and assets per country.

3.A.3 Positive and negative feelings

Our results may be biased due to time-varying optimism and pessimism correlated with health status lead to changes in answering styles. Although we cannot control for this in a proper way, we do want to check what happens when we include measures of positive and negative feelings in the model. This is not without problems, since by including these variables in our estimating equation we introduce simultaneity bias into our model. The results from this model should thus definitely by interpreted with caution.

Using questions from the EURO-D depression scale, we construct measures of positive and negative feelings, similar to Fischer and Sousa-Poza (2008). Our variable 'negative feelings' is the summation of dummy variables indicating whether an individual experiences feelings of sadness, guilt, and hostility. The variable 'positive feelings' is a summation of dummy variables indicating whether an individual experiences feelings of self-assurance, attentiveness, and joviality. As shown in table 3.9, individuals with ADL limitations tend to experience less positive feelings and more negative feelings. When including the measures of positive and negative feelings to the regressions, the estimated health state dependence parameter is still significant at a 5% level and only slightly smaller than the baseline estimate, 0.208 compared to 0.284 (column (1) of table 3.10).

[26] OECD.stat: Benefits, Taxes and Wages - Net incomes 2004 retrieved from http://stats.oecd.org/Index.aspx?DataSetCode=FIXINCLSA.

Table 3.9: Percentage of respondents experiencing positive and negative feelings.

	$\geq$ 1 ADL problem	no ADL problem
Positive feelings		
- has hopes for the future	66%	82%
- has enjoyed an activity recently	75%	88%
- has good concentration on reading/entertainment	62%	82%
Negative feelings		
- has been sad or depressed in the last month	60%	40%
- has felt to rather be death in the last month	21%	8%
- has feelings of guilt or self-blame	11%	7%
- has been irritable recently	31%	21%
- has cried in the last month	39%	25%

Table 3.10: Results including measures of positive and negative feelings

	(1) baseline	(2) incl. pos. and neg. feelings
a. Estimation results		
limitation	-0.110***	-0.0811**
	(0.0353)	(0.0361)
$\ln(Y)$	0.880***	0.857***
	(0.0323)	(0.0322)
age	-0.0482**	-0.0243
	(0.0217)	(0.0217)
positive feelings		0.0720***
		(0.0164)
negative feelings		-0.0333***
		(0.0121)
Observations	25,827	25,128
Number of id	10,943	10,658
b. Health parameter estimates		
unit change in limitations		
Health equivalence scale ($\widehat{\mu}$)	1.133	1.099
	(1.044, 1.223)	(1.008, 1.190)
Health state dependence ($\widehat{\phi}_1$)	0.284	0.208
	(0.081,0.487)	(0.008, 0.409)
sd change in limitations		
Health equivalence scale ($\widehat{\mu}$)	1.029	1.021
	(1.010, 1.047)	(1.002, 1.04)
Health state dependence ($\widehat{\phi}_1$)	0.058	0.043
	(0.020,0.096)	(0.005, 0.082)

Standard errors are clustered at the individual level. Income is in 2004 German euros and equivalized to a one-person household using the OECD equivalence scale. For sake of readability of the coefficients in column (2) and (3) Y is divided by 10,000. Mundlak terms and dummies for countries, waves, and wave participation are included. For the calculation of health state dependence in column (1) and the construction of the income measure in columns (2) and (3) we assume $\gamma = 3$. Panel A: standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Panel B: 95% confidence intervals between parentheses.

4 Graded Return-to-Work as a Stepping Stone to Full Work Resumption

Abstract

While there is increasing evidence that graded return-to-work is an effective tool for the rehabilitation of sick-listed workers, little is known on the optimal timing and level of grading in return-to-work trajectories. We use administrative data from a Dutch private workplace reintegration provider to fill this gap. In order to correct for the selection bias inherent to the evaluation of activation strategies, we exploit the discretionary room of the case managers in setting up treatment plans. We find that graded return-to-work reduces sick spells with eighteen weeks within the first two years after reporting sick. However, the probability of work resumption after two years remains unchanged. Work resumption can be achieved faster when graded return-to-work is started earlier or at a higher rate of work resumption. These findings however do not hold for individuals who have problems related to mental health.

A working paper version of this chapter is published as Kools and Koning (2018) and is currently under review. The chapter is co-authored by Pierre Koning. The authors gratefully acknowledge the financial support from Instituut Gak for this research project. They also thank Bénédicte Rouland, Rob Euwals and Bertjan Teunissen for detailed comments and suggestions on earlier versions of the paper, as well as seminar participants that gave feedback at the EALE conference in St. Gallen in 2017 and the research seminars at CPB Netherlands Bureau for Economic Policy Analysis, the University of Antwerp and RWI in Essen.

4.1 Introduction

In the past decades many Western countries have seen a rise in uptake of disability benefits (OECD 2010). In an effort to curb this trend, there has been an increased focus on what disabled individuals can do at work, rather than what they cannot. For example, in England sick notes have been replaced by a statement of fitness for work in 2010 (Wainwright et al. 2011), in Sweden general practitioners are recommended to subscribe part-time sick leave rather than full time sick leave (Kausto et al. 2008) and in Norway sick-listed employees are since 2004 required to work partially after eight weeks of sick leave unless a physician has stated this is impossible (Hernæs 2017). In a similar vein, part-time sick leave is often used as a workplace based intervention aimed at speeding up the rehabilitation process of sick-listed employees. In these interventions usually the amount of hours worked gradually increases over time, up to the moment that full work resumption is achieved. The idea is that graded work prevents the loss of working skills and may even speed up the recovery from certain injuries. For instance, Andren and Svensson (2012) argue that particularly individuals with musculo-skeletal problems benefit from graded work activities. Likewise, *Individual Placement and Support* (IPS) interventions for sick workers with mental impairments are built upon the idea that work activities may contribute to the recovery process.[1]

Research shows almost unanimously positive effects of graded work on work rehabilitation[2], whereas interventions like vocational rehabilitation and regular paramedical care rather seem to lengthen sick spells (Markussen and Røed 2014, Rehwald et al. 2016). This however does not mean that graded return-to-work is beneficial for all individuals (Andren

[1] Corrigan and McCracken (2005) argue that psychiatric problems can be addressed only for some workers in real-life settings, so as to identify the cause of them.

[2] See e.g. Bernacki et al. (2000), Bethge (2016), Hernæs (2017), Høgelund et al. (2010), Kausto et al. (2014), Markussen et al. (2012), Rehwald et al. (2016), Viikari-Juntura et al. (2012). The general finding that graded work increases work resumption is confirmed in peer reviewed papers on the effects of part-time sick leave, active sick leave, phased return to work, and graded return to work. Related to this literature, evidence on graded work exposure or light duties also points at positive results, see e.g. Krause et al. (1998).

2014, Andren and Svensson 2012, Høgelund et al. 2012). Starting graded work trajectories too soon or for too many hours may induce stress or strain on the body, hampering the recovery process. In light of these considerations, it is important to understand what separates an effective graded return-to-work trajectory from an ineffective one.

In this chapter we analyze how the specifics of the set-up of a graded return-to-work trajectory determine its effectiveness. More specifically, we analyze if work resumption rates change when the trajectory is started later or at a higher initial rate of work resumption. For this we make use of registered data from a private workplace reintegration provider, which performs case management for mostly small and medium sized firms. This provider offered reintegration services for about 12,000 long-term sick listed workers, of which 62% participated in graded work trajectories between the years 2011 and 2014. We observe detailed worker information on the timing and the degree of grading that is used, as well as information on impairment types, employer, and other individual characteristics. We enrich these data with information on the case managers that were assigned to them by the reintegration provider.

In order to correct for the selection bias inherent to the evaluation of activation strategies, we follow an instrumental variables approach for which we exploit the discretionary room of the case managers in setting up treatment plans. We use the tendency of a case manager to focus on early/intense graded work (graded work propensity) as an instrument to actually receiving such a strategy. In doing so, we follow a strand of literature applying this technique in the context of activation strategies for sick-listed employees, such as Dean et al. (2015), Markussen and Røed (2014) and Rehwald et al. (2016).[3] As case managers may learn on the job or change their preference for graded work, we allow graded work propensities to vary across years. Our key assumption is that the assignment of (new) sick-listed workers to case managers is

[3]For the Dutch case, where sick-listed employees have to follow a return-to-work plan established in the beginning of the sick-spell, we prefer this approach over the use of proportional hazard models, as used by for example Høgelund et al. (2010) for the case of Denmark, which relies on the non-anticipation assumption. Other methods used in the context of graded return-to-work are propensity score matching (Bethge 2016) and randomized control trials (Viikari-Juntura et al. 2012).

exogenous. We argue that this assumption is plausible, as the assignment is driven by the direct availability of case managers. Moreover, all the individual information on new sick-listed workers that is available to the case managers at the moment of intake is observed in our data. This means that any selection on observables can be controlled for. Reversely, we also can test for the importance of such selection effects by estimating model specifications without individual controls.

Our analysis also extends on earlier studies in this field of research by using alternative propensity measures that proxy the specifics of graded work trajectories. In line with earlier work, we will first define case managers' propensity measures as the likelihood of initiating a trajectory for sick-listed workers that haven't started one yet. With the information of workers that have effectively started a trajectory, we next construct propensity measures of case managers that proxy the timing of graded work during the sickness spell as well as the graded work percentage that is applied. This then enables us to evaluate the effects of differences in the timing and the degree of grading of the interventions on (full) work resumption for those individuals that have started a graded work program. We thus gain insight in the optimal timing of graded work and the importance of gradually increasing the degree of grading.

We also shed new light on the determinants of graded work propensities and the implications of this for the interpretation of our findings. Even though the case managers' tendencies to use graded work interventions can be considered as exogenous, we cannot be sure that they are uncorrelated with other case manager characteristics affecting the likelihood of work resumption. For instance, high graded work propensities may be a marker of high quality case managers that also show higher work resumption rates without the use of graded work interventions. If so, the effectiveness of graded work interventions will be overestimated. We therefore estimate model versions with various proxies for case manager quality as additional controls. Among others, these proxies include the current and past work resumption rates of (other) sick-listed workers that were assigned and work resumption rates of workers that are out of our sample. When controlling for these proxies, we are able to assess the

extent to which graded work effects are truly driven by the allocation to trajectories, rather than other case manager activities that are correlated with graded work.

In line with earlier literature, we find overall positive effects of graded return-to-work. Graded return-to-work speeds up the recovery process. At the same time, graded work does not necessarily help rehabilitate individuals who would otherwise have not rehabilitated. We find an increase in the number of weeks worked during the first two years after sick-listing of 18 weeks due to graded return-to-work, but no significant effects on the probability to return to work within two years. Moreover, we find that starting the graded return-to-work trajectory earlier and at a higher rate of work resumption speeds up the recovery process. Starting one week earlier raises the number of weeks worked in the first two years with two weeks. Starting a graded return-to-work trajectory at a work resumption rate which is 10 percentage point higher increases the probability to return to work within two years with 2.5 percentage point. Work resumption rates are more strongly affected by the moment that graded return-to-work is started than by the moments within the trajectory at which the level of work resumption is increased.

The positive effects of graded return-to-work are especially strong for individuals who have general medical conditions. For them the positive effects persist at the end of the waiting period. For individuals with problems related to mental health we find no significant effects of graded return-to-work. Moreover, and contrary to the overall findings, for these individuals starting the graded return-to-work trajectory one week earlier decreases the probability to return to work within two years with 3 percentage point.

In the following section, we explain the system of sick leave and disability insurance in the Netherlands. Then, in Section 4.3 we provide descriptive statistics on the sick-listed individuals in the data set, the graded return-to-work trajectories, and the case managers. In Section 4.4, we explain our empirical strategy and underlying assumptions. We present the results of the analysis in Section 4.5, followed by concluding remarks in Section 4.6.

4.2 Institutional setting

The Dutch disability system used to be notorious for its large and increasing number of beneficiaries; at its peak those receiving benefits amounted up to 12 percent of the insured individuals (Koning and Lindeboom 2015). Since the beginning of the 21st century disability insurance award rates have been steadily declining, due to a number of reforms to the system. Among these reforms was the introduction of the Gatekeeper Protocol, obliging employers and employees to engage in activities aimed at reintegrating sick-listed workers into the workforce. As a consequence of the Gatekeeper Protocol, disability insurance inflow was estimated to reduce by about 40 percent (van Sonsbeek and Gradus 2013). This positive effect can partly be attributed to improved screening, making it more difficult to use DI as an alternative exit root for Unemployment Insurance (de Jong et al. 2011). Moreover, increased employer responsibilities have played a crucial role in curbing the rise in DI beneficiaries, both as a stimulus to actively prevent sickness and as a way to accommodate activation strategies for sick-listed workers (Koning and Lindeboom 2015).

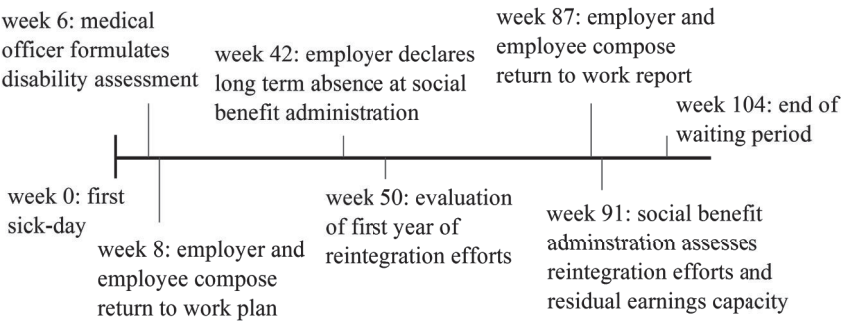
As a result of the reforms the Netherlands has a rather unique, largely privately organized sickness and disability system (Koning 2017). This section describes those elements of the system that are relevant for understanding the context within graded return-to-work is used.

4.2.1 Gatekeeper Protocol

In the Netherlands all workers are insured against income losses due to injuries, irrespective of having incurred the injury at the workplace or not. Individuals can apply for DI benefits after a two year waiting period, during which the employer is obliged to continue payments of at least 70% of the employees regular salary.[4] In practice, most Collective Labor Agreements stipulate full wage payments in the first year and 70% in the

[4]For comparison, in Scandinavian countries employers are responsible for two to three weeks of continued wage payments, after which the Social Insurance Administration (Sweden/Norway) or municipalities (Denmark) take over the burden (Andren 2014, Markussen and Røed 2014, Rehwald et al. 2016).

Figure 4.1: Time line of the gatekeeper protocol.



second year. During these two years, sick-listed workers can start with graded work or adapted work. As long as the waiting period proceeds and the worker has not fully recovered, wage payments are continued.

During the waiting period the employer and the employee are obliged to undertake efforts towards re-integration of the sick-listed employee. The Gatekeeper Protocol (in Dutch: *Wet verbetering Poortwachter*) gives directions as to what these efforts should entail. Figure 4.1 shows a time line of the concrete steps that need to be taken under the Gatekeeper Protocol. In the sixth week a disability assessment should be conducted by a medical officer (company doctor). This assessment is used as input for a reintegration plan, due in week 8. This plan is composed by the employer and employee, and should stipulate the reintegration aim[5] and the planned steps towards reaching this aim. A case manager should be appointed to keep track of the reintegration process and the return-to-work plan may be reevaluated at set dates.

After 42 weeks of sick-listing, the employer has to declare the sick-listed employee to the Social Benefit Administration (responsible for Disability Insurance) and after a year the reintegration efforts undertaken so far have to be evaluated. In the 87th week the employer and employee have to compose a return to work report, including all the reintegration efforts

[5]Preferably, the reintegration aim should be (partial/adapted) employment with the current employer 'first track' reintegration). Only if this is out of reach, one can aim at fitting employment with another employer ('second track' reintegration).

taken. This report will be assessed by the Social Benefit Administration in the 91th week, when also the residual earnings capacity of the individual is established. Finally, at the end of the waiting period the individual can apply for (wage-related) DI benefits granted that (1) both employer and employee can show they have taken adequate reintegration measures and (2) the individual has a residual earnings capacity of less than 65% of his/her pre-disability earnings. In case the employer has not shown sufficient collaboration, the waiting period can be extended with one more year at maximum.

4.2.2 Private insurance of continued wage payments and case management

Employers can insure themselves against the risk of the continued wage payments during the waiting period via private insurers. Approximately 76% of Dutch employers has such an insurance (de Jong et al. 2014). The employees of the uninsured and insured firms are similar in terms of age and gender, however insured firms are usually smaller than the uninsured firms. 78% of firms with 2-10 employees has an insurance for continued wage payments, whereas only 27% of firms with more than 100 employees has such an insurance. For small employers the risk of continued wage payments is similar to large firms, the relative burden however is higher. Insurers can offer the possibility to not only insure wages, but also insure all the costs that come with the obligations of the Gatekeeper Protocol. At least 67% of the insured firms have such a 'broad' insurance (de Jong et al. 2014). One such obligation is to assign a case manager that serves as a link between all the parties involved and keeps track of the progress of the sick-listed employee.[6]

[6]There are many variations possible when it comes to these insurances. There is freedom of choice in the percentage of wages insured (77% of firms chooses to insure 100% of the wages paid in the first year and 70% of the wages paid in the second year of sick leave), firms can opt-in for a deductible (77% of firms chooses to keep two weeks to two months of sick leave on their own account), and firms can choose for a stop-loss insurance (only chosen by 5% of firms of which most are firms with more than 100 employees). Of the firms surveyed by de Jong et al. (2014) 9% answered that their insurance only covers continued wage payments, 67% answered their insurance covers

During the waiting period, the sick-listed employee is allowed to work partially. The employee can either do therapeutic work wherein he or she is considered as an extra pair of hands, or do graded work. In the latter case, the employee engages in productive work and the employer pays for those productive hours worked and the insurer only pays the hours foregone. For example, if an employee engages for 20% in graded work, he gets paid 100% of his pre-sickness wage of which 80% is covered by the insurer and 20% by the employer. As the case managers are hired by the insurer, they have a direct financial incentive to actively keep track of the individuals' residual earnings capacity and to try to get the individual to participate in paid work for as much as deemed possible. With full insurance and full sick pay coverage, direct financial incentives are obviously less strong for employers and employees, but they do have an interest in work resumption anyway. For employers, sickness absence may be costly for other reasons than wage continuation, non cooperation may lead to an extension of the waiting period, and potential DI benefit costs after the waiting period are experience rated. Moreover, non-cooperation with reintegration plans inhibits the risk of getting fired or loosing eligibility to DI benefits for sick-listed employees.[7]

The data used in this chapter come from a private workplace reintegration provider that is the sole provider of case management for two large insurers, together holding a market share of about 30% of the insurances for continued wage payments (Dutch Association of Insurers 2016). The workplace reintegration provider offers different types of products, from the registration of sickness absence to case management for individuals at risk of long term absence. In the current study, we focus on the individuals assigned to case management. Employers who take out the 'broad' insurance package with either of the two insurers are automatically directed to our workplace reintegration provider for case management. Those who are only insured against continued wage payments can opt

wage payments and the costs for gatekeeper obligations, 4% has some other type of package, and 19% does not know what their insurance covers.

[7]The evidence also confirms that private workplace reintegration providers usually increase reintegration activities in the waiting period (Everhardt and de Jong 2011). This suggests that the provision of insurance does not (fully) remove the incentive to achieve work resumption.

to work with a case manager from within their own company, hire an external case manager, or hire the services of the case manager of our workplace reintegration provider.

In a typical case management trajectory a sick-listed employee can be directed to our workplace reintegration provider when a disability assessment is made by the company doctor. When there is an indication for imminent long-term absenteeism at that time and the contract with the provider includes case management, the employee gets assigned to a case manager who establishes a more detailed diagnosis and writes the return to work plan. The assignment of sick-listed employees to case managers is based on caseload, i.e. the case manager that has time takes on the sick-listed employee. Stated differently, case managers are not specialized in specific health problems, sectors, or regions.[8]

The case managers working at our workplace reintegration provider are not doctors. Usually, case managers have a background in law, HR, or (para)medical care. They purely serve as a manager of the reintegration process: consulting with the occupational physician, keeping in regular contact with the employer and sick-listed employee, identifying the steps to be taken by the employer and employee, putting together the return to work plan, and administrating the process. Based on cost-benefit analysis case managers can decide to buy rehabilitation products from external parties, such as paramedical care, job training, and coaching. They do not provide this care themselves.

4.3 Data

4.3.1 Characteristics of sick-listed employees

We have access to all files on sick-listed employees that were assigned to case management at our private workplace reintegration provider between the years 2011 and 2014. We exclude those individuals that hold specific

[8]The workplace reintegration provider has only one office, located in the center of the country. Contact with the sick-listed employee is mostly maintained via phone and email.

insurance contracts, which include extra services before case management and/or earlier entry into case management (when there is not yet a risk of long term sickness). These contracts are predominantly held by self-employed.[9] The client files include characteristics like gender, gross (pre-sickness) wage earnings, and age. Moreover, they include the exact dates of the first sick day, of the entry day at the workplace reintegration provider, and of (partial) recovery. These files are merged to a file containing the interventions applied to each sick-listed employee and a file containing information on the assigned case manager. The data covers 11,741 sick-listed employees that are assigned to 68 case managers.

Table 4.1 shows the characteristics of the sick-listed employees in our sample. Almost half of the individuals is female and they are on average 42 years old. The time between sick-listing and the sick leave file arriving at the provider is on average nine weeks, whereas individuals are legally obliged to start their return to work activities by the eight week. Figure 4.2 shows that roughly half of the individuals do enter case management before the eighth week of sickness absence. However, it also shows that there is quite some spread in the moment at which the individuals start case management. As the elapsed duration until intake is likely to affect both the likelihood of graded work and work resumption, we should take this into account in our empirical analysis. We have no information on possible reintegration efforts by the employer and employee between the moment of sick-listing and the moment of entry at the workplace reintegration provider.

Individuals earn on average 255.86 euro a day and mostly work in small to average sized firms. 32.7% of the individuals has a general medical condition, 10.7% has physical problems, 20.5% has musculo-skeletal problems, 30.6% has psychiatric, psychological, or social problems, 4.0% has a conflict at work, and 1.5% has some kind of other condition, such as flue or complaints due to pregnancy. When it comes to general medical conditions one must think of individuals who are recovering from surgery or suffer from chronic illness. The average individual has

[9]Table 4.12 of Appendix 4.A shows the selection of our data in more detail. As becomes apparent from the table, we also exclude observations that were assigned to caseworkers with less than 25 clients in a particular year.

Table 4.1: Descriptive statistics sick-listed employees.

	all	no graded rtw	graded rtw	p-value[a]
number of sick-listed employees	11,741	4,504 (38.4%)	7,237 (61.6%)	
% female	47.3%	49.6%	45.9%	0.000
age at start case management	42.4	41.9	42.8	0.000
weeks until start case management	9.2	9.3	9.1	0.207
gross pre-sickness wage (euro/day)	255.86	235.12	268.76	0.458
firm size				
- 1 employee	15.2%	17.0%	14.1%	0.000
- 2 to 9 employees	36.3%	37.5%	35.5%	0.031
- 10 to 49 employees	35.8%	32.8%	37.7%	0.000
- 50 or more employees	2.6%	1.9%	3.1%	0.000
- number of employees unknown	10.1%	10.9%	9.5%	0.020
type of condition				
- general medical - mild	7.7%	10.9%	5.7%	0.000
- general medical - medium	13.5%	11.7%	14.7%	0.000
- general medical - severe	11.5%	10.5%	12.1%	0.007
- physical - mild	7.1%	6.9%	7.3%	0.395
- physical - severe	3.6%	3.3%	3.8%	0.127
- neck, shoulder, arm complaints	6.9%	5.6%	7.7%	0.000
- hip, ankle, knee complaints	6.3%	4.7%	7.4%	0.000
- back complaints	7.3%	6.2%	8.1%	0.000
- psychiatric	1.8%	1.9%	1.7%	0.442
- psychological - mild	11.4%	10.4%	12.0%	0.007
- psychological - severe	2.8%	2.6%	2.9%	0.303
- psychosocial - mild	10.7%	10.1%	11.0%	0.106
- psychosocial - severe	1.8%	1.4%	2.0%	0.004
- social problems	2.1%	2.1%	2.0%	0.751
- conflict	4.0%	8.6%	1.1%	0.000
- other[b]	1.5%	3.2%	0.4%	0.000
time spent on claimant (min/week)	17.0	23.1	13.2	0.000
weeks until closing of file	42.1	36.0	45.9	0.000
returns to work within one year	59.6%	53.6%	63.3%	0.000
returns to work within two years	76.7%	59.3%	87.6%	0.000

[a] Two-sided t-test on difference between sample with graded work and no graded work, with unequal variances.
[b] Other contains conditions such as flu and complaints due to pregnancy.

Figure 4.2: Histogram of application moments.

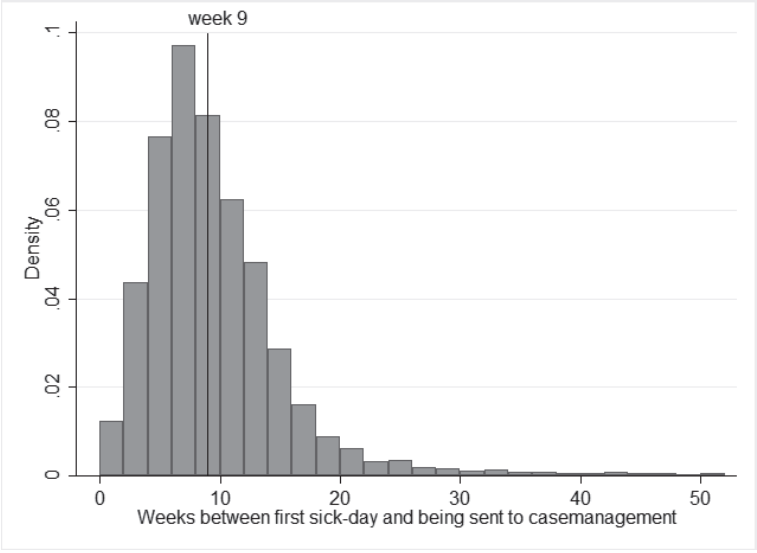
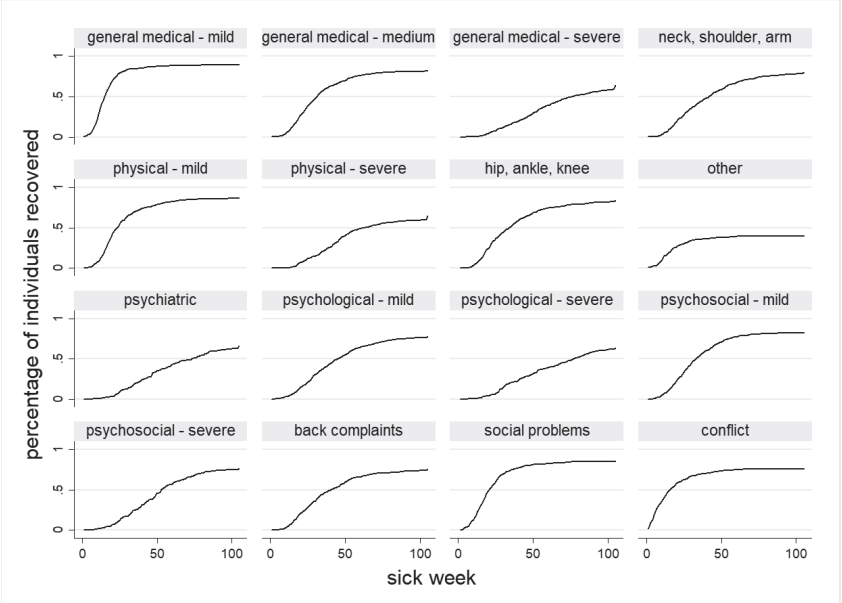


Figure 4.3: Recovery patterns by type of diagnosis

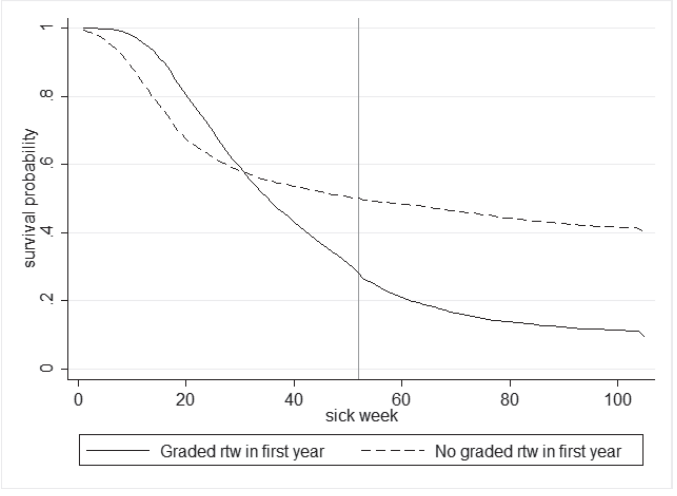


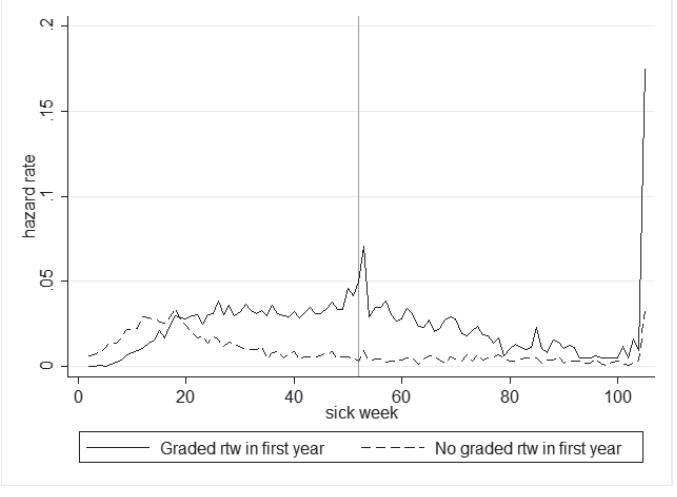
17 minutes per week allocated to him by the case manager. Individuals exit the trajectory on average after 42.1 weeks, with 59.6% of individuals returning to work within a year, and 76.7% of individuals returning to work within two years. Figure 4.3 shows the percentage of individuals that recovered in each sick week, stratified with respect to type of diagnosis. It should be stressed here that we only consider the individuals that were directed to the workplace reintegration provider after some period of sickness. As a result of this type of selection, recovery and work resumption rates remain close to zero in these first weeks. In line with expectations, we observe that individuals with less severe problems on average recover faster than those with more severe problems. The different type of musculo-skeletal problems (neck/shoulder/arm, hip/ankle/knee, and back) show similar recovery patterns.

Table 4.1 also shows the characteristics of the sick-listed employees for those who did and those who did not participate in a graded return-to-work arrangement. We define an individual to be in graded return-to-work when his wage value, e.g. the degree of pre-sickness productive work time resumption, exceeds 0%. Roughly 60% of the individuals in our data set participate in graded return-to-work at some point during their sick spell. The two groups are comparable in terms of age, gender, and moment of application; the differences in means are statistically significant in some cases, but not substantial. The graded individuals do not earn significantly more than the non-graded individuals. The compositions of the groups are slightly different when it comes to the diagnoses. For example, people who have a conflict at work rarely enter a graded return-to-work trajectory. Presumably, cooperation of the employer and possibly work place adaption is more troublesome in situations where there is a conflict.

Those in graded return-to-work have on average less time devoted to them by their case manager than those who are not in graded return-to-work. Despite the longer average sickness duration, those participating in graded return-to-work have a higher probability of returning to work in the longer run. This is also reflected in Figure 4.4 showing survival probabilities and hazard rates for individuals who started a graded return-

Figure 4.4: Survival and hazard rates for individuals with and without graded return-to-work in first year of absence





to-work in the first year of their sick leave and for individuals who did not start a graded return-to-work in the first year, respectively. Individuals participating in graded return-to-work have a lower probability to recover in the first weeks of illness, but start to perform better than those not participating in graded return-to-work from about the 25th week onward leading to substantially lower probabilities of non-recovery in the 70th week. From that point on the lines run roughly parallel to each other. The hazard rate spikes after the first year of sick-leave and at the end of the second year. These spikes correspond to the two annual evaluation moments in the Gatekeeper Protocol.

4.3.2 Characteristics of case managers

Table 4.2 shows case manager characteristics of our sample. We have information on 68 case managers, who are predominantly female (70.6%). They have on average about 68 sick-listed employees assigned to them per year. There is quite some spread however, with case managers treating up to 123 individuals a year at maximum. We dropped those case manager-years in which a case manager treated less than 25 individuals in a particular year.[10]

In principle individuals are assigned to case managers based on caseload. That is, new clients are directed to those who have time. However, there seems to be some clustering at certain case managers based on gender and type of diseases. More specifically, the spread of the case manager averages is relatively high for these variables. This could hint at some form of specialization, in the sense that case managers select those individuals that they know best how to deal with. However, when it comes to the diagnoses of the clients, the variation is more likely to be a result

[10] In the appendix, we present the results of robustness analyses that take different cutoffs (see Tables 4.20, 4.21, 4.22 and 4.23 in Appendix 4.B). When setting the cutoff too low, the average behavior of case managers with only a few clients is more likely to be a poor representation of grading practices. This will weaken the explanatory power of the instrument. When setting the cutoff too high, however, many observations need to be dropped, thus decreasing the efficiency of the estimations. As we will show, both the point estimates as the standard errors turn out to be hardly affected by the choice of cutoff.

Table 4.2: Descriptive statistics of the 68 case managers

	mean	sd	min	max
a. characteristics of case manager				
female	70.6%			
age on 1 Nov 2014	39.1	10.1	25	65
number of clients per year	68.4	23.1	25	123
b. characteristics of the clients of case managers				
fraction of clients female	48.5%	14.8%	20.9%	76.6%
average age at start of case management	42.4	1.7	37.6	46.1
weeks until start of case management	9.1	1.1	60.4	11.1
average gross pre-sickness wage (euro/day)	253.08	242.16	76.45	1317.26
median gross pre-sickness wage (euro/day)	108.12	5.04	84.36	110.00
fraction of clients from firm size categories				
- 1 employee	15.1%	5.7%	2.6%	30.6%
- 2 to 9 employees	36.5%	5.8%	24.0%	51.9%
- 10 to 49 employees	35.4%	7.8%	13.3%	56.0%
- 50 or more employees	2.8%	3.5%	0.0%	23.1%
- number of employees unknown	10.2%	3.4%	3.6%	18.3%
fraction of clients with condition type				
- general medical - mild	8.3%	6.8%	0.0%	28.8%
- general medical - medium	13.3%	6.0%	3.7%	41.0%
- general medical - severe	10.9%	4.6%	0.0%	25.3%
- physical - mild	7.4%	7.0%	0.0%	39.3%
- physical - severe	3.7%	2.8%	0.0%	17.6%
- neck, shoulder, arm complaints	6.7%	3.9%	0.0%	19.0%
- hip, ankle, knee complaints	6.4%	3.8%	0.0%	16.4%
- back complaints	7.2%	3.2%	0.0%	17.5%
- psychiatric	1.7%	1.4%	0.0%	6.3%
- psychological - mild	11.6%	7.8%	0.0%	40.7%
- psychological - severe	2.8%	3.0%	0.0%	19.3%
- psychosocial - mild	10.4%	7.2%	0.0%	33.1%
- psychosocial - severe	1.7%	1.8%	0.0%	8.5%
- social problems	2.3%	3.6%	0.0%	21.4%
- conflict	4.1%	2.5%	0.0%	11.9%
- other[a]	1.5%	2.0%	0.0%	8.0%
c. activities and results of case managers				
fraction of clients in graded work	60.2%	8.2%	33.9%	77.4%
average time allocated to client (min/week)	17.0	3.1	10.6	28.4
average weeks until closing of file	41.0	6.2	21.2	57.2
fraction of clients rtw within one year	60.8%	10.3%	23.3%	92.0%
fraction of clients rtw within two years	76.9%	8.5%	40.7%	94.1%

[a] 'Other' contains conditions such as flu and complaints due to pregnancy.

of the reporting behavior of the case managers than reflecting selection. This is because the diagnoses are established by the case managers after the clients are assigned to them. The results from the sensitivity checks reported in Section 4.5.4 will show that our results are unlikely to be driven by potential specialization of case managers.

Case managers differ substantially in their use of graded return-to-work, with some case managers only having 33.6% of their clients in graded return-to-work and others having up to 82.6% of their clients participating in graded return-to-work. Also the average time allocated to the individuals vary greatly among case managers.

4.3.3 Setup of graded return-to-work trajectories

Within the group of clients that started a graded work trajectory, relevant outcomes measures are the moment and the degree at which grading is started. The variable 'wage value', which we use to construct our graded return-to-work index, may contain any integer value ranging from 0 to 100 and can be updated up to 24 times at maximum in a two-year-trajectory. Case managers are encouraged to fill in the variable succinctly, as any degree of work resumption implies lower costs for the workplace reintegration provider. The extent to which we can use this detailed information depends on the variation in the graded return-to-work trajectories. In this section we explore the different trajectories in detail.

Figure 4.5 shows the percentage of individuals participating in graded return-to-work in a certain week, where we define five categories of graded work: 1-20%, 21-40%, 41-60%, 61-80%, and 81-100% of the pre-sickness wage value, respectively.[11] The figure shows that in the first weeks of sickness individuals usually work modest amounts of time (21-60% graded work). Towards the 20th week, individuals participate more often in high degrees of graded work resumption (81-100%) or very low degrees (<20%).

[11] When calculating this percentage, we include individuals from the first sick day up until the end of the 105th sick week (so also after recovery). As a result, the numerator remains unchanged.

Figure 4.5: Percentage of individuals participating in graded return-to-work per week.

Table 4.3: Variation in grading practices across case managers.

	mean	sd	min	max
average weeks waited until start graded rtw	20.85	2.83	12.56	25.92
average degree of grading at start graded rtw	36.01%	4.24%	28.26%	55.15%
fraction of graded rtw that started:				
1 - 8 weeks	13.90%	5.82%	3.85%	31.34%
9 -16 weeks	35.14%	6.39%	22.95%	55.56%
16 - 24 weeks	22.42%	6.07%	8.96%	36.84%
24 - 32 weeks	11.97%	3.84%	3.70%	23.08%
after 32 weeks	16.56%	6.51%	0.00%	28.32%
fraction of graded rtw started at a grade between:				
1 - 20% of pre-sickness wage value	26.4%	8.5%	0.0%	47.4%
21 - 40% of pre-sickness wage value	34.6%	7.3%	7.1%	60.0%
41 - 60% of pre-sickness wage value	31.3%	8.7%	17.9%	78.6%
61 - 80% of pre-sickness wage value	4.0%	2.9%	0.0%	15.6%
81 - 100% of pre-sickness wage value	3.7%	2.9%	0.0%	14.3%

In the later weeks (when most have recovered), those who are still in graded return-to-work mostly work modest amounts of time, i.e. $< 20\%$ graded work resumption.

Table 4.3 shows the variation in grading practice of the different case managers. On average case managers wait 20.85 weeks before starting the graded return-to-work and do so at a degree of 36.01%. The fastest case manager waits on average 12.56 weeks and the slowest 25.92. The case manager that starts grading at the lowest degree does so at 28.26% on average and the one that starts the highest does so at 55.15% on average. There are some case managers that never start a graded return-to-work arrangement after 32 weeks, while others start almost a third of the trajectories that late. Also, some case managers never start a graded return-to-work arrangement at 1-20% of pre-sickness wage value, whereas others start almost half the arrangements at this level. We thus conclude there is quite some variation in the grading practice of the different case managers.

4.4 Estimation strategy

To identify the effectiveness of graded return-to-work at reducing the length of sick spells, we use an instrumental variable (IV) method which was introduced by Duggan (2005). Duggan analyzes how expenses on new drugs affect total medical expenditures by exploiting the variation in psychiatrists' preferences in drugs prescription as an instrument for individual expenses on certain types of new drugs. In a similar fashion, more recent applications exploit variation in strictness of disability examiners and judges in awarding disability benefits (French and Song 2014, Maestas et al. 2013) and the propensities of employment offices or individual caseworkers to use certain interventions (Dean et al. 2015, Markussen and Røed 2014, Markussen et al. 2017, Rehwald et al. 2016). Our approach is most similar to Markussen et al. (2012), who exploit variation in physicians' use of graded absence certificates to identify the effect of part-time sick leave on absence duration.

In our case, employees are send to the reintegration provider after some weeks of absence. The provider assigns them to case managers that have substantial discretionary room in choosing specific treatments. Case managers are encouraged to use graded return-to-work whenever possible. However, the actual grading practice may vary among the case managers. First, this is because different case managers may make different assessments of when an individual is ready to start graded return-to-work and the individuals' ability to work. Second, one cannot simply assign an individual to graded return-to-work in all relevant work environments. The case manager has to negotiate the possibilities of adapted work duties with the employer, who is not always willing to allow for such flexibility (Wainwright et al. 2011).[12] One case manager may be better in this negotiation process than the other, speeding up the process towards graded return-to-work. Hence, whether an individual participates in graded return-to-work and when he starts to do so, may depend on the case manager he is assigned to. This means the case manager's *propensity to grade* can be used to instrument the graded return-to-work variable.

Within the context of the current analysis, the validity of instrumental variables estimation essentially requires four conditions to be met. First, the probability of graded work should be affected by the concerning case managers' propensity to use a graded work for all other individuals that are assigned to him ('relevance'). In light of the time span of four years that is covered, assuming the tendency to use graded work to be constant over time may be too restrictive. Case managers may change their behavior over time, as they may learn from earlier experiences. We therefore construct propensities by case manager for each year in our sample. This also potentially increases the efficiency of our estimates.[13]

Our second condition for IV to work is that sick listed individuals are assigned randomly to case managers. Stated differently, this implies that sick-listed individuals with long and short expected sick durations

[12]When performing a decomposition analysis of the observed variation in graded-work applications across case managers and employers, we see indeed that the individual's employer is more important than the individual's case manager. As long as individual's are randomly assigned to case managers, however, this does not burden our analysis. At most, it decreases the efficiency of our method.

[13]At the same time, the sample size per case manager per time unit should be sufficient.

do not cluster among certain case managers. With the information on sick-listed workers in our data, we can test for randomness by excluding client characteristics in our model. If this yields different coefficient values for graded work, this suggests there is clustering on worker types. In a similar vein, we can re-run the analyses while excluding case managers who have abnormal client group compositions. The results of both of these analyses are reported in Section 4.5.4. Obviously, testing for clustering on unobservable characteristics is more complex, but it should be stressed that case managers did not receive more information than the registered data we have. This renders it plausible there was no selection on unobservables.

Third, we rely upon the assumption that graded work effects are not correlated with the general ability of case managers in getting individuals back to work (i.e., the 'exclusion restriction'). For instance, if high quality case managers have a strong tendency to use graded work, the IV model will overestimate the effect of graded work. We therefore will conduct various sensitivity tests that use proxies for the overall quality of case managers. In particular, such proxies include both current and lagged work resumption rates for clients that were assigned to case managers or work resumption rates for individuals that were on graded work already at the moment of intake. The results of these analyses are reported in Section 4.5.4

Finally, individuals who would not be treated by a high propensity case manager, should also not be treated by a low propensity case manager. This monotonicity assumption implies that the graded work propensities should impact all individuals equally in our sample. For instance, this assumption may be violated if one case manager is more inclined to use graded work for individuals with mental issues, but less inclined to use graded work for individuals with musculo-skeletal problems. With this in mind, we will conduct tests on the equality of graded work propensity impacts on the actual use of graded work – i.e, the first stage estimates. The results of these checks are reported in Section 4.5.4.

Specification of the effect of graded work 4.4.1

When specifying the IV model that estimates the effect of graded-work on the incidence of work resumption and the number of sickness weeks, we closely follow Markussen and Røed (2014) and Rehwald et al. (2016). In these analyses, the aim is to estimate the effect of the provision of graded work (G). As we will show later on, we extend their analysis in two ways. First, we will develop propensities for the weeks waited until the start of graded work (W). For the individuals with graded work, this enables us to estimate the impact of the timing of graded work on full work resumption. Second, we will focus on the impact of the level of graded work at the start of a graded work trajectory (S). For ease of exposition, we will consider a single time period for which we construct case manager propensities. As argued above, we can extend this by allowing for case manager propensities for each year in our sample.

To start with, we structure the cross sectional data on the sick-listed individuals to a panel where every period t corresponds to one week. We include all individual-weeks in the first year of the sick-spell up to and including the week in which graded work started or, in case of the absence of a graded work treatment, until the sick spell ended (i.e., individual went back to work or entered the DI scheme). Then, we run an OLS regression on a dummy indicating whether the individual is or is not starting to participate in graded work that week. In this regression we control for time constant individual characteristics x_i for individual i (e.g. age, age squared, sex, sick type, log gross (pre-sickness) wage, log gross (pre-sickness) wage squared, firm size, year of application, type of insurance contract, sick duration until application at the re-integration office), together with period dummies ($date_{it}$), and dummies for all possible outcomes of elapsed sick weeks (d_{it}):

$$
\begin{aligned}
graded_{ijt} &= \mathbf{x}_i' \, \boldsymbol{\theta}^g + \delta_1^g \, d_{it} + \delta_2^g \, date_{it} + u_{ijt}^g, \qquad (4.1)\\
&\quad i = 1, \ldots n \;\; \text{(individuals)},\\
&\quad j = 1, \ldots J \;\; \text{(case managers)},\\
&\quad t = 1, \ldots T \;\; \text{(periods)},
\end{aligned}
$$

where u_{ijt} is i.i.d. and clustered at the level of case manager-year combinations. The parameters θ^g, δ_1^g and δ_2^g describe the effects of individual characteristics, the elapsed sick weeks and period dummies, respectively.

Using the estimated individual errors $\hat{u}_{ijt}^g$, we next construct the case manager propensities to treat ψ_i^g. We sum the errors over the periods for every individual i, i.e.

$$\hat{u}_{ij}^g = \sum_{t=1}^{T_i} \hat{u}_{ijt}^g, \tag{4.2}$$

where T_i is the last period individual i is at risk of making a transition into treatment. Following Markussen and Røed (2014) and Rehwald et al. (2016), $\hat{u}_{ij}^g$ can be interpreted as the difference between the duration until treatment of individual i and the average duration until treatment for individuals with the same pre-treatment characteristics as individual i. We next take the average of all $\hat{u}_{ij}^g$ per case manager, while leaving out $\hat{u}_{ij}^g$ for the sick-listed employee concerned, i.e.

$$\psi_i^g = \frac{1}{n_j - 1} \sum_{k \in N_j^{-i}} \hat{u}_{kj}^g, \tag{4.3}$$

where N_j is the set of individuals corresponding to case manager j. For ease of interpretation, we rescale these ψ_i^g from 0 to 1, with 0 indicating the lowest propensity to use graded work and 1 indicating the highest propensity to use graded work.

In order to estimate the effect of graded return-to-work on the probability to return to work (y_i), we collapse the data to one observation per individual. This observation may either be the probability of work resumption or the number of weeks that have been worked over a certain time window. We estimate the effect of having participated in graded return-to-work on the return-to-work probability, using the propensity to grade (ψ_i^g) as an instrumental variable. We control for the same individual characteristics as in the propensity regressions. This yields the following

IV model:

$$y_i = \mathbf{x}_i' \, \boldsymbol{\beta}^g + \gamma^g \, \widehat{G}_i + \epsilon_i^g, \tag{4.4}$$

$$G_i = \mathbf{x}_i' \, \boldsymbol{\pi}^g + \alpha^g \, \psi_i^g + \eta_i^g. \tag{4.5}$$

4.4.2 Specification of the effect of timing and initial degree of graded work

Following the IV estimation procedures as in equations (1) to (5), the variation in graded-work propensities of case managers that we exploit essentially stems from two sources. First, case managers show differences in the likelihood of starting graded work interventions. Second, there is variation in the timing of treatments across case managers for those individuals that start graded work. To estimate the isolated impact of the duration until graded return-to-work on absence duration, we select only those individuals that enter graded return-to-work at some point during their sickness absence, next recalculate the propensities as explained in equations (1), (2), and (3) and denote these as ψ_i^w. Next, for this sub-sample of individuals, we define the variable W_i as the number of weeks until the start of graded return-to-work for individual i and estimate the effect of this variable on the absence durations using ψ_i^w as an instrument:

$$y_i = \mathbf{x}_i' \, \boldsymbol{\beta}^w + \gamma^w \, \widehat{W}_i + \epsilon_i^w, \tag{4.6}$$

$$W_i = \mathbf{x}_i' \, \boldsymbol{\pi}^w + \alpha^w \, \psi_i^w + \eta_i^w. \tag{4.7}$$

As with any IV model, it is important to stress at this point that our parameter of interest in the above equation, γ^w, should be interpreted as a local average treatment effect (LATE). This parameter denotes the effect of waiting one week extra before starting the trajectory on absence duration for those individuals. This result does not necessarily extrapolate to all individuals or to the whole support of the weeks waited variable, W_i.

Our data also allow us to focus on the tendency of case managers to start graded work at a high or low degree. For this purpose, we calculate a propensity based only on the percentage of pre-sickness hours worked

during the first week of graded return-to-work, i.e. the starting level denoted by S_{ij}, for the selected sample of individuals with graded work. We estimate a regression corresponding to equation (1),

$$S_{ij} = \mathbf{x}_i' \, \boldsymbol{\theta}^s + \delta_1^s \, d_i + \delta_2^s \, date_i + u_{ij}^s, \tag{4.8}$$

$$i = 1, \ldots n \quad \text{(individuals)},$$

$$j = 1, \ldots J \quad \text{(case managers)}, \tag{4.9}$$

where u_{ij}^s is i.i.d. and clustered with respect to case manager-year combinations. Based on the outcomes of this regression, we calculate similar propensities as in equation (4.3) for individual i with case manager j. We denote these as ψ_i^s. We instrument the initial degree of grading with the average initial degree of grading for all other sick listed workers that were assigned to this case manager. This enables us to conduct an IV regression as above using the degree of graded work resumption rate at the start of graded return-to-work as the intervention:

$$y_i = \mathbf{x}_i' \, \boldsymbol{\beta}^s + \gamma^s \, \widehat{S}_i + \epsilon_i^s, \tag{4.10}$$

$$S_i = \mathbf{x}_i' \, \boldsymbol{\pi}^s + \alpha^s \, \psi_i^s + \eta_i^s. \tag{4.11}$$

4.5 Results

4.5.1 Overall effects of graded return-to-work

Table 4.4 shows the effects of graded return-to-work trajectories on (1) a dummy variable indicating whether the sick-listed employee returned to work within 1 year; (2) a dummy variable indicating whether the sick-listed employee returned to work within 2 years; (3) the number of weeks worked in the first year; (4) the number of weeks worked in the first two years. Panel a shows the OLS results, panel b shows the IV results and panel c shows the reduced form or 'Intention-to-treat' estimates for the case manager propensity measure. The results for the regressions underlying

the propensities and the estimated coefficients for the control variables of the regressions are shown in Tables 4.15 and 4.16 of Appendix 4.B.

Columns 1 to 4 of Table 4.4 present the baseline results, where we consider an individual as treated if he enters a graded return-to-work trajectory within the first year of sick leave.[14] Based on the OLS results, one would conclude that graded return-to-work trajectories have substantial and positive rehabilitation effects. The IV estimates however show only moderate and statistically insignificant effects, suggesting positive selection into the treatment. This is best illustrated by the outcomes at the end of the second year. The OLS estimates indicate a 30 percentage point increase in return to work probabilities for individuals on a graded return-to-work trajectory, whereas the IV estimates show only a 7.5 percentage point (insignificant) increase. Similarly, the reduced form estimates indicate that individuals assigned to a case manager with the highest propensity to use graded return-to-work are only 2 percentage point (insignificant) more likely to rehabilitate within two years than those assigned to the case manager with the lowest propensity to use graded work.

Columns 5 to 8 of Table 4.4 show the results when only considering graded return-to-work trajectories which started in the first 26 weeks of sick leave as a treatment (individuals who started a graded return-to-work trajectory after the 26 weeks are considered untreated). Compared to the earlier results with 52 weeks as a maximum, there are noticeable differences in the effects. The probability to return to work increases with 30.8 percentage point compared to 12.7 percentage point and the number of weeks worked increases with 8.9 weeks compared to 1.2. One explanation for this difference in outcomes is that graded return-to-work trajectories are more effective when started earlier, which is the hypothesis we will further explore in Section 4.5.2. Another explanation is that there is a lock-in for graded return-to-work trajectories that occurs in the first

[14] 5.3% of untreated individuals do start a graded return-to-work trajectory in the second year of sick leave. Since these trajectories start later in time than outcome variables (1) and (3), we consider these individuals as untreated. When we do consider them as treated and estimate the effects at the end of the two year waiting period, outcome variables (2) and (4), we find slightly smaller effects: return to work probabilities increase by 0.0488 (0.112), the number of weeks worked increases by 1.728 (8.680).

Table 4.4: Overall effects of graded return-to-work on full work resumption

Intervention:	Graded rtw started in week 1-52				Graded rtw started in week 1-26			
	Returned to work		Weeks worked in		Returned to work		Weeks worked in	
	1 year	2 years	week 1-52	week 1-104	1 year	2 years	week 1-52	week 1-104
a. OLS estimates								
Graded rtw	0.184***	0.300***	0.251	14.98***	0.280***	0.225***	4.865***	17.78***
	(0.010)	(0.009)	(0.287)	(0.719)	(0.009)	(0.008)	(0.264)	(0.636)
n	11,741	11,741	11,741	11,741	11,741	11,741	11,741	11,741
R^2	0.198	0.181	0.296	0.244	0.239	0.131	0.319	0.262
b. IV estimates								
Graded rtw	0.127	0.075	1.173	6.642	0.380***	0.070	8.901**	18.30**
	(0.122)	(0.109)	(3.581)	(8.531)	(0.125)	(0.104)	(3.759)	(8.624)
n	11,741	11,741	11,741	11,741	11,741	11,741	11,741	11,741
R^2	0.195	0.117	0.296	0.231	0.230	0.101	0.303	0.262
stage 1: Ψ[graded rtw]	0.270***				0.268***			
	(0.0268)				(0.0267)			
c. Reduced form estimates of propensity								
Ψ[graded rtw]	0.034	0.020	0.317	1.793	0.102***	0.019	2.386**	4.907**
	(0.033)	(0.030)	(0.970)	(2.333)	(0.035)	(0.028)	(1.014)	(2.372)
n	11,741	11,741	11,741	11,741	11,741	11,741	11,741	11,741
R^2	0.167	0.068	0.296	0.201	0.168	0.068	0.297	0.202

Control variables include gender, age, wage, sick weeks until application, year dummies, medical conditions, contract types, and firm size.
Claimants are excluded when their assigned case manager treated fewer than 25 claimants in the same year as the claimant.
Clustered (case manager - year) standard errors between parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

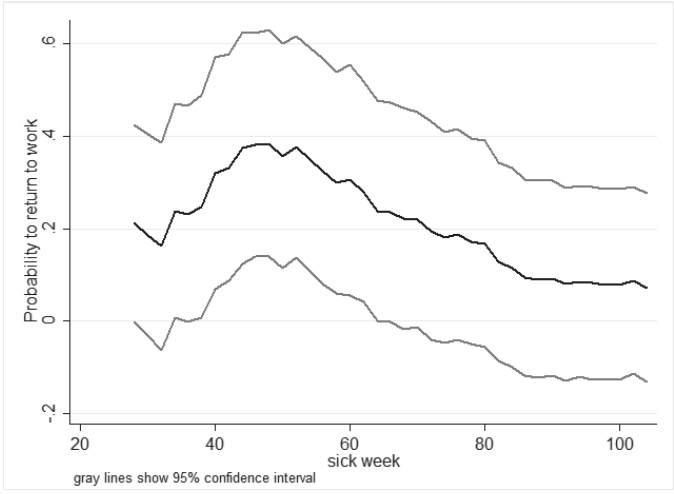
weeks of grading. If so, we would expect differences in effectiveness of graded work to fade out over time. This is confirmed when comparing the long-term effects that are shown in column 2 and 6 of Table 4.4.

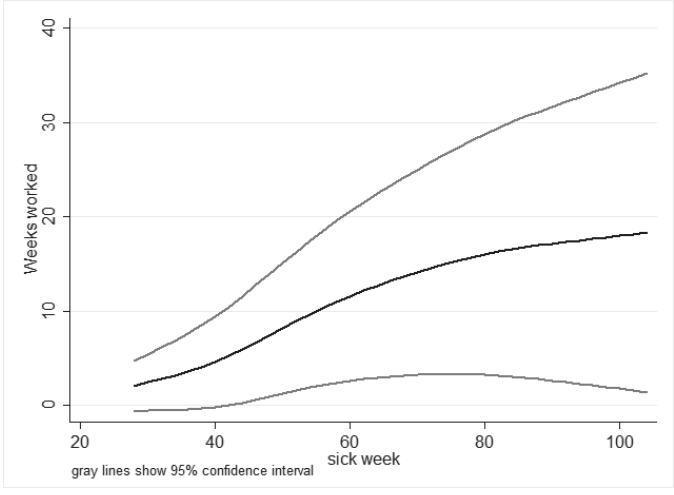
To illustrate the evolution of the effects in more detail, Figure 4.6 shows the effects of graded return-to-work trajectories that started in the first half year on the return to work probability as well as the number of weeks worked. The effect on the return-to-work probabilities is increasing up to week 46, after which the effect declines. It appears that graded return-to-work speeds up the recovery process, with the return-to-work probabilities being almost equal after two years. In line with this, the steep increase in weeks worked between week 40 and 60 does not persist, such that the line flattens out towards the end of the second year.

The effect of graded return-to-work spells started in the first half year on weeks worked in the first year is comparable to the effect found in Markussen et al. (2012) with data from sick-listed workers in Norway. They find that part-time sick leave decreases the absence spells with eight to ten weeks. Rehwald et al. (2016) find substantially bigger results, amounting to a 30 week increase in weeks in regular employment in the first year.[15] Contrary to our results, both Markussen et al. (2012) and Rehwald et al. (2016) find positive long run effects. The first shows that employment two years after sick listing increases with 16 to 21 percentage point, the latter finds a increase of 27 weeks worked during the second year and an increase of 26 weeks in the third year. When comparing these outcomes with ours, one should bear in mind that employers in the Netherlands are committed to facilitate the return-to-work for the sick-listed workers for at least two years. Accordingly, we may expect that individuals in the control group – i.e., those without graded return-to-work – are likely to receive other services. This in turn may explain why the long-term impacts we find are smaller and insignificant. Still, our evidence also suggests that graded return-to-work may speed up the recovery process, particularly when starting early.

[15]Markussen et al. (2012) only consider grading decisions made within the first eight weeks of sick leave. In the field experiment of Rehwald et al. (2016), graded return-to-work should be started within four weeks after a meeting which is held in the first eight weeks.

Figure 4.6: Cumulative effects of graded work per sick weeks





4.5.2 Effects of the timing and initial level of graded work

We argued earlier that both the timing and the initial level of graded work may determine the effectiveness of graded work trajectories. To investigate the importance of these two parameters, we select the sample of individuals who entered a graded return-to-work trajectory in the first year of their sick leave. Using a similar setup as for our benchmark model, we first estimate the effect of starting a graded return-to-work trajectory one week later on the outcome measure. The results are reported in Panel a of Table 4.5. The first stage results show that being assigned to the case manager with the highest propensity of graded work leads to a four week reduction in waiting time until graded return-to-work, as compared to the case manager with the lowest propensity. The second stage results indicate that waiting one week extra before starting graded return-to-work, decreases the probability to rehabilitate within one year with 4.4 percentage point, whereas the probability to return to work within two years is not affected. This again suggests that graded return-to-work speeds up the recovery process, rather than increasing the long-term probability of recovery. Starting graded return-to-work one week later results in a trajectory that lasts half a week longer, so that the number of weeks worked in the first year decreases by 1.5. When taking a time horizon of two years, the number of weeks worked even decreases by 2.2.

In panel b of Table 4.5 we consider the effect of the level of work resumption at the start of the graded return-to-work trajectory on work resumption. In this setup, the instrument orders case managers in terms of their preference to start a return-to-work trajectory at high levels of graded work resumption. As the table shows, being assigned to a case manager that tends to start trajectories at high rates rather than to one that tends to start at low rates, increases the starting level of work resumption by 25 percentage point. From the second stage estimates we infer that starting at a 10 percentage point higher level of grading results in a 6 percentage point higher chance of recovering in the first year.[16] The return to work

[16]Note that more than 90 % of the trajectories have an initial degree of graded work that is less than 60%. Accordingly, variation in the degrees we study typically reflects differences between one, two or three days of working at the start of graded work.

Table 4.5: Effect of starting graded return-to-work one week later or at a higher starting level: IV estimates.

	Returned to work		Weeks worked in	
	1 year	2 years	week 1-52	week 1-104
a. Duration until start of graded return to work trajectory				
Sick weeks until grading start	-0.044***	-0.001	-1.497***	-2.177***
	(0.010)	(0.005)	(0.257)	(0.489)
n	5,906	5,906	5,906	5,906
R^2	-0.028	0.033	0.105	0.111
stage 1: ψ[weeks waited]	-3.935***			
	(0.791)			
b. Level of work resumption at start (linear specification)				
Starting level (0-100)	0.006***	0.003**	0.135***	0.318***
	(0.002)	(0.001)	(0.052)	(0.111)
n	5,913	5,913	5,913	5,913
R^2	0.142	0.018	0.314	0.199
stage 1: Ψ[degree grading]	25.37***			
	(0.564)			

Control variables include gender, age, wage, sick weeks until application, year dummies, medical conditions, contract types and firm size.
Propensities are calculated on the sample of graded individuals. Claimants are excluded when their assigned case manager graded fewer than 25 claimants in the same year as the claimant.
Clustered (case manager - year) standard errors between parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

probability for the first two years is increased by 2.5 percentage point. The number of weeks worked in the first year increases by 1.4, whereas individuals work 3.2 additional weeks in the first two years. This suggests that a higher initial level of grading also improves long-term recovery rates. It may thus be that giving sick-listed individuals an easy start by re-introducing them to work for a very limited amount of hours, may actually harm them. It may be that the potential positive effects of graded work cannot be established if the individual cannot properly participate in work processes and is not viewed as a full-fledged employee.

Effects for different types of medical conditions 4.5.3

Table 4.6 shows IV estimates for samples of specific medical conditions that are registered by the reintegration provider. Panel a shows the baseline estimates for all graded return-to-work trajectories and panel b those for all graded return-to-work trajectories started in the first 26 weeks. The first stage estimation results are similar in size across medical conditions, suggesting that the extent to which case managers can affect the use of graded work is equal across groups in the first 26 weeks of absence. The second stage estimates however vary across medical conditions. While graded return-to-work increases first year return-to-work probabilities substantially for general medical as well as musculo-skeletal problems, it seems to have little effect on workers with mental problems. This corresponds with the findings of Høgelund et al. (2010) and Andren (2014) who both find no effects of graded return-to-work for individuals with mental disorders, but positive effects for individuals with other disorders. Also Hernæs (2017) finds larger effects for individuals with musculo-skeletal problems, than for individuals with psychological problems. After two years of sickness, the effect for individuals with musculo-skeletal problems tends to zero, whereas the effect for individuals with general medical problems remains high. This suggests that graded work can be meaningful for individuals with chronic illnesses or individuals that recover from medical treatments.

Table 4.6: IV estimation results on work resumption for different medical conditions.

	General medical		Musculo-skeletal		Mental	
	Returned to work		Returned to work		Returned to work	
	1 year	2 years	1 year	2 years	1 year	2 years
a. Overall effect: trajectories started week 1-52						
Graded rtw	0.572*	0.563**	0.477	-0.203	-0.0234	-0.108
	(0.327)	(0.244)	(0.540)	(0.413)	(0.352)	(0.373)
stage 1:	0.191***		0.155		0.170**	
Ψ*[graded rtw]*	(0.072)		(0.095)		(0.074)	
b. Overall effect: trajectories started week 1-26						
Graded rtw	0.789***	0.468**	0.539*	-0.061	0.051	-0.259
	(0.238)	(0.205)	(0.323)	(0.261)	(0.261)	(0.296)
stage 1:	0.281***		0.229***		0.266***	
Ψ*[graded rtw]*	(0.066)		(0.076)		(0.079)	
c. Duration until start trajectory						
Sick weeks until	-0.040**	-0.0003	-0.016	0.016	-0.010	0.030*
start grading	(0.017)	(0.008)	(0.011)	(0.010)	(0.018)	(0.017)
stage 1:	-5.025***		-5.179***		-3.559**	
Ψ*[weeks waited]*	(1.704)		(1.738)		(1.402)	
d. Initial degree of grading						
Starting level (%)	0.005*	0.002	0.010***	0.003	0.007***	0.001
	(0.002)	(0.001)	(0.003)	(0.002)	(0.002)	(0.001)
stage 1:	30.59***		23.31***		26.71***	
Ψ*[degree graded]*	(0.859)		(1.141)		(0.841)	

The group 'general medical' consists of individuals with the conditions general medical - mild/medium/severe. The group musculo-skeletal consists of individual with the conditions neck, shoulder, arm, hip, ankle, knee or back complaints. The group mental consists of individuals with the conditions psychiatric, psychological - mild/severe, psychosocial - mild/severe or social problems. Individuals with physical mild/severe conditions are not considered because of the small sample size. Also individuals labels as 'other' or having a conflict are excluded. Control variables include gender, age, wage, sick weeks until application, year dummies, medical conditions, contract types and firm size. Claimants are excluded when their assigned case manager treated fewer than 10 claimants of the same type in the same year as the claimant. Panels a and b are based on 3,971 observations with general medical conditions, 1,947 with musculo-skeletal conditions, and 3,380 with conditions related to mental health. Panels c and d are based on 1,667 observations with general medical conditions, 982 with musculo-skeletal conditions, and 1,807 with conditions related to mental health. Clustered (case manager - year) standard errors between parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Panels c and d of Table 4.6 show the effects of the timing and initial level of graded work on work resumption for different medical conditions, respectively. Starting the trajectory one week later decreases the probability to return to work for individuals with general medical problems within one year with 4 percentage point, but the effect becomes small and insignificant after two years of absence. For musculo-skeletal conditions we do not find any significant effects on the probability to return to work. There also is no evidence of effects of starting later in the short run for individuals with mental conditions. In the long run, it even seems harmful to start graded work early for this group. In particular, the effect of starting the graded return-to-work trajectory one week later amounts to an increase in the probability to return to work within two years that is equal to about 3 percentage point. Finally, starting at a higher initial level of graded work resumption results in higher probabilities to return to work after one year and no significant effect after two years for all three types of medical conditions. The results for weeks worked correspond to the results for the return-to-work probability and can be found in Table 4.19 of Appendix 4.B.

Sensitivity tests 4.5.4

Endogeneity and specialization

We stated earlier that new clients were assigned to case managers based only on their caseload. As a result, there would be no specialization of case managers that could result in a positive correlation between the propensity to grade and the likelihood to return to work for reasons other than graded return-to-work itself. Based on observed characteristics of clients that are assigned to the same case managers, we can test for this assumption and biases that may stem from specialization. At the same time, we should bear in mind that for analyses based on samples of graded individuals only, there may be non-random selection – even if there is no specialization to start with. This may occur if for instance case managers that only grade few individuals do this because they only grade the most easy cases, which are also easy to grade early and at a high starting degree.

In that case, there will be a positive correlation between the weeks waited propensity (or the degree grading propensity) and the likelihood to return to work, for other reasons than starting the trajectory early (or starting at a high degree).

To ensure that potential non-random distribution of clients over case managers does not affect our results, we run a set of sensitivity analyses which are reported in Tables 4.7 and 4.8. As the most prominent effects on graded work were found in the first year of absence, we focus on the return-to-work dummy within one year as the variable that is to be explained. First, we re-run the regressions while excluding specific sets of covariates. If these covariates are correlated with both the probability to recover as well as the propensity to grade, the baseline analysis is subject to omitted variable bias. We exclude sick types in column (2), sick weeks until application in column (3), and all covariates expect the time dummies in column (4). Overall, the results are similar to the baseline. As for the overall effect estimates for trajectories that started in weeks 1 to 52, we find the most substantial difference from the baseline occurs when excluding all variables, with a decrease in the point estimate from 0.127 (0.122) to -0.0270 (0.132); see table 4.7 panel a. Coefficient estimates are hardly affected when we concentrate on trajectories starting in the first 26 weeks of absence; see table 4.7 panel b. It is only when we exclude the sick type dummies that the point estimate increases from 0.380 to 0.477. To put these findings in perspective, it is important to bear in mind that sick type is not known at the start of the case management trajectory, but determined by the case manager after the client is assigned to him. Differences in reporting a condition for example as general medical or physical, may be influenced by case manager beliefs. These beliefs may in turn be correlated with the propensity to grade. A similar explanation may also hold for the change in the effect estimate of the initial degree of grading that occurs when we exclude sick type dummies – see table 4.8 panel b. In all cases, the sizes of the difference in point estimates do not lead to concerns about the validity of our approach.

As a second sensitivity check, we exclude case managers with abnormal client group compositions. We define a group composition by the

Table 4.7: Sensitivity tests for specialization effects – Return-to-work within one year – Overall effect

Dependent: **rtw within 1 year**	(1) Baseline	(2) Exclude covariates	(3) Exclude covariates	(4) Exclude covariates	(5) Exclude abnormal groups	(6) Exclude abnormal groups	(7) Include
		Sick type	Weeks until application	All except year dummies	>3 sd from mean	>2 sd from mean	graded work propensity
a. Overall effect: trajectories started in week 1-52							
Graded rtw	0.127	0.130	0.099	-0.027	0.253**	0.263	
	(0.122)	(0.137)	(0.124)	(0.132)	(0.118)	(0.279)	
stage 1: Ψ*[graded rtw]*	0.270***	0.254***	0.266***	0.271***	0.285***	0.233***	
	(0.027)	(0.024)	(0.026)	(0.023)	(0.032)	(0.050)	
b. Overall effect: trajectories started in week 1-26							
Graded rtw	0.380***	0.477***	0.370***	0.391***	0.331***	0.340*	
	(0.125)	(0.140)	(0.121)	(0.138)	(0.093)	(0.192)	
stage 1: Ψ*[graded rtw]*	0.268***	0.239***	0.274***	0.255***	0.301***	0.322***	
	(0.027)	(0.024)	(0.028)	(0.025)	(0.027)	(0.061)	

Claimants are excluded when their assigned case manager treated fewer than 25 claimants in the same year as the claimant.
Panels a and b are based upon 11,741 observations of which 8,464 remain in column (5) and 3,807 in column (6).
Clustered (case manager - year) standard errors between parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.8: Sensitivity tests for specialization effects – Return-to-work within one year – Timing and intensity

Dependent: **rtw within 1 year**	(1) Baseline	(2) Exclude covariates	(3) Exclude covariates	(4) Exclude covariates	(5) Exclude abnormal groups	(6) Exclude abnormal groups	(7) Include
		Sick type	Weeks until application	All expect year dummies	>3 sd from mean	>2 sd from mean	graded work propensity
a. Duration until start trajectory							
Weeks waited	-0.044***	-0.040***	-0.050***	-0.043***	-0.037***	-0.026***	-0.049***
	(0.010)	(0.007)	(0.013)	(0.007)	(0.009)	(0.009)	(0.013)
Ψ[graded rtw]							-0.060
							(0.053)
stage 1: Ψ*[weeks waited]*	-3.935***	-5.345***	-3.351***	-5.079***	-4.230***	-5.450***	-3.550***
	(0.791)	(0.769)	(0.837)	(0.776)	(0.781)	(1.012)	(0.853)
b. Initial degree of grading							
Degree grading (0-100)	0.006***	0.005**	0.006***	0.004*	0.007***	0.005	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.002)
Ψ[graded rtw]							0.107**
							(0.045)
stage 1: Ψ*[degree grading]*	25.37***	24.57***	25.54***	25.07***	25.66***	26.13***	25.46***
	(0.564)	(0.576)	(0.560)	(0.596)	(0.670)	(1.479)	(0.537)

Claimants are excluded when their assigned case manager treated fewer than 25 claimants in the same year as the claimant. Panels a and b are based upon 5,913 observations of which 4,492 remain in column (5) and 2,105 in column (6). Clustered (case manager - year) standard errors between parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

group averages of the characteristics of the clients per case manager-year combination. If a group average of one of the characteristics is more than three (column 4) or two (column 5) standard deviations away from the mean of the group averages the group composition is defined as abnormal. In effect, it means that if a case managers has an extremely high or low number of clients of the same sick type or gender or extremely high or low average ages, sick duration until application at the office, or wage levels among his clients, the clients belonging to this case manager in the respective year are removed from the sample. Excluding these observations results in slightly smaller point estimates than the baseline, but not statistically significantly different.

Finally, we have conducted similar sensitivity tests that apply to the sample with individuals with graded work only – i.e., the analyses on timing and initial degree of graded work. In particular, we add the propensity used in table 4.7 panel a to the regressions in table 4.8, to check if the weeks waited and probability to recover are correlated with the overall propensity to grade.[17] With results that are virtually identical, the picture that emerges is that sample selection effects are negligible.

Case manager quality

Graded return-to-work is only one of the pieces in the case manager's toolbox. He may also make use of other interventions such as paramedical care, job training, or coaching. Or he may assert control by contacting the employee at the right moments and giving valuable advice. A case manager thus can be effective in many different ways. For the validity of our approach, we assume that the case managers' propensity to grade is not correlated with overall case manager quality – i.e., the exclusion restriction. This assumption may not hold when, for example, high quality case managers are also better at motivating the employer and employee in setting up graded return-to-work arrangements, so that there exists a

[17]The correlation between the propensity to grade and the weeks waited propensity equals 0.3433. The correlation between the propensity to grade and the degree grading propensity equals -0.240.

positive correlation between the propensity to grade and the likelihood to return to work for reasons other than graded return-to-work itself. Or, on the contrary, it could be that lower quality case managers tend to overestimate the ability of individuals to participate in graded return-to-work and enter individuals in graded return-to-work too early, leading to a negative correlation between propensities to grade and the likelihood to return to work.

A straightforward measure of case manager quality is his success: does he or she manage to get individuals back to work quickly? Similar to the propensity to grade, we therefore define a 'propensity to cure' that measures the ability of the case manager to get individuals, other that the individual concerned, back to work quickly. Column (2) of Table 4.9 shows the results of the IV regressions when we control for case manager quality using this propensity to cure. Again, we take the return-to-work within one year as the relevant outcome measure. Being appointed to the highest quality case manager rather than the lowest quality case manager increases the likelihood to return to work with 36 percentage point. At the same time, the effect of graded return-to-work itself decreases to -0.027 when all trajectories are considered and to 0.191 when only the trajectories in the first half year are considered. Following these results, one could conclude that half of the effect of the graded return-to-work trajectories started in the first half year could actually be ascribed to general case manager quality and grading itself is less effective on its own.

That being said, using case manager quality in the way described may be problematic for the same reason as not controlling for quality at all. In particular, the overall quality that we measure may partly be due to the appropriate usage of graded return-to-work, such that the propensity may absorb a too large part of the effect.[18] Under the assumption that case manager quality is relatively constant[19], but the usage of graded work may vary, we could partly resolve this problem by using a lagged quality

[18]The correlation between the propensity to grade (within the first half year) and the propensity to cure is 0.200.

[19]In principle some may have more natural ability at the job than others. At the same time, there may be room to increase job performance by learning from past cases or through training.

Table 4.9: Sensitivity tests for the importance of case manager quality

Dependent: **rtw within 1 year**	(1) Baseline	(2)	(3)	(4)
		Quality: Propensity to cure		
		Regular	Lagged	Graded at start
a. Overall effect: trajectories started in week 1-52				
Graded rtw	0.127	-0.027	0.266*	0.080
	(0.122)	(0.088)	(0.143)	(0.133)
Ψ[cure]		0.356***	0.202***	0.143***
		(0.044)	(0.071)	(0.038)
stage 1: Ψ*[graded rtw]*	0.270***	0.277***	0.267***	0.268***
	(0.027)	(0.027)	(0.032)	(0.027)
stage 1: Ψ*[cure]*		-0.059*	-0.091**	-0.008
		(0.030)	(0.037)	(0.023)
b. Overall effect: trajectories started in week 1-26				
Graded rtw	0.380***	0.191**	0.396***	0.323**
	(0.125)	(0.085)	(0.123)	(0.133)
Ψ[cure]		0.342***	0.188***	0.137***
		(0.043)	(0.060)	(0.035)
stage 1: Ψ*[graded rtw]*	0.268***	0.268***	0.284***	0.257***
	(0.027)	(0.027)	(0.030)	(0.031)
stage 1: Ψ*[cure]*		-0.0004	-0.036	0.004
		(0.028)	(0.034)	(0.021)
c. Duration until start of trajectory				
Weeks waited	-0.044***	-0.041***	-0.048***	-0.050***
	(0.010)	(0.009)	(0.010)	(0.011)
Ψ[cure]		0.153**	0.074	0.063
		(0.061)	(0.083)	(0.053)
stage 1: Ψ*[weeks waited]*	-3.935***	-3.970***	-4.125***	-3.490***
	(0.791)	(0.791)	(0.833)	(0.848)
stage 1: Ψ*[cure]*		0.398	0.165	-0.324
		(0.894)	(1.094)	(0.741)
d. Initial degree of grading				
Degree grading (0-100)	0.006***	0.005***	0.005**	0.006***
	(0.002)	(0.002)	(0.002)	(0.002)
Ψ[cure]		0.157**	0.068	0.065
		(0.066)	(0.080)	(0.046)
stage 1: Ψ*[degree grading]*	25.37***	25.43***	25.51***	25.25***
	(0.564)	(0.574)	(0.757)	(0.608)
stage 1: Ψ*[cure]*		-0.712	0.560	1.146**
		(0.709)	(0.810)	(0.505)

Claimants excluded when the assigned case manager treated fewer than 25 claimants that year. Panels a and b based upon 11,741 observations of which 8,319 remain in column (3) and 10,244 in column (4). Panels c and d based upon 5,913 observations of which 4,408 remain in column (3) and 5,591 in column (4). Baseline results for the subsample of column (3): a. 0.180 (0.131); b. 0.355*** (0.122); c. -0.0483*** (0.00973); d. 0.00510** (0.00236). Baseline results for the subsample of column (4): a. 0.107 (0.138); b. 0.360*** (0.139); c. -0.0515*** (0.0114); d. 0.00596*** (0.00209). Clustered (case manager - year) se in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

measure. Using a lagged quality measure as a control variable, we find point estimates of the effect of graded work – as shown in panels a and b – that are virtually equal to those in the baseline model. Moreover, the maximum effect of case manager quality on work resumption is about 20 percentage point.[20]

Lastly, the original data set also included individuals that already participated in graded return-to-work before entering case management. These individuals were excluded from the sample for the baseline regressions, as the case manager had no influence on their graded return-to-work status. As an auxiliary source of information, we calculated the case managers' propensity to cure on the sample of individuals graded at the start, for each case manager-year combination with at least ten observation in the sample graded at start. As such, we have a proxy of case manager quality apart apart from the ability to appropriately use graded return-to-work.[21] Using this measure of quality we find only a slight decrease in the point estimate of graded work. Based on these estimates, it thus seems that the quality of the case manager does not drive the effects we find. Being appointed to the highest quality case manager rather than the lowest quality case manager does increase the likelihood to return to work with about 14 percentage point.

We have also conducted the above-mentioned sensitivity tests on the regressions for the effects of the timing and the initial degree of grading. As panels c and d of Table 4.9 show, including proxies for case manager quality does not change these results considerably. This suggests that high quality case managers may be inclined to use graded work more often, but not at an earlier stage or at a higher starting level than low quality case managers.

[20] When we assume case manager quality is fully constant over time, we could also control for it using case manager fixed effects. In that case the effect of graded return-to-work trajectories started in the first half year on work resumption within one year is estimated at 0.268.

[21] The case manager can still exert influence on the level of grading throughout the rest of the trajectory.

Table 4.10: First stage results for detailed subcategories

subgroup	overall: started in week 1-52	week 1-26	N	duration until start trajectory	initial degree of grading	N
general medical - mild	0.174*	0.175	907	-1.040	24.17***	348
	(0.104)	(0.109)		(1.393)	(6.003)	
general medical - medium	0.343***	0.334***	1,588	-6.012**	24.65***	895
	(0.067)	(0.067)		(2.458)	(3.580)	
general medical - severe	0.198***	0.202***	1,350	-8.520***	32.40***	632
	(0.075)	(0.075)		(2.387)	(4.770)	
physical - mild	0.177*	0.120	838	-0.757	25.16***	455
	(0.093)	(0.092)		(2.647)	(7.668)	
physical - severe	0.488***	0.495***	427	1.166	29.52***	216
	(0.099)	(0.101)		(2.103)	(5.026)	
neck, shoulder, arm complaints	0.179	0.228*	810	-0.311	32.03***	463
	(0.146)	(0.127)		(4.670)	(10.29)	
hip, ankle, knee complaints	0.318***	0.391***	743	-7.416***	20.91***	447
	(0.106)	(0.105)		(2.158)	(5.683)	
back complaints	0.081	0.036	860	5.988	50.97	477
	(0.232)	(0.222)		(9.330)	(35.04)	
psychiatric	0.043	0.082	210	-10.67	11.64	86
	(0.196)	(0.195)		(8.060)	(9.802)	
psychological - mild	0.140*	0.172**	1,338	-1.664	24.19***	687
	(0.077)	(0.085)		(1.924)	(4.475)	
psychological - severe	0.042	-0.116	328	-10.58	-2.960	146
	(0.173)	(0.162)		(6.954)	(12.57)	
psychosocial - mild	0.370***	0.288***	1,254	-2.800	20.93***	706
	(0.084)	(0.098)		(2.144)	(3.369)	
psychosocial - severe	0.445***	-0.001	209	1.262	33.69***	127
	(0.156)	(0.169)		(7.123)	(8.998)	
social problems	0.423***	0.481***	244	-7.968***	29.91***	137
	(0.096)	(0.101)		(2.301)	(4.790)	
conflict	0.228	0.440**	464	-1.023	14.54*	67
	(0.197)	(0.187)		(3.003)	(8.334)	
other[a]	0.269**	0.247**	171	-0.255	38.80**	24
	(0.110)	(0.106)		(5.230)	(15.41)	
F-test on equality of coefficients:						
F(15, 181) / F(15,130)	1.98	1.94		1.70	1.53	
p-value	0.0190	0.0220		0.0578	0.1049	

[a] 'Other' contains conditions such as flue and complaints due to pregnancy.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.11: First stage results for rough subcategories

subgroup	overall: started in week 1-52	started in week 1-26	N	duration until start trajectory	initial degree of grading	N
general medical	0.261***	0.176	3,845	-5.752***	27.18***	1,875
	(0.049)	(0.108)		(1.344)	(2.166)	
musculo-skeletal	0.298***	0.333***	2,413	-5.367***	25.85***	1,387
	(0.060)	(0.067)		(1.377)	(2.621)	
mental	0.246***	0.207***	3,373	-2.527*	21.40***	1,803
	(0.048)	(0.075)		(1.430)	(2.623)	
F-test on equality of coefficients:						
F(2,181) / F(2, 130)	0.21	1.15		1.68	1.00	
p	0.8094	0.3186		0.1909	0.3701	

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Monotonicity

For the interpretation of our results as local average treatment effects, we need instrumental monotonicity to hold. That is, an individual who would not be treated when assigned to a high propensity case manager, should also not be treated when assigned to a low propensity case manager. This would be violated if certain case managers are more likely to grade individuals with psychological problems, whereas other case managers are more likely to grade individuals with musculo-skeletal complaints.

The correlation between the case managers propensity to grade and the individuals grading status should thus be roughly equal for each subgroup of individuals. Table 4.10 shows the first stage correlation coefficients for each subgroup of diagnosis. For most diagnoses these are comparable, but not for all. When clustering sick types to the subgroups that we have used earlier in section 4.5.3, however, we get first-stage results that are very similar (Table 4.11). It thus appears that the differences in first-stage estimates for the detailed subcategories largely stem from small group size.

4.6 Conclusion

In this chapter we investigate the conditions under which graded return-to-work arrangements are most effective at rehabilitating sick-listed employees. We use administrative data from a Dutch private rehabilitation provider and exploit the differences in grading practices between case managers to identify the effect of graded return-to-work. Our analysis relies on the fact that the assignment of new sick-listed clients to case managers is based on caseload. Based on this assumption, we effectively compare the full work resumption rates of case managers with a high propensity to grade to those with a low propensity to grade. We extend this method by also constructing propensities for the speed of starting graded work and the initial level of graded work.

Generally, we find positive effects of graded return-to-work on the number of weeks worked by sick-listed workers. When initiated in the first half year of sick leave, graded return-to-work increases the probability to return to work within one year by 38 percentage point. After two years of absence, however, we do not find statistically significant effects on the probability to return to work. Graded return-to-work increases the number of weeks worked in the first two years after sick-listing with 18 weeks. Overall, these results suggest that graded return-to-work speeds up the recovery process, rather than having a permanent impact on work resumption.

Our evidence suggests that the timing of graded work and the initial level of graded work are crucial determinants of the success of trajectories. Broadly speaking, graded work trajectories should start early and at an initial level that should be sufficiently substantial. Even though starting graded work one week earlier does not affect the return to work rate after two years, it does raise the number of weeks worked in the first two years after sick-listing with two weeks. In addition, starting a graded return-to-work trajectory at a work resumption rate which is 10 percentage point higher increases the probability to return to work within two years with 2.5 percentage point.

The positive effects of graded return-to-work we find are especially strong for individuals who have general medical conditions, such as chronic illnesses. For these the positive effects persist at the end of the waiting period. For individuals with problems related to mental health, however, we find no significant effects of graded return-to-work. For these individuals, speeding up the start of graded work even causes work resumption rates to decrease.

Additional data descriptives 4.A

Table 4.12: Data selection steps

Selection step	Observations
Total number of clients	35,040
Selection on contract type and insurer [a]	- 14,156
Individual died or left because of problems with insurance	-139
No case management/reported goal different than back to work	- 2,093
Intervention/graded rtw took place before application at Keerpunt	-5,030
Implausible dates	- 121
Individual could not have been observed for two years at October 7, 2016	- 5
year 2009 - 2010 deleted (only few observations)	- 221
Observations left	13,275
Individuals excluded due to missing values or being assigned to Case managers with less than 25 clients that year	- 1,534
Observations used for analysis	11,741

[a] Different contract types follow different processes leading onto application at the workplace reintegration provider. The selected contract types follow similar procedures. The main criterion for selection was that the individuals should not have been in contact with the workplace reintegration provider before the application date.

Table 4.13: Additional case manager characteristics

average working hours per year:	
- less than 800	33.8%
- 800 - 1000	20.6%
- more than 1000	26.5%
- unknown	19.1%
senior reintegration specialist	7.4%
education:	
- secondary/vocational	10.3%
- bachelor	54.4%
- master/doctorate	35.3%
workplace education	
- less than 10 courses	16.2%
- 10 - 19 courses	44.1%
- 20 - 29 courses	26.5%
- 30 - 39 courses	10.3%
- 40 or more courses	1.5%
- unknown	0.0%

Table 4.14: Reasons for sick-listing

	Freq.	Percent
general medical - mild	1,509	8.75
general medical - medium	2,423	14.05
general medical - severe	1,827	10.59
physical - mild	1,304	7.56
physical - severe	625	3.62
neck, shoulder, arm complaints	1,199	6.95
hip, ankle, knee complaints	1,092	6.33
back complaints	1,324	7.68
psychiatric	291	1.69
psychological - mild	1,895	10.99
psychological - severe	440	2.55
psychosocial - mild	1,787	10.36
psychosocial - severe	298	1.73
social problems	384	2.23
conflict	557	3.23
other	290	1.68
total	17,245	100

4.B Additional results

The propensities are calculated based on an an OLS regression using all individual-week observations up to and including the first week into graded work or up to the end of the first year of sickness, using a dummy indicating whether one entered graded work as left hand side variable and individual characteristics as right hand side variables. The results are shown in panels a of Tables 4.15, 4.16, 4.17, and 4.18. Females are more likely to be assigned to graded work. The likelihood to participate is hump shaped in age and income. Those who apply later are more likely to participate in graded work. In the year 2012 people were less likely to participate in graded work. The results do not seem to differ substantially for the different categories. The coefficients for the sick week dummies are plotted in Figure 4.7. Using the errors from these regressions we calculate the year-case manager propensities to treat. The distribution of these propensities, before rescaling, is shown in Figure 4.8.

Panels b and c of Tables 4.15, 4.16, 4.17, and 4.18 show all the coefficient estimates for the control variables of the regressions underlying the baseline IV estimates in Tables 4.4 and 4.5.

Table 4.19 show the results for different medical conditions corresponding to Table 4.6 for the other to outcome measures. Tables 4.20, 4.21, 4.22, and 4.23 show the results using different cut-offs for the minimum number of clients per case manager.

Table 4.15: Effect of graded return-to-work when started in week 1-52, including coefficients on control variables.

a. Stage 0 - dependent: participates in graded return-to-work

sex	0.000	(0.001)	condition:			contract type:		
age at application	0.001***	(0.000)	general medical - medium	0.010***	(0.001)	B	0.001	(0.002)
age at application^2	0.000***	(0.000)	general medical - severe	0.000	(0.001)	C	0.003	(0.002)
ln(gross wage)	0.005***	(0.001)	neck, shoulder, arm	0.007***	(0.001)	D	0.001	(0.002)
ln(gross wage)^2	0.000***	(0.000)	physical - mild	0.008***	(0.001)	E	0.003**	(0.002)
weeks until application	0.001***	(0.0001)	physical - severe	0.004**	(0.002)	F	0.008***	(0.002)
weeks until application^2	0.000	(0.000)	hip, ankle, knee	0.012***	(0.002)	G	0.003	(0.002)
application year:			other	-0.011***	(0.002)	H	0.004**	(0.002)
2012	-0.001	(0.002)	psychiatric	-0.002	(0.002)	I	0.004**	(0.002)
2013	0.001	(0.002)	psychological - mild	0.004***	(0.001)	firm size:		
2014	0.002	(0.003)	psychological - severe	-0.002	(0.002)	2-9 employee	0.001*	(0.001)
Constant	-0.016	(0.088)	psychosocial - mild	0.006***	(0.001)	10-49 employees	0.004***	(0.001)
			psychosocial - severe	0.005*	(0.002)	50+ employees	0.010***	(0.002)
			back complaints	0.007***	(0.001)	unknown	0.001	(0.001)
Observations	290,929		social problems	0.008***	(0.002)			
R-squared	0.011		conflict	-0.012***	(0.002)			

b. Stage 1 - dependent: participates in graded return-to-work

ψ	0.270***	(0.027)	condition:			contract type:		
sex	0.002	(0.009)	general medical - medium	0.200***	(0.024)	B	0.027	(0.027)
age at application	0.013***	(0.003)	general medical - severe	0.107***	(0.024)	C	0.049*	(0.027)
age at application^2	0.000***	(0.000)	neck, shoulder, arm	0.193***	(0.024)	D	0.035	(0.025)
ln(gross wage)	0.102***	(0.018)	physical - mild	0.166***	(0.025)	E	0.059**	(0.026)
ln(gross wage)^2	-0.006***	(0.002)	physical - severe	0.159***	(0.032)	F	0.115***	(0.035)
weeks until application	0.000	(0.002)	hip, ankle, knee	0.236***	(0.026)	G	0.039	(0.040)
weeks until application^2	0.000***	(0.000)	other	-0.254***	(0.039)	H	0.070***	(0.026)
application year:			psychiatric	0.046	(0.039)	I	0.066**	(0.031)
2012	-0.048***	(0.012)	psychological - mild	0.148***	(0.022)	firm size:		
2013	-0.063***	(0.012)	psychological - severe	0.078**	(0.031)	2-9 employee	0.021	(0.015)
2014	-0.070***	(0.013)	psychosocial - mild	0.163***	(0.024)	10-49 employees	0.060***	(0.014)
Constant	-0.311***	(0.076)	psychosocial - severe	0.211***	(0.037)	50+ employees	0.129***	(0.028)
			back complaints	0.181***	(0.023)	unknown	0.007	(0.018)
Observations	11,741		social problems	0.152***	(0.038)			
			conflict	-0.300***	(0.027)			

c. Stage 2 - dependent: returned to work within 1 year

intervention	0.127	(0.122)	condition			contract type:		
sex	-0.031***	(0.009)	general medical - medium	-0.169***	(0.030)	B	0.047*	(0.025)
age at application	0.0003	(0.003)	general medical - severe	-0.531***	(0.024)	C	0.042	(0.027)
age at application^2	0.000	(0.000)	neck, shoulder, arm	-0.271***	(0.032)	D	-0.018	(0.028)
ln(gross wage)	0.017	(0.020)	physical - mild	-0.100***	(0.028)	E	-0.002	(0.027)
ln(gross wage)^2	-0.003*	(0.002)	physical - severe	-0.446***	(0.035)	F	0.030	(0.037)
weeks until application	-0.010***	(0.002)	hip, ankle, knee	-0.197***	(0.036)	G	0.034	(0.035)
weeks until application^2	0.000	(0.000)	other	-0.439***	(0.066)	H	-0.023	(0.027)
application year:			psychiatric	-0.478***	(0.037)	I	0.027	(0.031)
2012	0.171***	(0.017)	psychological - mild	-0.319***	(0.029)	firm size:		
2013	0.181***	(0.020)	psychological - severe	-0.510***	(0.035)	2-9 employee	0.012	(0.014)
2014	0.149***	(0.021)	psychosocial - mild	-0.170***	(0.026)	10-49 employees	0.017	(0.016)
Constant	0.713***	(0.076)	psychosocial - severe	-0.416***	(0.045)	50+ employees	0.054	(0.035)
			back complaints	-0.274***	(0.030)	unknown	0.017	(0.019)
Observations	11,741		social problems	-0.073**	(0.032)			
R-squared	0.195		conflict	-0.108**	(0.047)			

[a] baseline category: general medical light (for final results make other the baseline)
[b] baseline category: 0-2 week
[c] baseline category: 2011
Cluster robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.16: Effect of graded return-to-work when started in week 1-26, including coefficients on control variables.

a. Stage 0 - dependent: participates in graded return-to-work

sex	-0.001	(0.001)	condition:			contract type:		
age at application	0.001***	(0.000)	general medical - medium	0.006***	(0.002)	B	0.003	(0.003)
age at application^2	0.000***	(0.000)	general medical - severe	-0.017***	(0.002)	C	0.005*	(0.003)
ln(gross wage)	0.008***	(0.002)	neck, shoulder, arm	0.001	(0.003)	D	0.002	(0.003)
ln(gross wage)^2	0.000***	(0.000)	physical - mild	0.008***	(0.003)	E	0.006**	(0.003)
weeks until application	0.001**	(0.000)	physical - severe	-0.009***	(0.003)	F	0.011***	(0.004)
weeks until application^2	0.000	(0.000)	hip, ankle, knee	0.007**	(0.003)	G	0.007*	(0.004)
application year:			other	-0.027***	(0.004)	H	0.006*	(0.003)
2012	-0.003	(0.004)	psychiatric	-0.020***	(0.004)	I	0.008**	(0.003)
2013	0.000	(0.005)	psychological - mild	-0.005**	(0.002)	firm size:		
2014	0.002	(0.007)	psychological - severe	-0.017***	(0.003)	2-9 employee	0.002	(0.001)
Constant	-0.014	(0.106)	psychosocial - mild	0.001	(0.002)	10-49 employees	0.006***	(0.001)
			psychosocial - severe	-0.013***	(0.004)	50+ employees	0.014***	(0.003)
			back complaints	0.000	(0.003)	unknown	0.003	(0.002)
Observations	147,713		social problems	0.009**	(0.004)			
R-squared	0.007		conflict	-0.028***	(0.003)			

b. Stage 1 - dependent: participates in graded return-to-work

ψ	0.268***	(0.027)	condition:			contract type:		
sex	-0.009	(0.009)	general medical - medium	0.088***	(0.025)	B	0.038	(0.028)
age at application	0.007**	(0.003)	general medical - severe	-0.114***	(0.024)	C	0.065**	(0.029)
age at application^2	0.000***	(0.000)	neck, shoulder, arm	0.059**	(0.024)	D	0.035	(0.028)
ln(gross wage)	0.078***	(0.017)	physical - mild	0.104***	(0.027)	E	0.065**	(0.030)
ln(gross wage)^2	-0.005***	(0.002)	physical - severe	-0.037	(0.032)	F	0.085**	(0.034)
weeks until application	-0.016***	(0.002)	hip, ankle, knee	0.096***	(0.027)	G	0.069*	(0.039)
weeks until application^2	0.000	(0.000)	other	-0.269***	(0.037)	H	0.063**	(0.028)
application year:			psychiatric	-0.163***	(0.036)	I	0.069**	(0.032)
2012	-0.038***	(0.010)	psychological - mild	0.004	(0.023)	firm size:		
2013	-0.033***	(0.011)	psychological - severe	-0.113***	(0.030)	2-9 employee	0.015	(0.014)
2014	-0.052***	(0.012)	psychosocial - mild	0.053**	(0.024)	10-49 employees	0.048***	(0.013)
Constant	-0.005	(0.072)	psychosocial - severe	-0.066	(0.041)	50+ employees	0.124***	(0.029)
			back complaints	0.048**	(0.024)	unknown	0.022	(0.018)
Observations	11,741		social problems	0.103**	(0.040)			
			conflict	-0.347***	(0.027)			

c. Stage 2 - dependent: returned to work within 1 year

intervention	0.380***	(0.125)	condition			contract type:		
sex	-0.027***	(0.009)	general medical - medium	-0.176***	(0.020)	B	0.036	(0.026)
age at application	-0.001	(0.003)	general medical - severe	-0.473***	(0.024)	C	0.023	(0.028)
age at application^2	0.000	(0.000)	neck, shoulder, arm	-0.267***	(0.022)	D	-0.028	(0.028)
ln(gross wage)	-0.001	(0.020)	physical - mild	-0.117***	(0.024)	E	-0.020	(0.028)
ln(gross wage)^2	-0.002	(0.002)	physical - severe	-0.411***	(0.026)	F	0.011	(0.035)
weeks until application	-0.004	(0.003)	hip, ankle, knee	-0.203***	(0.025)	G	0.011	(0.035)
weeks until application^2	0.000	(0.000)	other	-0.370***	(0.064)	H	-0.039	(0.027)
application year:			psychiatric	-0.408***	(0.039)	I	0.008	(0.032)
2012	0.179***	(0.017)	psychological - mild	-0.301***	(0.021)	firm size: (CHECK)		
2013	0.186***	(0.018)	psychological - severe	-0.455***	(0.034)	2-10 employee	0.008	(0.014)
2014	0.160***	(0.019)	psychosocial - mild	-0.168***	(0.020)	10-49 employees	0.006	(0.016)
Constant	0.634***	(0.073)	psychosocial - severe	-0.362***	(0.035)	50+ employees	0.022	(0.035)
			back complaints	-0.268***	(0.021)	unknown	0.010	(0.018)
Observations	11,741		social problems	-0.090***	(0.026)			
R-squared	0.230		conflict	-0.014	(0.053)			

[a] baseline category: general medical light (for final results make other the baseline)
[b] baseline category: 0-2 week
[c] baseline category: 2011
Cluster robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.17: Effect of starting moment of graded return-to-work, including coefficients on control variables

a. Stage 0 - dependent: participates in graded return-to-work

sex	-0.008***	(0.002)	condition:			contract type:		
age at application	-0.001**	(0.001)	general medical - medium	-0.095***	(0.007)	B	0.008	(0.007)
age at application^2	0.000	(0.000)	general medical - severe	-0.149***	(0.007)	C	0.005	(0.007)
ln(gross wage)	0.006	(0.005)	neck, shoulder, arm	-0.120***	(0.007)	D	-0.005	(0.007)
ln(gross wage)^2	-0.001	(0.000)	physical - mild	-0.087***	(0.008)	E	0.003	(0.007)
weeks until application	-0.002***	(0.001)	physical - severe	-0.137***	(0.008)	F	0.003	(0.009)
weeks until application^2	0.000***	(0.000)	hip, ankle, knee	-0.114***	(0.007)	G	-0.001	(0.009)
application year:			other	-0.087***	(0.020)	H	-0.001	(0.007)
2012	-0.007	(0.008)	psychiatric	-0.144***	(0.010)	I	0.004	(0.008)
2013	0.006	(0.010)	psychological - mild	-0.122***	(0.007)	firm size:		
2014	-0.008	(0.013)	psychological - severe	-0.153***	(0.008)	2-9 employee	-0.004	(0.003)
Constant	0.273	(0.275)	psychosocial - mild	-0.106***	(0.007)	10-49 employees	-0.002	(0.003)
			psychosocial - severe	-0.149***	(0.009)	50+ employees	0.000	(0.006)
			back complaints	-0.113***	(0.007)	unknown	0.007	(0.004)
Observations	71,670		social problems	-0.055***	(0.011)			
R-squared	0.027		conflict	-0.076***	(0.013)			

b. Stage 1 - dependent: participates in graded return-to-work

ψ	-3.935***	(0.791)	condition:			contract type:		
sex	0.838***	(0.268)	general medical - medium	4.687***	(0.478)	B	-0.822	(0.897)
age at application	0.123	(0.079)	general medical - severe	12.37***	(0.563)	C	-0.882	(0.998)
age at application^2	-0.001	(0.001)	neck, shoulder, arm	7.064***	(0.582)	D	0.607	(0.933)
ln(gross wage)	-0.705	(0.654)	physical - mild	3.876***	(0.463)	E	-0.234	(0.965)
ln(gross wage)^2	0.062	(0.056)	physical - severe	10.28***	(0.755)	F	0.059	(1.079)
weeks until application	0.841***	(0.058)	hip, ankle, knee	6.403***	(0.550)	G	-0.239	(1.170)
weeks until application^2	-0.003*	(0.002)	other	4.237**	(1.834)	H	0.156	(0.991)
application year:			psychiatric	10.77***	(1.368)	I	-0.337	(0.962)
2012	-0.073	(0.338)	psychological - mild	7.849***	(0.572)	firm size:		
2013	-1.238***	(0.326)	psychological - severe	13.21***	(1.166)	2-9 employee	0.290	(0.378)
2014	0.127	(0.377)	psychosocial - mild	5.751***	(0.492)	10-49 employees	0.066	(0.383)
Constant	5.933**	(2.566)	psychosocial - severe	12.50***	(1.203)	50+ employees	-0.068	(0.838)
			back complaints	6.620***	(0.610)	unknown	-0.753	(0.531)
Observations	5,906		social problems	1.747**	(0.687)			
			conflict	2.983***	(0.939)			

c. Stage 2 - dependent: returned to work within 1 year

intervention	-0.044***	(0.010)	condition			contract type:		
sex	0.005	(0.015)	general medical - medium	0.131**	(0.054)	B	-0.064	(0.049)
age at application	0.003	(0.005)	general medical - severe	0.105	(0.128)	C	-0.071	(0.056)
age at application^2	0.000	(0.000)	neck, shoulder, arm	0.134*	(0.077)	D	-0.050	(0.050)
ln(gross wage)	0.006	(0.051)	physical - mild	0.098*	(0.051)	E	-0.091*	(0.051)
ln(gross wage)^2	-0.003	(0.004)	physical - severe	0.118	(0.116)	F	-0.020	(0.059)
weeks until application	0.028***	(0.009)	hip, ankle, knee	0.143*	(0.074)	G	-0.016	(0.059)
weeks until application^2	0.000***	(0.000)	other	0.000	(0.127)	H	-0.093*	(0.052)
application year:			psychiatric	0.114	(0.121)	I	-0.084	(0.053)
2012	0.142***	(0.030)	psychological - mild	0.095	(0.083)	firm size:		
2013	0.129***	(0.034)	psychological - severe	0.137	(0.142)	2-10 employee	0.010	(0.020)
2014	0.126***	(0.031)	psychosocial - mild	0.146**	(0.066)	10-49 employees	-0.003	(0.020)
Constant	1.101***	(0.167)	psychosocial - severe	0.204	(0.139)	50+ employees	0.034	(0.043)
			back complaints	0.122	(0.075)	unknown	-0.010	(0.027)
Observations	5,906		social problems	0.059	(0.044)			
R-squared	-0.028		conflict	0.025	(0.078)			

[a] baseline category: general medical light (for final results make other the baseline)
[b] baseline category: 0-2 week
[c] baseline category: 2011
Cluster robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.18: Effect of initial degree of graded return-to-work, including coefficients on control variables.

a. Stage 0 - dependent: initial degree of graded return-to-work

sex	-1.566***	(0.566)	condition:			contract type:		
age at application	-0.016	(0.171)	general medical - medium	-6.155***	(1.290)	B	-2.310	(1.702)
age at application^2	0.000	(0.002)	general medical - severe	-11.13***	(1.410)	C	-2.014	(1.793)
ln(gross wage)	-1.440	(1.328)	neck, shoulder, arm	-6.723***	(1.456)	D	-3.437**	(1.737)
ln(gross wage)^2	0.019	(0.117)	physical - mild	-4.538***	(1.444)	E	-2.251	(1.715)
weeks until application	0.101	(0.152)	physical - severe	-11.81***	(1.777)	F	-0.986	(2.162)
weeks until application^2	0.003	(0.005)	hip, ankle, knee	-5.661***	(1.463)	G	-0.400	(2.347)
application year:			other	6.618	(4.229)	H	-2.102	(1.784)
2012	2.477	(1.891)	psychiatric	-12.34***	(2.448)	I	0.262	(1.928)
2013	3.799	(2.481)	psychological - mild	-11.64***	(1.366)	firm size:		
2014	3.855	(3.172)	psychological - severe	-10.35***	(2.029)	2-9 employee	1.162	(0.826)
Constant	85.950***	(12.63)	psychosocial - mild	-9.015***	(1.344)	10-49 employees	1.321	(0.824)
			psychosocial - severe	-10.02***	(2.133)	50+ employees	1.320	(1.636)
			back complaints	-8.308***	(1.441)	unknown	0.956	(1.111)
Observations	5,913		social problems	-4.814**	(2.024)			
R-squared	0.098		conflict	4.507*	(2.676)			

b. Stage 1 - dependent: initial degree of graded return-to-work

ψ	25.373***	(0.564)	condition:			contract type:		
sex	-1.499***	(0.512)	general medical - medium	-8.477***	(1.443)	B	-2.953*	(1.727)
age at application	-0.125	(0.167)	general medical - severe	-14.59***	(1.407)	C	-2.777	(1.727)
age at application^2	0.002	(0.001)	neck, shoulder, arm	-9.749***	(1.653)	D	-4.255**	(1.823)
ln(gross wage)	-1.943	(1.759)	physical - mild	-6.011***	(1.435)	E	-3.643**	(1.780)
ln(gross wage)^2	0.066	(0.140)	physical - severe	-14.91***	(1.629)	F	-1.534	(2.162)
weeks until application	-0.189	(0.136)	hip, ankle, knee	-8.306***	(1.515)	G	-0.950	(2.370)
weeks until application^2	0.009*	(0.005)	other	2.851	(5.991)	H	-3.603**	(1.712)
application year:			psychiatric	-16.02***	(2.167)	I	-0.883	(1.957)
2012	4.470***	(0.258)	psychological - mild	-14.56***	(1.404)	firm size:		
2013	4.994***	(0.309)	psychological - severe	-14.75***	(2.229)	2-9 employee	1.660**	(0.833)
2014	2.090***	(0.334)	psychosocial - mild	-11.20***	(1.239)	10-49 employees	2.013***	(0.750)
Constant	42.303***	(6.564)	psychosocial - severe	-13.31***	(2.395)	50+ employees	1.501	(1.688)
			back complaints	-10.44***	(1.472)	unknown	1.985*	(1.075)
Observations	5,913		social problems	-6.074***	(1.721)			
			conflict	3.661	(3.570)			

c. Stage 2 - dependent: returned to work within 1 year

intervention	0.006***	(0.002)	condition:			contract type:		
sex	-0.023**	(0.013)	general medical - medium	-0.033	(0.025)	B	-0.012	(0.032)
age at application	-0.002	(0.004)	general medical - severe	-0.367***	(0.035)	C	-0.024	(0.033)
age at application^2	0.000	(0.000)	neck, shoulder, arm	-0.133***	(0.033)	D	-0.053	(0.037)
ln(gross wage)	0.042	(0.032)	physical - mild	-0.049*	(0.025)	E	-0.061*	(0.034)
ln(gross wage)^2	-0.006**	(0.003)	physical - severe	-0.258***	(0.049)	F	-0.025	(0.044)
weeks until application	-0.007**	(0.003)	hip, ankle, knee	-0.101***	(0.033)	G	-0.002	(0.046)
weeks until application^2	0.000**	(0.000)	other	-0.225***	(0.085)	H	-0.080**	(0.035)
application year:			psychiatric	-0.282***	(0.064)	I	-0.067*	(0.039)
2012	0.121***	(0.027)	psychological - mild	-0.174***	(0.039)	firm size:		
2013	0.155***	(0.030)	psychological - severe	-0.377***	(0.052)	2-9 employee	-0.010	(0.018)
2014	0.110***	(0.031)	psychosocial - mild	-0.051*	(0.030)	10-49 employees	-0.019	(0.018)
Constant	0.641***	(0.152)	psychosocial - severe	-0.282***	(0.054)	50+ employees	0.030	(0.041)
			back complaints	-0.115***	(0.030)	unknown	0.016	(0.028)
Observations	5,913		social problems	0.007	(0.035)			
R-squared	0.142		conflict	-0.132**	(0.060)			

[a] baseline category: general medical light (for final results make other the baseline)
[b] baseline category: 0-2 week
[c] baseline category: 2011
Cluster robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Figure 4.7: Duration coefficients

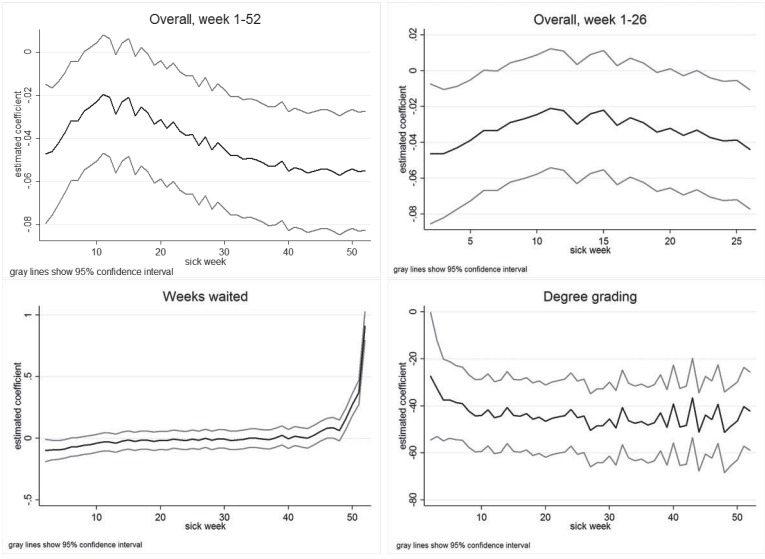


Figure 4.8: Propensities to treat before scaling.

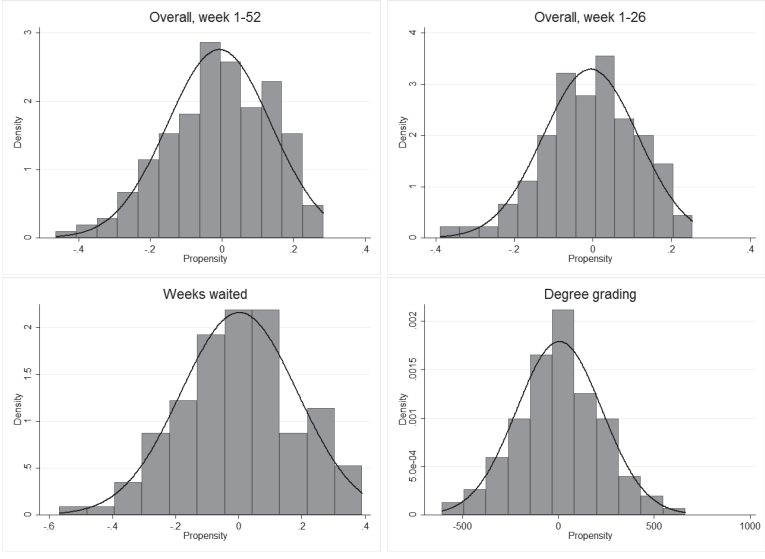


Table 4.19: IV estimation results for different medical conditions – weeks worked.

	General medical		Musculo-skeletal		Mental	
	Weeks worked in		**Weeks worked in**		**Weeks worked in**	
	week 1-52	week 1-104	week 1-52	week 1-104	week 1-52	week 1-104
a. Baseline, all trajectories started in week 1-52						
Graded rtw	9.066	30.54	8.680	6.885	7.341	6.824
	(9.855)	(20.87)	(17.21)	(29.57)	(10.85)	(24.74)
stage 1:	0.191***		0.155		0.1703**	
Ψ*[graded rtw]*	(0.072)		(0.095)		(0.074)	
b. Baseline, all trajectories started in week 1-26						
Graded rtw	20.06***	42.12***	14.31	18.34	5.354	-0.427
	(7.555)	(15.51)	(11.33)	(21.25)	(7.752)	(19.92)
stage 1:	0.281***		0.229***		0.266***	
Ψ*[graded rtw]*	(0.066)		(0.076)		(0.079)	
c. Duration until start of graded return to work trajectory						
Weeks waited	- 1.577***	-1.924**	-0.867**	-0.791	-0.532	0.490
	(0.464)	(0.767)	(0.356)	(0.684)	(0.442)	(1.116)
stage 1:	- 5.025***		- 5.179***		-3.559**	
Ψ*[weeks waited]*	(1.704)		(1.738)		(1.402)	
d. Level of work resumption at start						
Degree grading (0-100)	0.156**	0.303**	0.250***	0.546***	0.193***	0.355***
	(0.066)	(0.133)	(0.078)	(0.174)	(0.062)	(0.125)
stage 1:	30.59***		23.31***		26.71***	
Ψ*[degree grading]*	(0.859)		(1.141)		(0.841)	

The group general medical consists of individuals with the conditions general medical - mild/medium/severe. The group musculo-skeletal consists of individual with the conditions neck, shoulder, arm, hip, ankle, knee or back complaints. The group mental consists of individuals with the conditions psychiatric, psychological - mild/severe, psychosocial - mild/severe or social problems. Individuals with physical mild/severe conditions are not considered because of the small sample size. Also individuals labels as 'other' or having a conflict are excluded. Control variables include gender, age, wage, sick weeks until application, year dummies, medical conditions, contract types and firm size. Claimants are excluded when their assigned case manager treated fewer than 10 claimants of the same type in the same year as the claimant. Panels a and b are based on 3,971 observations with general medical conditions, 1,947 with musculo-skeletal conditions, and 3,380 with conditions related to mental health. Panels c and d are based on 1,667 observations with general medical conditions, 982 with musculo-skeletal conditions, and 1,807 with conditions related to mental health. Clustered (case manager - year) standard errors between parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.20: Overall results (1-52 weeks) using different cut-offs for the minimum number of clients per case manager

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
a. 15 clients or more per caseworker (N=12,534)				
Graded rtw	0.093	0.086	-0.784	4.090
	(0.117)	(0.103)	(3.489)	(8.245)
stage 1: Ψ*[graded rtw]*	0.385***			
	(0.033)			
b. 20 clients or more per caseworker (N=12,258)				
Graded rtw	0.129	0.079	0.487	5.821
	(0.115)	(0.109)	(3.375)	(8.189)
stage 1: Ψ*[graded rtw]*	0.343***			
	(0.032)			
c. 25 clients or more per caseworker (N=11,741)				
Graded rtw	0.127	0.075	1.173	6.642
	(0.122)	(0.109)	(3.581)	(8.531)
stage 1: Ψ*[graded rtw]*	0.270***			
	(0.027)			
d. 30 clients or more per caseworker (N=11,343)				
Graded rtw	0.145	0.041	1.243	5.922
	(0.121)	(0.108)	(3.626)	(8.469)
stage 1: Ψ*[graded rtw]*	0.268***			
	(0.029)			
e. 35 clients or more per caseworker (N=10,810)				
Graded rtw	0.188	0.054	2.757	7.682
	(0.124)	(0.110)	(3.683)	(8.734)
stage 1: Ψ*[graded rtw]*	0.271***			
	(0.030)			

Only graded rtw spells started in the first half year are considered
Contains results of IV regressions
Clustered (case manager - year) standard errors between parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.21: Overall results (1-26 weeks) using different cut-offs for the minimum number of clients per case manager

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
a. 15 clients or more per caseworker (N=12,534)				
Graded rtw	0.344***	0.0736	7.605**	16.02**
	(0.115)	(0.100)	(3.596)	(8.113)
stage 1: Ψ*[graded rtw]*	0.386***			
	(0.034)			
b. 20 clients or more per caseworker (N=12,258)				
Graded rtw	0.348***	0.061	7.331**	15.56*
	(0.111)	(0.103)	(3.355)	(7.945)
stage 1: Ψ*[graded rtw]*	0.386***			
	(0.035)			
c. 25 clients or more per caseworker (N=11,741)				
Graded rtw	0.380***	0.070	8.901**	18.30**
	(0.125)	(0.104)	(3.759)	(8.624)
stage 1: Ψ*[graded rtw]*	0.268***			
	(0.027)			
d. 30 clients or more per caseworker (N=11,343)				
Graded rtw	0.337***	0.031	7.803**	14.84*
	(0.119)	(0.103)	(3.699)	(8.227)
stage 1: Ψ*[graded rtw]*	0.268***			
	(0.028)			
e. 35 clients or more per caseworker (N=10,810)				
Graded rtw	0.335***	0.0272	8.125**	13.99
	(0.123)	(0.109)	(3.745)	(8.520)
stage 1: Ψ*[graded rtw]*	0.244***			
	(0.026)			

Clustered (case manager - year) standard errors between parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.22: Weeks waited results using different cut-offs for the minimum number of clients per case manager

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
a. 15 clients or more per caseworker (N=6,672)				
Weeks waited	-0.046***	-0.005	-1.397***	-2.234***
	(0.008)	(0.004)	(0.200)	(0.435)
stage 1: Ψ[weeks waited]	-6.751***			
	(0.995)			
b. 20 clients or more per caseworker (N=6,436)				
Weeks waited	-0.0428***	-0.003	-1.402***	-2.066***
	(0.008)	(0.004)	(0.218)	(0.417)
stage 1: Ψ[weeks waited]	-5.064***			
	(0.826)			
c. 25 clients or more per caseworker (N=5,906)				
Weeks waited	-0.044***	-0.001	-1.497***	-2.177***
	(0.010)	(0.005)	(0.257)	(0.489)
stage 1: Ψ[weeks waited]	-3.935***			
	(0.791)			
d. 30 clients or more per caseworker (N=5,411)				
Weeks waited	-0.042***	0.001	-1.359***	-1.891***
	(0.010)	(0.005)	(0.229)	(0.454)
stage 1: Ψ[weeks waited]	-4.230***			
	(0.825)			
e. 35 clients or more per caseworker (N=4,658)				
Weeks waited	-0.040***	0.002	-1.337***	-1.854***
	(0.011)	(0.006)	(0.245)	(0.457)
stage 1: Ψ[weeks waited]	-4.290***			
	(0.903)			

Clustered (case manager - year) standard errors between parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.23: Degree grading results using different cut-offs for the minimum number of clients per case manager

	Returned to work		**Weeks worked in**	
	1 year	2 years	week 1-52	week 1-104
a. 15 clients or more per caseworker (N=6,679)				
Degree grading	0.006***	0.002**	0.150***	0.327***
	(0.002)	(0.001)	(0.047)	(0.093)
stage 1: Ψ[degree grading]	28.650***			
	(0.481)			
b. 20 clients or more per caseworker (N=6,443)				
Degree grading	0.006***	0.003***	0.148***	0.335***
	(0.002)	(0.001)	(0.049)	(0.099)
stage 1: Ψ[degree grading]	25.178***			
	(0.459)			
c. 25 clients or more per caseworker (N=5,913)				
Degree grading	0.006***	0.003**	0.135***	0.318***
	(0.002)	(0.001)	(0.052)	(0.111)
stage 1: Ψ[degree grading]	25.373***			
	(0.564)			
d. 30 clients or more per caseworker (N=5,415)				
Degree grading	0.007***	0.003**	0.183***	0.408***
	(0.003)	(0.001)	(0.055)	(0.119)
stage 1: Ψ[degree grading]	23.244***			
	(0.638)			
e. 35 clients or more per caseworker (N=4,661)				
Degree grading	0.007***	0.002	0.203***	0.387***
	(0.003)	(0.001)	(0.065)	(0.137)
stage 1: Ψ[degree grading]	19.691***			
	(0.670)			

Clustered (case manager - year) standard errors between parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

5 One-stage versus two-stage cluster sampling, a simulation study

Abstract

Two-stage cluster sampling is a widely used method to sample households for large-scale face-to-face household surveys. Developments in survey sampling methodology, such as gridded sampling, can make it easier to define smaller Primary Sampling Units and adopt a one-stage cluster sampling approach. This approach may mitigate the risk of excluding mobile populations and reduce operational costs per cluster by combining the listing and interviewing phases. However, one-stage cluster sampling may require larger sample sizes if households of the same type tend to live close to each other. Based on a synthetic population of Oshikoto, Namibia, we analyze the potential increase in the required number of clusters under a one-stage design to achieve the statistical power of a typical two-stage cluster sample. We find that under moderate assumptions sample sizes at most double. However, in some extreme cases the required number

The chapter is co-authored by Dana Thomson. The authors thank Felicity Cutts and Dhale Rhoda for their feedback at several stages of the chapter. Furthermore, the authors thank Chris Jochem for his guidance in setting up the synthetic data set. The authors would also like to thank Jeremiah J. Nieves for assembling and sharing the WorldPop geospatial datasets used in this study. The geospatial datasets were produced by David Kerr, Heather Chamberlain, Chris T. Lloyd, Maksym Bondarenko (WorldPop, University of Southampton), Gregory Yetman, and Linda Pistolesi (Center for International Earth Science Information Network, Columbia University) in the framework of the WorldPop âĂIJGlobal High Resolution Population DenominatorsâĂİ Project funded by the Bill & Melinda Gates Foundation (OPP1134076). Lieke Kools received financial support from the Leiden University Fund/Kroese-Duijsters Fonds for conducting a research visit to Southampton University during which a large part of the work for this chapter has been conducted.

of clusters can increase by up to thirteen times. The potential increase depends on both prevalence of the characteristic and the intracluster correlation at the level of Enumeration Areas. The differences between extreme and moderate scenario's fade out when segment sizes are increased.

5.1 Introduction

Large multi-topic household surveys, such as the Living Standards Measurement Study (LSMS) and Demographic and Health Surveys (DHS) in low and middle income countries (LMICs) and EU-SILC in Europe are an important tool for monitoring socio-economic progress. When deciding how to select the respondents for such surveys, one has to ensure that the resulting sample is representative of the population and large enough to estimate key characteristics at the (sub)national level with sufficient precision. On the other hand, the survey should be affordable and the approach easy to implement in the field. A two-stage cluster sampling design was historically the only available study design and is seen as the gold standard for survey sampling in LMICs, because it offers a good balance between these requirements. However, in cases where clusters contain few household and the target population is rare, recent WHO guidelines also suggest one-stage cluster sampling as a suitable sampling method (World Health Organization 2015). Thanks to novel ways of defining clusters this alternative method now even becomes feasible for general household surveys. Compared to two-stage sampling, one-stage cluster sampling may reduce operational costs per cluster and may mitigate the risk of excluding hard-to-survey populations. However, it is more sensitive to spatial clustering of household characteristics and may therefore come at the cost of larger sample size requirements. The potential usefulness of one-stage cluster sampling in the field depends on the required increase in sample size to maintain the same statistical precision as a two-stage cluster sample. Therefore, an effort to quantify this increase is needed.

Two-stage cluster sampling is a common sampling approach for face-to-face household surveys. The first stage consists of defining primary sampling units (PSUs), i.e. mutually exclusive subsets of the population,

and then selecting a subset of these PSUs.[1] PSUs can for example be defined based on administrative boundaries, such as the Enumeration Areas (EAs) used in the most recent Census, or by overlaying a map by a raster (gridded sampling, Galway et al. (2012), Thomson et al. (2017)). In the second stage a subset of households within each PSU is selected, who together form the sample of the survey. Often this is done by visiting the selected PSU, listing all the households in that region, taking a systematic or random sample from this list, and then revisiting these households for an interview. Two-stage cluster sampling has practical benefits compared to taking a Simple Random Sample (SRS) of the population: building the sampling frame does not require complete population registries and field work can be concentrated in a few areas. However, the sampling approach may result in non-random selection of household types due to various reasons. Among these is the substantial time lag between the listing phase and interviewing phase, which has as a consequence that mobile populations, such as seasonal workers, are at risk of being excluded from the sample.

One could forgo on revisits by defining PSUs in such a way that each contains only few households and interview all the households in the selected PSU, i.e. one-stage cluster sampling. For example, a grid cell of $100m^2$ is often smaller than a EA, such that gridded sampling offers opportunities for one-stage cluster sampling. Also, when taking a gridded sampling approach, PSUs of different sizes can be established by combining neighboring cells or segmenting cells. A one-stage cluster approach could potentially lead to cost savings because the area to cover in the sampled clusters is much smaller, so that listing and interviewing can be executed on the same day. Moreover, the one-stage setup does allow to capture mobile and non-standard households, as shown by Himelein et al. (2014). However, if similar types of households tend to live close to each other, each one-stage cluster adds less new information to the sample than each two-stage cluster of the same sample size, so that one should sample

[1] There also exist two-stage methods in which the PSUs are not a mutually exclusive subset of the population. For example when points on a map are sampled and the PSU is defined as the region in a specified radius around that point. In such cases survey weights should be adapted for the possibility that two selected PSUs overlap.

more clusters for a one-stage cluster sample to achieve the same precision as a two-stage cluster sample. How many more depends on the spatial clustering of characteristics and thus on the variable of interest and the context these are measured in.

There is little evidence on the difference in precision of one- and two-stage cluster samples and how different forms of spatial clustering affect these differences. However, literature on other sampling procedures may give guidance for the direction of the results, in particular the literature on two-stage EPI sampling. For this approach, the second stage consists of selecting a random starting point in the PSU and from there taking a 'random walk' through the PSU on which households are selected until the required number of households are interviewed. The most prominent critique on this sampling method is that is does not lead to a true probability sample[2], however another critique is that households living close together are more likely to be sampled than households living further apart. Therefore, a two-stage EPI sample is affected by spatial clustering in a similar way as a one-stage cluster sample. Indeed, Milligan et al. (2004) show that implementation of two-stage EPI sampling and one-stage cluster sampling[3] lead to equivalent point estimates of vaccination coverage in the Western region of Gambia. The EPI approach has been shown to be sensitive to pocketing of vaccinated individuals (Lemeshow et al. 1985) and to perform poorly for socio-economic variables (Bennett et al. 1994). However, the approach does usually lead to estimates within

[2]When selecting households using a 'random walk', the households in the EA are not listed. As most two-stage sampling methods, the EPI approach depends on the last census for its sampling frame. These sampling frames are usually outdated, so that without listing all the households currently living in the EA one cannot establish second stage sampling probabilities. Therefore, despite the ease of implementation, the current survey guidelines of the WHO recommend against the use of this approach and in favor of systematic two-stage cluster sampling or one-stage cluster sampling (World Health Organization 2015). Contrary to the EPI approach, one-stage cluster sampling does generally lead to a true probability sample.

[3]Actually Milligan et al. (2004) compare the EPI approach to an approach they call two-stage compact segment sampling. This approach was introduced by Turner et al. (1996) and consists of first sampling PSUs, dividing these PSUs in x equal segments, and subsequently interview all households in one segment per PSU. Practically this gives a similar sample to a one-stage cluster sample, with the only exception that in a one-stage cluster sample two segments belonging to one 2-stage PSU could in principle both be sampled.

10% of the population mean (the EPI criterion for a good sample) and has shown to provide similar estimates for mortality and vaccination status as more systematic sampling procedures (Luman et al. 2007, Rose et al. 2006). To our knowledge, specific guidelines with respect to the difference in the number of clusters to sample are not given in the literature.

In this chapter we aim to find out how many additional clusters need to be sampled when using a one-stage cluster design to achieve the statistical power of a typical two-stage cluster sample in LMIC household surveys. In order to answer this question we create a synthetic population of households in Oshikoto, Namibia. We argue that this population has the same properties as the true population in the sense that both population averages and the distribution of EA-level prevalences are equivalent for key characteristics.[4] Next we adopt several scenarios for the spatial distribution of individuals within each EA, while keeping EA-level prevalences fixed. For each of these different scenarios we calculate the minimal number of clusters to be sampled to achieve a given statistical precision based on bootstrapped measures of performance. We focus on three measures (1) household wealth index, (2) womens use of modern contraception, and (3) 0-5 years old children's DPT3 vaccination coverage. These measures show different distributions of EA-level prevalences and each cover a different subsample of the population.

The results show that under moderate assumptions sample size requirements at most double. However, under extreme assumption of within EA clustering sample size requirements can increase dramatically, especially for variables with low EA-level ICCs and prevalence levels near 50%. The most extreme case showed an increase in the minimal number of clusters to sample of almost thirteen times. The differences between extreme and moderate scenario's fade out when segment sizes are increased. Before implementing one-stage cluster sampling one should carefully examine the likely scenarios of within EA clustering, to assess feasibility of the approach.

[4]That is, we expect the intra cluster correlations found in the synthetic population is close to those in the true population.

5.2 Method

We evaluate the sampling procedure using a synthetic population of the region Oshikoto in Namibia. This region was selected because of both the availability of high quality data and the diversity of the region. The region covers $38,653 km^2$ including planned and unplanned city neighborhoods, rural settled agriculture, rural nomadic populations and large unpopulated areas. In this section we briefly explain how the synthetic population is generated and argue why this provides a valid testing ground for the question at hand. Next, we explain how we constructed the different scenarios of within EA clustering and how we calculate the minimal number of clusters to be sampled when using either a one- or two-stage cluster sampling approach. For a more detailed description of the data and methods used to construct the synthetic population, we refer the reader to Thomson et al. (2018).

5.2.1 Generating a realistic synthetic population

The synthetic population is constructed using the 2013 Namibian Demographic Household Survey (DHS), the 2011 Namibian census Public Use Microdata Sample (PUMS), a set of publicly available spatial covariates, and a household point location file constructed by visual inspection of satellite imagery of the region. Table 5.1 provides an overview of the datasets and variables used to generate the synthetic population. The population is created in three steps: first we predict the spatial distribution of household types, which we then use to assign a realistic set of synthetic households to realistic household locations, and finally we predict some extra characteristics of the individuals. The steps constitute of a series of random processes, which will be further explained below, so that executing them once results in one of many possible realizations of the synthetic population. For the analysis we generate five realizations of the synthetic population, run the analysis on each of these realizations, and base our conclusions on the combined results of the different analyses.

Table 5.1: Overview of data sources used for simulations

Dataset	Information retrieved	Original source (unit)
Demographic and Health Survey 2013[a](MoHSS and ICF 2014)	geo-displaced cluster coordinates, urban/rural (hv025), cluster (v001), hhsize (derived), water source (hv201), toilet facility (hv205, hv225), space (hv216), structure (hv213), cooking fuel (hv226), relationship (hv101), age (hv105), sex (hv104), education (hv109), wealth index (hv270) , contraception (v313), DPT3 vaccination[b](h7)	
Census 2011 PUMS (NSA 2013)	admin-3 level indicator (constituency), urban/rural (urban_rural), hhsize (derived), water source (H9), toilet facility (H10), space (H4), structure (H7), cooking fuel (H8a), relationship (B3), age (B5), sex (B4), education (D3)	
2011 Census EA boundaries (NSA 2011a)	EA boundaries	
2011 Census main report (NSA 2011b)	Constituency populations totals	
2014-2016 DigitalGlobe Quickbird imagery, 50cm (DigitalGlobe 2014)	(estimated) household point locations	
ccilc_dst011_2012	Distance to cultivated terrestrial lands[c]	2012 ESA CCI annual LC maps v2.0.7 (≈300m)[d]
ccilc_dst040_2012	Distance to woody areas[c]	""
ccilc_dst130_2012	Distance to shrub areas[c]	""
ccilc_dst140_2012	Distance to herbaceous areas[c]	""
ccilc_dst150_2012	Distance to terrestrial vegetation areas[c]	""
ccilc_dst190_2012	Distance to urban area[c]	""
ccilc_dst200_2012	Distance to bare areas[c]	""
cciwat_dst	Distance to water bodies[c]	ESA CCI, Water bodies v4.0 (≈150m)[d]
dmsp_2011	Nighttime lights intensity[c]	2011 inter-calibrated version of the v4 DMSP-OLS Nighttime Lights Time Series (≈1km)[d]
gpw4coast_dst	Distance to coastline[c]	GPWv4 input administrati units (≈100m)[d]
osmint_dst	Distance to road intersections[c]	2016 OSM highways[d]
osmriv_dst	Distance to major waterways[c]	2016 OSM waterways[d]
osmroa_dst	Distance to major roads[c]	2016 OSM highways[d]
slope	Slope[c]	2000 Viewfinder Panoramas (≈100m)[d]
topo	Elevation[c]	2000 Viewfinder Panoramas (≈100m)[d]
tt50k2000	Travel time to populated places (pop more 50k)	2000 EC-JRC Travel time to major cities (≈1km)[d]
urbpx_prp_1_2012	Proportion of settlement pixels within 1 cell radius[c]	2012 DLR Global Urban Footpring (≈12.5m) & 2000 EC-JRC Global Human Settlement layer; 38m[d]
2010 MODIS (≈1km) (Running et al. 2014)	Annual net primary productivity[c]	
2001 education facilities (UN-OCHA ROSA 2001b)	Distance to schools[c]	
2001 health facilities (UN-OCHA ROSA 2001a)	Distance to health facilities[c]	

[a] The household variables are taken from the household recode file, the women and child variables come from the individual recode file.

[b] We only measure DPT3 vaccination coverage for children living with their mother. In Oshikoto 29 children under 5 (9.9%) are reported to live away from their mother in the DHS. The vaccination coverage of children living away from their mother is slightly lower than that of children living with there parents, though not statistically significantly lower (72.41% vs 78.41%, p-value of one-side t-test: 0.2312).

[c] Unit of measurement: 3 arc seconds (≈ 100m).

[d] Spatial covariate was processed by the "Global High Resolution Population Denominators" Project.

Step 1: Predict the spatial distribution of household types.

In order to generate a synthetic population which is realistically distributed over space, we need to understand which types of households are likely to live in which areas of our region. We can do this by estimating relationships between spatial covariates and household types. For this we rely on the 2013 DHS survey, which not only provides information on households but also GPS coordinates of the surveyed clusters. To establish this relationship we need to find out what the typical household looks like in each surveyed cluster. We start by summarizing the individual characteristics to the household-level, so that we have a household file with the following dummy variables: 1[is rural], 1[head has any formal education], 1[has any children under 5 years old], 1[does not have access to an improved water source][5], 1[does not have access to improved toilet facility][6], 1[lives in a non-durable structure][7], 1[lives in house with inadequate space][8], 1[cooks on solid fuel][9]. The choice for these characteristics is based on the availability of information in both the DHS and the census PUMS, which we use in step 2 to generate our synthetic population. Next, we take the cluster average of these household characteristics, resulting in one typical household per cluster. Finally, we construct a single variable

[5]i.e. the water source is labeled as well unprotected, river/dam/stream in cases of census data and the water source is labeled as unprotected well, river/dam/lake/ponds/stream/canal/irrigation, unprotected spring, tanker truck, cart with small tank (hv205), or shared (hv225) in case of DHS data (UN-HSP 2003). For both the DHS as the census the category other is set to missing.

[6]i.e. the toilet facility is labeled as uncovered pit latrine without ventilation, bucket toilet, or no facility in case of census data and the toilet facility is labeled as pit latrine without slab/open pit, flush to somewhere else, bucket toilet, hanging toilet/latrine, or no facility/bush/field in case of DHS data (UN-HSP 2003). For both the DHS as the census the category other is set to missing.

[7]i.e. the floor material is labeled as sand/earth, cement, mud/clay or wood in case of census data and the floor material is labeled as earth/sand, dung, mud/clay, wood planks, palm/bamboo in case of DHS data Fink et al. (2014). For both the DHS as the census the category other is set to missing.

[8]i.e. on average more than 3 individuals share one sleeping room. For both the census as the DHS data this was derived from household size and a variable measuring the number of sleeping rooms in the house (UN-HSP 2003).

[9]i.e. cooking fuel is labeled as wood/charcoal from wood, charcoal-coal, or animal dung in case of census data and cooking fuel is labeled as charcoal, wood, agricultural crop or animal dung in case of DHS data. For both the DHS as the census the category other is set to missing.

Table 5.2: Household types

Type	Name	Description
1	rural rich	educated and high access to facilities
2	rural poor 1	No formal education and low access to facilities, except water
3	urban rich	educated, few under fives, and high access to facilities
4	urban average	No formal education and average access to facilities
5	rural average 1	few under fives, average to high access to facilities
6	rural poor 2	many under fives and low access to facilities
7	rural average 2	average access to facilities, low access to fuel.

High:= above regional average, Low:= below regional average, Regular:= close to regional average. Summary statistics of each type are given in Thomson et al. (2018).

that summarizes the information of the different characteristics by means of k-means clustering. K-means clustering is a form of unsupervised clustering aiming to partition observations into a pre-defined number (k) of groups or clusters. The clusters are formed such that the sum of squares between the points and the cluster centroids (middle points) is minimized (Hartigan and Wong 1979). We denote the resulting variable as the 'household type'. The algorithm results in the 7 types depicted in table 5.2.

In order to describe the relationship between these household types and the spatial covariates we fit a Random Forest model predicting the cluster household type using information related to the location of the clusters, such as elevation and distance to roads. A Random Forest is a supervised learning technique that can be used for both classification as regression (Breiman 2001). It is an ensemble method meaning that it combines information from multiple fitted models so to obtain better predictive performance. In the case of a Random Forest these building blocks are called Decision Trees. A Decision Tree is a classification or regression model aiming to predict the class or value of a certain outcome measure by generating splits on the input variables.[10] Splits are chosen so to minimize a cost function, e.g. sum of squares in case of regression.

For sake of anonymity the DHS provides GPS coordinates that are displaced by up to two kilometer in urban areas, up to five kilometer

[10]Would we for example have one input variable x and output variable y a decision tree could hold the following information: if $x < 3$ then $\bar{y} = 2$, if $x \geq 3$ then $\bar{y} = 7$.

in rural areas, and up to ten kilometer in a random one percent of the rural cases (Perez-Heydrich et al. 2013). Therefore, rather than using the actual value of the spatial covariates at the given GPS locations, we extract for each cluster the average, minimum, and maximum value of the spatial covariates within a radius of five kilometers around the cluster coordinates and use those generated covariates in the Random Forest model. When applying the k-means method for clustering we selected only the clusters in Oshikoto to ensure that we only define household types which are meaningful to our region. When fitting the Random Forest model however, we use all the information available for Namibia, to avoid overfitting due to the small sample of clusters in Oshikoto (N=38). We thus apply the household type definition constructed using Oshikoto data to all the clusters in the DHS survey before running the model.

Finally, using the estimated Random Forest model, we predict for each 100m grid cell h in Oshikoto the probability p_{hk} that the average household is of household type $k, k = 1, \ldots, 7$. We combine these probabilities in seven grids, one for each household type, which can be seen as probability surfaces. Because the spatial covariates give little extra information about the possible variation within very dense areas (in our case the city Tsumeb), we inspected satellite imagery of those areas to create an extra probability layer with subjective probabilities of the presence of rich households in those areas. This layer is multiplied with the probability surfaces for urban household types, to force a more realistic assignment of household types within cities.

Step 2: Generate a realistic synthetic population.

We build a realistic synthetic base population from the Public Use Microdata Sample (PUMS) of the 2011 Namibian Census, using the R-package simPop. We first generate a set of households by sampling household ids from the Census PUMS file using the provided household weights recalibrated to the total number of observations per constituency in the household point location file. The variables age, gender, and relationship (i.e. head, child etc.) of the household members in the households corre-

sponding with these household ids are replicated from the Census PUMS. Next, the household attributes water, toilet, structure, space, and fuel and the individual attribute education are predicted using multinomial models. We thus create synthetic households with combinations of characteristics that are similar to those in the census PUMS, while allowing for combinations of characteristics not present in this sample. In this way anonymity of the real households is preserved (Templ et al. 2017).

Using the characteristics of our synthetic population and the probability surfaces we can now assign the synthetic households to a household location. We start by assigning a household type, as defined in step 1, to each household in our synthetic population. Next, we want to assign to each household location j the probability q_{jk} that it holds a household of household type k, $k = 1, \ldots, 7$. Assuming that within a single grid of 100 m^2 there is no clustering of household types[11], we can set the probabilities for the household locations (j) equal to the probabilities of the grid cell (h) that they fall in. That is,

$$q_{jk} := p_{hk} \quad \text{if } j \in h, \quad \text{for } j = 1, \ldots, N_j, h = 1, \ldots, N_h, k = 1, \ldots, 7.$$

Now, if household i is of household type k, the probability that it is located at household location point j is given by

$$r_{ij} = \sum_{k=1}^{7} 1_{[\text{hhtype}=k]} \frac{q_{jk}}{\sum_{j=1}^{N} q_{jk}} \quad \text{for } i, j = 1, \ldots, N_j.$$

Based on this information we iteratively assign individuals to locations by applying the following steps for each constituency x urban-rural group of household points/households separately:

0. Let H be a list of household ids and corresponding household types, and let L be a list of locations and corresponding probabilities q_{hk}.
1. Randomly select a household i from H.

[11] We should note here that this does not mean that we do not assume any spatial clustering, that is, each different grid cell does have a different prevalence of household types.

2. Select a location j from L by sampling with probability weights r_{ij}.
3. Remove household i from H and remove location j from L.
4. Repeat step 1-3 until all households are allocated.

Step 3: Clustered household characteristics

We now have a base synthetic household population representative of space, but it does not yet contain the characteristics of interest to us, that is household wealth index, womens use of contraception, DPT3 vaccination coverage. The PUMS does not contain any information on these variables, so that they could not be added in step 2. However, the DHS does contain information on these characteristics as well as the base characteristics of our synthetic population. Therefore, for each variable we thus fit a multinomial model[12] on the DHS using the base characteristics as dependent variables, and subsequently predict these variables for our synthetic population.

Degree of realism of the population

For the generated synthetic household populations to be useful for our research it should have realistic population and EA-level properties. We are confident that this is the case because (1) population means of characteristics in the synthetic populations are equivalent to the population means of the 20% census PUMS; (2) Constituency-level means of characteristics in the synthetic population are equivalent to constituency-level means of the 20% census PUMS; (3) EA-level maps of prevalences show realistic spatial distributions of characteristics; (4) density plots of EA prevalences based on the synthetic population look sufficiently similar to plots based on DHS data; (5) the DHS sample is a potential sample from our synthetic populations. For the tables and figures supporting these claims we refer the reader to Thomson et al. (2018).

[12]The SimPop package used in step 1 employs the same model to build up the synthetic population.

5.2.2 Calculating the minimal number of clusters

We will calculate the minimal number of clusters required for one- and two-stage cluster sampling by means of bootstrapped samples from our synthetic population. In this section we explain (1) the set-up of our one- and two-stage sampling procedures, (2) the different scenarios of within EA clustering, and (3) the algorithm used to search for the minimal number of clusters for each combination of scenario and sampling method.

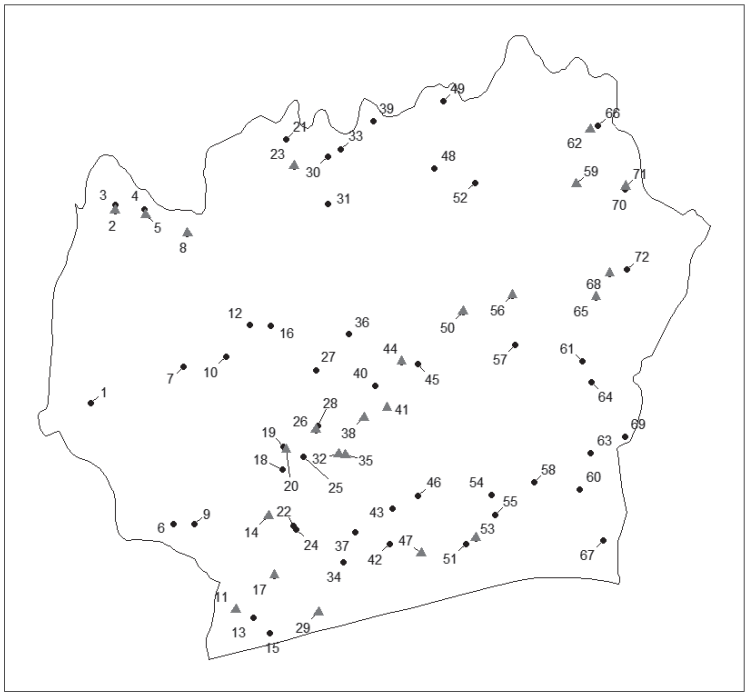
Cluster sampling set-ups

We design the two-stage cluster sampling approach such that it corresponds to the approach adopted for the 2013 DHS in Namibia. The Enumeration Areas from the 2011 Census are used as PSUs, holding on average 86 households. Since we know the coordinates of the household locations, we can easily retrieve the accompanying EA using a shapefile containing the boundaries of the EAs used in the 2011 Census. First a given number of EAs is randomly selected, after which from each of the selected EAs 25 households are systematically selected for interviews. That is, we order households within a given EA first by longitude and then by latitude. Then we randomly select one of the first $n := floor(EAsize/25)$ households on the list and from there on select every nth next household on the list.[13] An example of a two-stage sample is given in Figure 5.1

When we design our one-stage sampling set-up we aim to design PSUs of approximately 25 households, so that the size of the sample taken from one cluster is the same in the one- and two-stage cluster sampling approaches. We will call these groups of 25 households *segments*, since we define them by 'segmenting' the EAs in blocks of 25 households. That is, we order the households within a given EA first by longitude and then by latitude. Then we assign the first $m := floor(EAsize/n)$ households to one segment, the next m households to a second segment and so on. As a result the whole region will be divided into small segments of approximately 25 households. The one-stage cluster sample results from randomly selecting

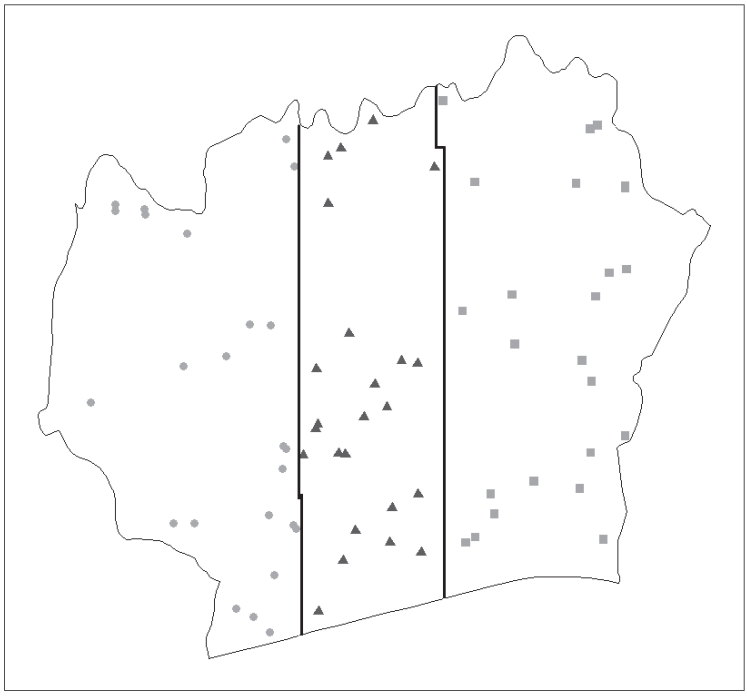
[13] We take every nth household on the list rather than a completely random set of households, with the aim to mimic as close as possible the sampling process in the field.

Figure 5.1: An example of 2-stage sampling of ordered household points and selection.



The numbers in the figure represent the ordering of households first by longitude and then by latitude. The selected households in the two-stage cluster sample are indicated by a triangle.

Figure 5.2: An example of 1-stage sampling segments.



For a one-stage cluster sample the given EA would be divided into three segments, the households indicated with dots would fall in segment 1, those indicated with triangles in segment 2, and those indicated with squares in segment 3. This is stressed by the two lines dividing the EA in three parts.

a given number of these segments. An example of segments for one-stage sampling are given in Figure 5.2.

Scenarios of spatial clustering

In order to analyze the performance of one- and two-stage cluster samples under several (extreme) cases of spatial clustering within EAs, we relocate households within EAs according to several scenarios. We start by ordering all household locations within one EA first by longitude and then by latitude. Next, we order the list of households in the same EA according to one of the scenarios given in table 5.3. For example, for scenario 3 we order the households from low wealth to high wealth. Then we attach the ordered list of households to the ordered list of coordinates and define the one-stage cluster segments as explained above. Figure 5.3 shows an example of the spatial distribution of wealth within an EA for the baseline case (scenario 1) and the case where wealth index is perfectly clustered (scenario 3).

The assumption of perfect clustering by our key outcome variables is rather extreme. A more moderate assumption would be that households are clustered by other household variables, such as access to improved toilet or water facilities, which are related to our outcome variables but do not necessarily lead to perfect clustering in these variables (scenarios 6-10). Scenarios 11-13 also represent more moderate configurations of clustering, by randomly relocating 50% of the households from the extreme scenarios 3-5 within each EA.

Calculating minimal number of clusters

A common way to express the requirements for sample estimates is by defining a bandwidth of B percentage points from the population mean in which $(100 - \alpha)\%$ of sample means should fall. For example, the EPI framework was once developed with the aim in mind that it should provide estimates of immunization coverage which are with 95% certainty within 10 percentage points from the true value (Bennett et al. 1991). In

Figure 5.3: An example of the spatial clustering of wealth within one EA for scenario 1 (top) and scenario 3 (bottom).

Table 5.3: Within EA clustering scenarios

Scenario		Description
1.	Baseline	Original ordering of households (HHs)
2.	Perfectly homogeneous	HHs randomly reordered
3.	Wealth index perfectly clustered	HHs ordered from low to high wealth
4.	Contraception perfectly clustered	HHs ordered from low to high prevalence of contraception, where HHs without women aged 15-64 positioned at random
5.	DPT3 perfectly clustered	HHs ordered from low to high prevalence of DPT3 vaccination, where HHs without children under 5 positioned at random
6-10.	Perfect clustering by underlying variables	HHs ordered by access to improved toilet facility (6), improved water (7), adequate structure (8), adequate space (9), or non-solid fuel (10).
11-13.	Moderate clustering	From scenarios 2-4 randomly replace 50% of HHs

that case $B = 10$ and $\alpha = 5$. We find the minimal number of clusters for which this condition holds in an iterative way, by partitioning the search space. First, we choose the maximum number of clusters to be included in the sample (n_{max}), for example n_{max} could be set to 15% of all clusters. Then, we calculate a bootstrapped $(100 - \alpha)\%$ confidence interval based on 10.000 samples using the maximum number of clusters and the relevant cluster sampling approach. If this confidence interval is wider than the benchmark, we set the sample size requirements equal to n_{max}. If it is smaller however, we recalculate the bootstrapped $(100 - \alpha)\%$ confidence interval using $0.5 * n_{max}$ clusters. If this confidence interval is wider than the benchmark, we next evaluate the bootstrapped confidence interval when using $0.75 * n_{max}$ clusters, if it is smaller, we next evaluate the bootstrapped confidence interval at $0.25 * n_{max}$ clusters and so on until we find a bootstrapped confidence interval that is (almost) equal to the desired size.

We repeat this process using both cluster sampling approaches and for each scenario on five realizations of our synthetic population.

5.3 Data description

We perform the analysis on a set of five synthetic populations. In this section we show the data description of one of the five synthetic populations. The tables and figures for the other populations are provided in Appendix 5.B.

Table 5.4 provides summary statistics for the whole region. Synthetic population 1 consists of 179,931 individuals living in 37,298 households. The households are mostly located in rural area's. A quarter of the household heads has had no formal education, almost a third has an incomplete primary degree and only 14.2% has completed secondary or tertiary schooling. Many are lacking access to improved toilet facilities (79.9%), adequate structure (61.4%), or non-solid fuel (83.8%). However, only few lack access to improved water facilities (26.8%) or have a house with inadequate space (7.5%). Compared to the rest of Namibia, the area is rather poor, with 62.1% of households falling in the lowest two wealth quintiles, and only 17.9% in the highest two wealth quintiles. The individuals are rather young, although compared to the average of Sub-Saharan African countries there are relatively few children under five (14.0% compared to 16.4%, UN (2017)) and relatively many individuals older than fifty (13.6% compared to 9.8%, UN (2017)). The use of modern contraception under 15-49 year old women is at 43.1% and 80.3% of children under five have received all three DPT vaccinations.

Figure 5.4 shows a map of the EA-level prevalences as measured in synthetic population 1 for each of the three key characteristics: individuals in the poorest wealth quintile, women 15-49 using modern contraception, and children under 5 who have received three DPT vaccinations.[14] All three maps show substantial spatial variation. The top graph shows the prevalence of individuals in the poorest wealth category. There is a clear relationship between the prevalence of poverty measured by asset indexes and accessibility of areas. The poorest never live in the urban areas and rarely live in the regions near a large road or in more densely populated

[14]Note, these maps do not depend on the chosen scenarios, as these scenarios only result in within EA relocations of households. The EA-level prevalences are thus not affected by the choice of scenario.

Figure 5.4: Prevalence of characteristics per EA - Synthetic population 1

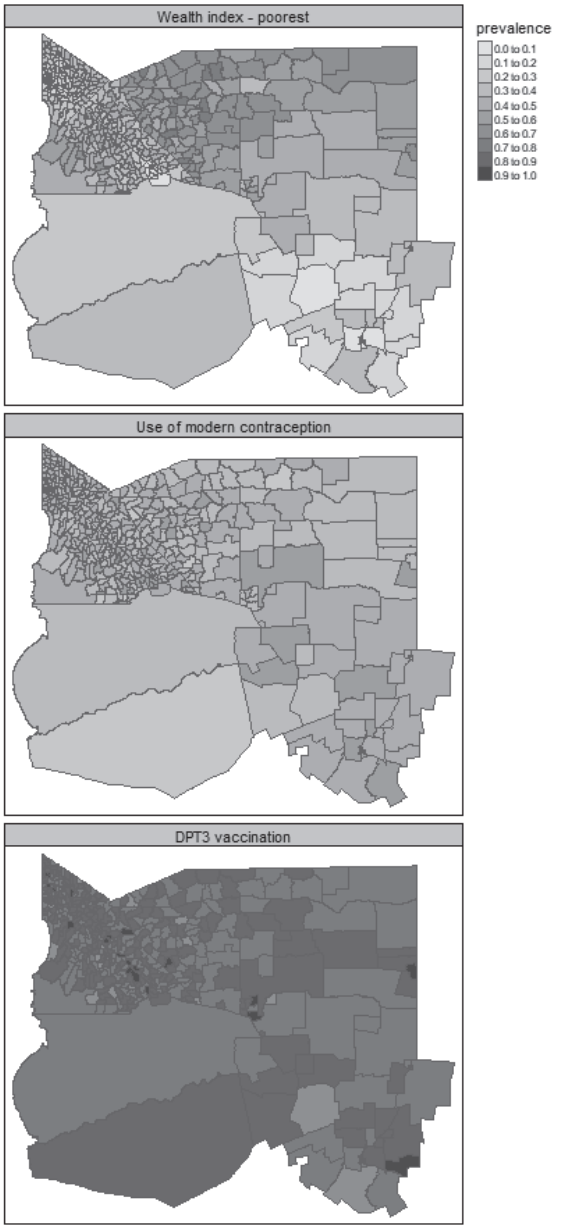


Table 5.4: Summary statistics - Synthetic population 1

household-level variables		individual-level variables	
nr households	37298	nr individuals	179931
average household size	4.82	male	47.8%
urban	15.7%	age:	
education head		- 0 - 4	14.0%
- no formal	25.1%	- 5 - 14	26.3%
- incomplete primary	30.1%	- 15 - 49	46.1%
- complete primary	30.6%	- 50 plus	13.6%
- complete secondary	10.7%		
- complete tertiary	3.5%	nr women 15-49	42785
unimproved water	26.8%	modern contraception	43.1%
unimproved toilet	79.9%		
inadequate space	7.5%	nr children under 5	25249
inadequate structure	61.4%	DPT3 vaccination	80.3%
solid fuel	83.8%		
wealth index			
- poorest	33.9%		
- poorer	28.2%		
- middle	19.7%		
- richer	13.5%		
- richest	4.6%		

areas.[15] The prevalence of the use of modern contraception shows a less clear spatial pattern. Prediction models also show that the use of modern contraception is only weakly related to indicators like education and access to improved water facilities (which do show a distinct spatial pattern) and more so to variables such as age (which is more uniformly distributed over space). The prevalence of DPT3 vaccination seems to be somewhat higher in more densely populated areas. However, also in the case of DPT3 vaccination there is substantial variation unrelated to the observed factors that show distinct spatial patterns.

The different scenarios defined in table 5.3 are likely to lead to fairly different segment-level ICCs. Figure 5.5 shows the segment-level ICCs together with boxplots of the segment-level prevalences. We can compare these to the EA-level ICC and boxplots reported in the same figure. The most left boxplot shows the spread of EA prevalence and confirms the image painted in Figure 5.4. There is substantial spread in the prevalence

[15] Figures of population density and distance to roads can be found in Appendix 5.A.

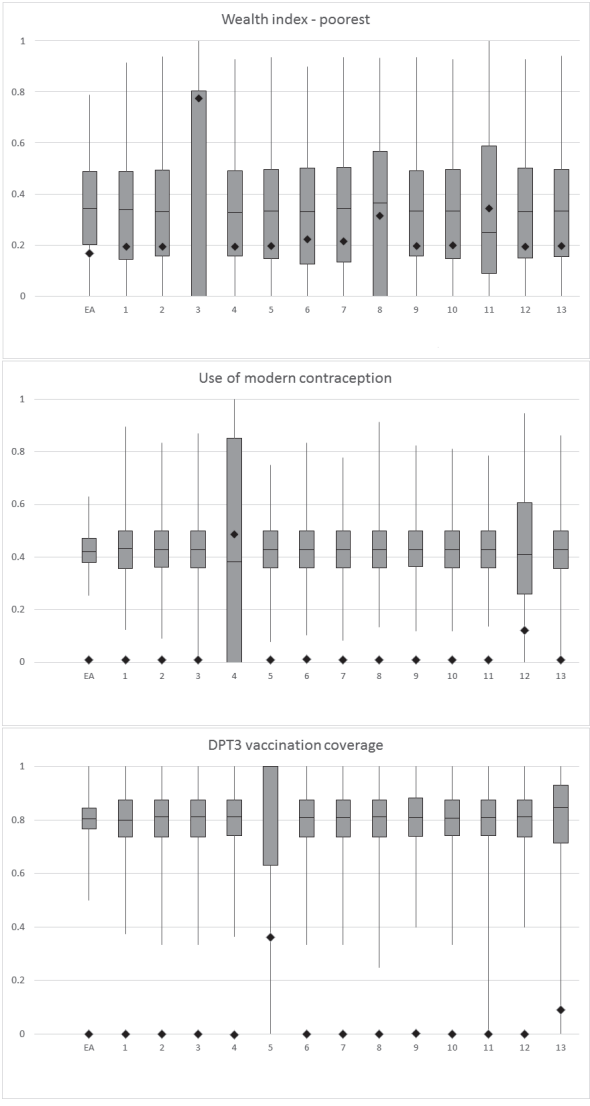
of being in the poorest wealth category, with half of EA prevalence levels falling within about 14 percentage point of the median EA prevalence. The spread of prevalence of modern contraception and DPT3 vaccination is more centered, with half of EAs having a prevalence within about 4 percentage point of the median EA prevalence. This is also reflected in the ICC, which is 0.17 for the poorest wealth category, but respectively 0.01 and 0.001 for use of modern contraception and DPT3 vaccination. The EA-level will be the basis of our two-stage cluster sampling procedure. The ICCs found in our synthetic population are close to the observed ICCs in the DHS 2013 sample, which are 0.23 for the poorest wealth category, -0.01 for contraception, and 0.02 for DPT3.[16]

For the one-stage cluster sampling procedure we cut the EAs into segments holding about 25 households (based on a list ordered by the (x,y) coordinates of households), relocated within their EA according to different scenarios. In the baseline case households are kept at their original location (scenario 1, second boxplot from the left). The spread of segment prevalences is somewhat larger than the spread of EA prevalences. ICC levels are similar in case of contraception and DPT3 and only slightly higher for the wealth indicator, such that we do not expect large increases in sample size requirements when moving from a two-stage cluster sampling design to a one-stage cluster sampling design.

The next four scenarios depict four possible extreme cases of spatial clustering within EAs: (2) complete homogeneity, (3) perfect clustering by wealth, (4) perfect clustering by contraception, (5) perfect clustering by DPT3 vaccination. The spread and ICC in scenario 2 are almost equivalent to the baseline scenario. When we assume perfect (within EA) clustering by wealth index (scenario 3) or contraception (scenario 4) the spread on the respective variable increases substantially. This is what one would expect, since within every EA the poorest households, those with high fractions of women using modern contraception, or those with high fractions of

[16]These figures are slightly lower than the national-level ICCs for Namibia, which are respectively 0.41, 0.04, and, 0.09. There is quite some variation in regional-level ICCs for the different regions in Namibia and Oshikoto represents a rather average regional case for Namibia. The national-level ICCs and regional-level ICCs for all regions in Namibia are reported in Appendix 5.D.

Figure 5.5: ICC (♦) and boxplots of prevalences per EA/segment - Synthetic population 1



children with a DPT3 vaccination were placed close together, so that they end up in the same segment. Segments are thus likely to have either very high prevalence rates on the respective variables or very low prevalence rates. In scenario 4, the prevalence of use of modern contraception falls within 40 percentage point of the median prevalence level for 50% of the segments. The difference between the minimum and maximum prevalence of DPT3 also increases under the assumption of perfect clustering (scenario 5). In this case however the bulk of prevalences is centered at the top, as the median prevalence is equal to 100%. The average prevalence of DPT3 vaccination is quite high, such that clustering and segmenting will likely lead to many segments with 100% prevalence and only few with slightly lower prevalences.[17] ICC levels increase substantially after clustering, most extremely for wealth (from 0.19 under the baseline scenario to 0.77 under scenario 3) and contraception (from 0.01 to 0.49 under scenario 4), and somewhat more moderately for DPT3 (from 0.00 to 0.36 under scenario 5). Clustering by wealth also slightly affects the spread of prevalences of contraception and DPT3, though the ICC levels are not affected. Based on the reported increases in ICC we expect that there may be large differences in the sample size requirements of one- and two-stage cluster sampling, under the assumption of perfect clustering.

Clustering by underlying factors (scenarios 6-10) leads to a slight increase in the spread of the prevalences. The increase is substantially larger for the wealth indicator when we assume households are perfectly clustered by structure (scenario 8). This may be because the percentage of households with an inadequate structure is closer to 50%, so that there is more potential for clustering or because the wealth and structure are more closely related. Also when replacing 50% of households from scenarios 3-5 (scenarios 11-13) ICC levels only increase slightly.

[17] For example suppose we have an EA with 75 households and a household characteristic with a prevalence rate of 50%. If we would order the households by prevalence and then segment into three groups, the segments would have prevalence rates of 0%, 50%, and 100%. However, would the EA prevalence have been equal to 90%, the resulting segment-level prevalences would equal 70%, 100%, and 100%.

Results 5.4

Baseline analysis 5.4.1

For each of the five synthetic populations we calculated the minimum number of clusters necessary to obtain a sample estimate which is with 95% certainty within 5 percentage points from the population mean. Table 5.5 reports the averages of the minimum number of clusters found in the five synthetic populations. The results for the five synthetic populations separately can be found in Appendix 5.C. There is only little variation between the results for the different synthetic populations.

The minimum number of clusters under a two-stage cluster sample are given in row 1. In case of a two-stage cluster sample we need to sample sixty-five clusters to obtain a sufficiently precise estimate of the proportion of individuals in the poorest wealth quintile (column 2), fifteen clusters for the proportion of women using modern contraception (column 4), and also fifteen clusters for the proportion of under five year old children with a DPT3 vaccination (column 6). The relatively large requirements for the wealth indicator stem from the fact that this measure has a higher ICC than the other variables. Moreover, the variable is perfectly correlated within households, so that each extra household virtually only adds one extra data point (like in the case of perfectly clustered EAs/segments). Would we want to estimate all three variables sufficiently precise in one survey, we would need sixty-five clusters (column 8).

The minimum number of clusters to obtain a precise enough one-stage cluster sample estimate are given in rows 2 to 14. Under the baseline scenario (scenario 1) these are slightly higher than the two-stage cluster sampling requirements. The average required number of clusters increase by 1.1 times to seventy clusters for the wealth indicator, increases by 1.2 times to eighteen clusters for contraception, and remain fifteen clusters for DPT3 vaccination status. This is not too different from the results for the case of perfect homogeneity within EAs (scenario 2), indicating that our baseline scenario is likely very similar to a perfect homogeneous within EA setting. In our model we assumed perfect within grid cell

Table 5.5: Required number of clusters - Baseline

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	65	-	15	-	15	-	65	-
1. Baseline	70	1.1	18	1.2	15	1.0	70	1.1
b. Extreme scenario's								
2. Homogeneous	70	1.1	18	1.2	16	1.0	70	1.1
3. Clustered by wealth index	282	4.3	19	1.3	17	1.2	282	4.3
4. Clustered by contraception	70	1.1	188	12.5	16	1.1	188	2.9
5. Clustered by DPT3	75	1.1	19	1.3	99	6.6	99	1.5
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	89	1.4	18	1.2	16	1.0	89	1.4
7. Clustered by improved water	75	1.1	18	1.2	16	1.1	75	1.1
8. Clustered by adequate structure	94	1.4	20	1.4	17	1.2	94	1.4
9. Clustered by adequate space	70	1.1	18	1.2	15	1.0	70	1.1
10. Clustered by non-solid fuel	70	1.1	18	1.2	16	1.1	70	1.1
d. Moderate clustering								
11. 50% replaced from 3.	141	2.2	18	1.2	17	1.1	141	2.2
12. 50% replaced from 4.	70	1.1	58	3.9	16	1.1	70	1.1
13. 50% replaced from 5.	70	1.1	19	1.3	43	2.8	70	1.1

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.
The minimum number of clusters is the average of the minimum number of clusters found in the five synthetic populations. The results for each of the five synthetic populations seperately can be found in Appendix 5.C.

homogeneity when assigning households to household locations. Since an EA usually consists of more than one grid cell this assumption does not need to lead to perfect within EA homogeneity. However, the different cells within one EA likely have rather similar spatial characteristics, so that they also get assigned a similar mix of households. Given this model, it is thus reasonable to expect moderate within EA homogeneity. However, households may also be located at certain locations for reasons that cannot be captured by observables in a model, so that also perfect within EA clustering is still a reasonable assumption.

Assuming perfect clustering by wealth index (scenario 3) increases the one-stage sample size requirements to on average 282 clusters. This is 4.3 times higher than the two-stage sample size requirements. Perfect clustering by contraception (scenario 4) increases the the sample size requirements to 188 clusters, an increase of 12.5 times the two-stage

sample size. Perfect clustering by DPT3 vaccination increases the sample size requirements from 15 to 99. In all cases the sample size requirements for the other variables are hardly affected. Although the sample size requirements are the largest for the wealth indicator, the increase in sample size requirements is starkest for contraception. The EA-level ICC on this variable is very low, so that clustering and segmenting within the EA can have large impacts. Also DPT3 vaccination has a low EA-level ICC, but contrary to contraception it has a high overall prevalence rate, so that even after within EA clustering and segmenting, the different EAs look pretty similar.[18] Even though perfect clustering leads to large increases in the sample size requirements for contraception and DPT3, the number of clusters needed for a complete household survey are affected only moderately. Clustering makes these variables look more similar to the wealth index in terms of ICC. Sample size requirements for the complete survey are at most 4.3 times the requirements for a two-stage cluster sample.

It may be more reasonable to assume that not wealth, contraception, and DPT3 are perfectly clustered, but that the underlying characteristics are perfectly clustered, i.e. if one does not have access to improved water, his or her neighbor probably does not either. Under that assumption the differences between one- and two-stage clustering are less extreme. DPT3 vaccination is hardly correlated with water, toilet, structure, space, or fuel, so that clustering on these variables is 'as if' there is complete homogeneity for the variable of interest. The wealth index is more strongly correlated with the underlying variables so that an increase in clusters is required to estimate the prevalence of poorest sufficiently precise, from 70 to at most 94 clusters. The largest difference is found when clustering by the adequacy of structure. When we assume that there is only moderate

[18]For the scenario 4 and 5 we assumed that the households without respectively women aged 15-49 or children under 5 (those with missing values on use of contraception/DPT3) were randomly located in between the households ordered by use of modern contraception/DPT3 vaccination. We could also assume that not only families with vaccinated children (women using modern contraception) live close together, but also families with young children (women aged 15-49) in general. When we would apply this assumption, the sample size requirements for contraception remain at 188 clusters, but the sample size requirements for DPT3 decrease to 65.

within EA clustering (scenario 11-13), the number of clusters increase slightly: a doubling of clusters in case of wealth, almost four times the amount in case of contraception, and almost tripling for DPT3.

The results for the different populations are comparable, though variation exist. For example, for the scenario of moderate clustering by contraception, increases in sample sizes lie between 3.1 and 4.7 times the sample size of two-stage cluster sampling. But the differences in panel c. are at most a factor 0.7.

5.4.2 Increasing sample sizes per cluster

The situation described above fits well to the case of multi-topic LMIC household surveys. However, in case of a topic specific survey focusing on a specific subsample of the population, for example immunization surveys, one may be tempted to enroll more households per cluster, because not every household will have an eligible household member.

Table 5.6 shows the required number of clusters for DPT3 vaccination under the assumption of (1) a two-stage cluster sample enrolling 25 households, (2) a one-stage cluster sample enrolling all households per cluster, (3) a one-stage cluster sample enrolling 50 households per cluster for the scenario's as described above, (4) a one-stage cluster sample enrolling 75 households per cluster for the scenario's as described above. Would one opt for generating segments of on average 50 households, rather than 25, eight clusters should be enrolled to achieve an accurate coverage estimate under the baseline scenario. In case of 75 households per segment this reduces to six clusters and would we enroll all households per cluster it reduces further to five clusters. The larger the segments, the less relevant the different scenario's become. As we have on average 86 households per EA, segments of size 75 are as if we are sampling the whole EA in most cases.

Extreme clustering by DPT3 status still results in a rather large sample size in the case where segments hold 50 households: 24 segments should be sampled, which is on average equal to 1200 households, three times

Table 5.6: Required number of clusters - increasing cluster size

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	5	0.4	1.2	-	-	-
1. Baseline	8	0.5	1.0	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	7	0.5	1.0	6	0.4	1.2
3. Clustered by wealth index	7	0.5	0.9	6	0.4	1.2
4. Clustered by contraception	7	0.5	1.0	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	8	0.5	1.0	6	0.4	1.2
7. Clustered by improved water	8	0.5	1.1	6	0.4	1.2
8. Clustered by adequate structure	7	0.5	1.0	6	0.4	1.2
9. Clustered by adequate space	7	0.5	0.9	6	0.4	1.2
10. Clustered by non-solid fuel	7	0.5	0.9	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.9	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.9	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.
The minimum number of clusters is the average of the minimum number of clusters found in the five synthetic populations. The results for each of the five synthetic populations seperately can be found in Appendix 5.C.

more than in the baseline two-stage sample. The number of households at most doubles when taking segments of 75 households.

5.5 Discussion

The results provide us with guidelines of the possible requirements of one-stage cluster sampling under extreme situations. Whereas the variety within the modeled region (Oshikoto, Namibia) makes it an interesting case to look at, there are also some limitations to the region, which may result in more moderate outcomes then we would find elsewhere. First of all, the EAs in Oshikoto are relatively small, holding on average 86 households. Typical EAs in other countries consist of 200 or even 400/500 households. In larger EAs, the effects of clustering on the sample size requirements of one-stage cluster sampling could potentially be larger. Secondly, whereas there are relatively many households in Oshikoto falling in the poorest wealth category, none of these seem to live in the urban areas. In many other LMIC settings, you would expect to find the poorest in cities, possibly leading to more extreme clustering effects.

Another limitation of the analysis is that it does inform us about what could happen in different extreme situations, but not how likely it is to encounter such situations. Although literature gives guidance on likely levels of ICCs and more general spatial patterns, little is written about how households are located within EAs. Intuitively, the amount of within EA clustering will depend on the type of EA. One may for example expect more clustering in an EA that covers a rural village, then in an EA covering multiple farms or a city block. Similarly, it would not be unreasonable to assume that wealth levels vary from street to street, due to the difference in types of houses, whereas contraception and vaccination levels are less likely to be linked to such specific locations. There does exist a line of research devoted to the effect of scale on measurements of racial segregations in large metropolitan cities in industrial countries where there is a higher availability of geo-coded micro-data. This research indicates that there is some variation between egocentric measurements based on a 100m radius versus a 1000m radius around households or individuals (Petrović et al. 2018), and slight variations in egocentric measurements including the 50 nearest neighbors compared to 200 nearest neighbors of households or individuals (Östh et al. 2015). However, it is not clear

how these results would generalize to (rural) LMIC settings or to socio-economic and health variables.

5.6 Conclusion

In this chapter we compared the commonly used two-stage cluster sampling approach to the alternative one-stage cluster sampling approach. We generated a synthetic population of Oshikoto, Namibia to provide as a testing ground for both sampling approaches. The households in this population were assigned to realistic (x,y) coordinates, in order to simulate realistic spatial patterns of the different household characteristics. To facilitate one-stage cluster sampling we created smaller Primary Sampling Units by segmenting Enumeration Areas (EAs) under different scenarios of within EA clustering. We searched for the minimum number of clusters to obtain an adequate sample in an iterative way based on bootstrapped confidence intervals of the sample means.

The results show that in most moderate scenario's the required number of clusters for one-stage cluster sampling is fewer than twice the required number of clusters of two-stage sampling, under the assumption that the same number of households per cluster are sampled with both methods. Under extreme clustering scenarios, the required number of clusters can increase by up to thirteen times. Especially when the EA-level intracluster correlation is moderate and prevalence is close to 50%, extreme assumptions about within EA clustering have large impact on the required number of clusters. When measuring variables focusing on small subsamples of the population, it can be beneficial to apply one-stage cluster sampling with larger segment sizes. This can lower the required number of clusters to visit, while enrolling the same number of households in the survey. By increasing segment sizes, clustering scenarios also become less relevant.

Whether one-stage or two-stage clustering is more cost-effective will depend on the type of survey and the regional context. Aspects such as the length of the survey, the type of survey (does it only include a questionnaire or also the collection of biomarkers?), and the accessibility of the regions will determine the relative costs of the two types of sampling

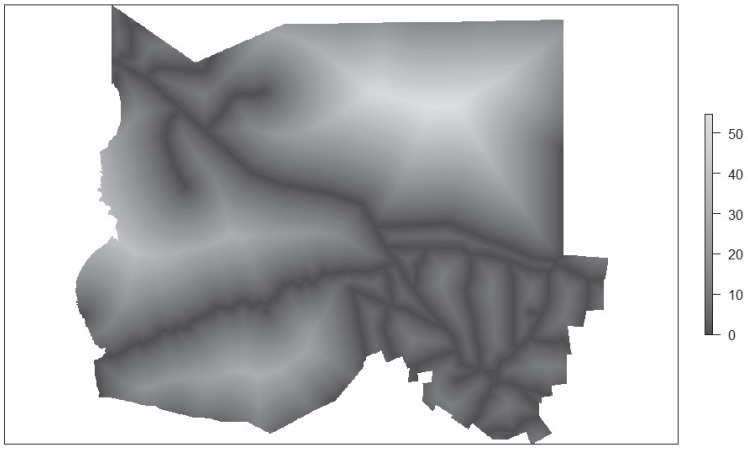
methods. The numbers provided in this chapter can serve as input when estimating these costs.

Maps of Oshikoto 5.A

Figure 5.6: Population density in Oshikoto expressed in people per pixel (roughly $100m^2$).

source: www.worldpop.org, Linard et al. (2012)

Figure 5.7: Distance to major roads (km).



source: Spatial covariate processed by the "Global High Resolution Population Denominators" Project (original source: 2016 OSM highways).

5.B Data description for synthetic populations 2-5

Table 5.7: summary statistics for synthetic populations 2-5

	pop 2	pop 3	pop 4	pop 5
nr households	37298	37298	37298	37298
average household size	4.82	4.83	4.83	4.83
urban	15.7%	15.7%	15.7%	15.7%
education head				
- no formal	24.9%	25.1%	24.8%	25.0%
- incomplete primary	30.0%	30.5%	30.3%	30.2%
- complete primary	31.2%	30.5%	31.1%	31.1%
- complete secondary	10.2%	10.6%	10.2%	10.5%
- complete tertiary	3.6%	3.4%	3.5%	3.3%
unimproved water	26.8%	27.1%	26.9%	27.3%
unimproved toilet	79.9%	80.1%	80.3%	80.5%
inadequate space	7.8%	7.7%	7.5%	7.7%
inadequate structure	61.3%	61.4%	61.5%	62.1%
solid fuel	83.6%	84.0%	84.0%	84.1%
wealth index				
- poorest	33.6%	34.0%	34.1%	34.4%
- poorer	28.7%	28.5%	28.6%	28.4%
- middle	19.6%	19.6%	19.2%	19.3%
- richer	13.4%	13.1%	13.7%	13.4%
- richest	4.7%	4.6%	4.3%	4.4%
nr individuals	179854	180233	180164	180111
male	48.0%	48.1%	48.2%	48.0%
age:				
- 0 - 4	14.2%	14.3%	14.0%	14.1%
- 5 - 14	26.2%	26.1%	26.1%	26.3%
- 15 - 49	46.2%	46.1%	46.4%	46.1%
- 50 plus	13.5%	13.5%	13.5%	13.5%
nr women 15-49	43007	42930	42903	42708
modern contraception	0.4350222	0.4329839	0.438967	0.436007
nr children under 5	25466	25742	25258	25463
DPT3 vaccination	80.4%	80.6%	80.2%	80.2%

Figure 5.8: Prevalence of characteristics per EA - Synthetic population 2

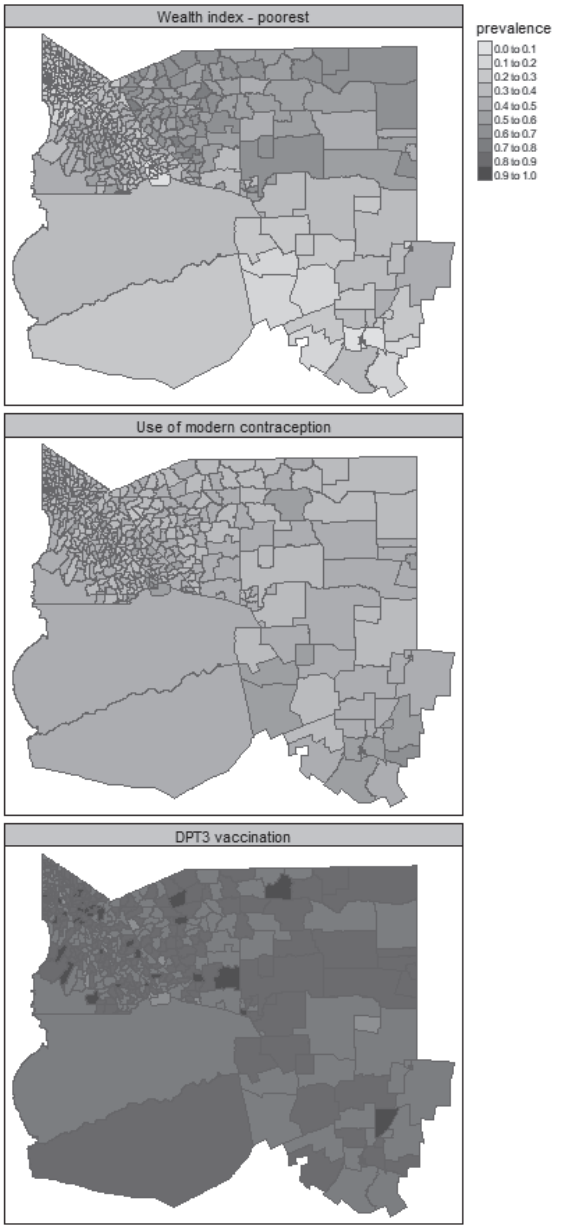


Figure 5.9: Prevalence of characteristics per EA - Synthetic population 3

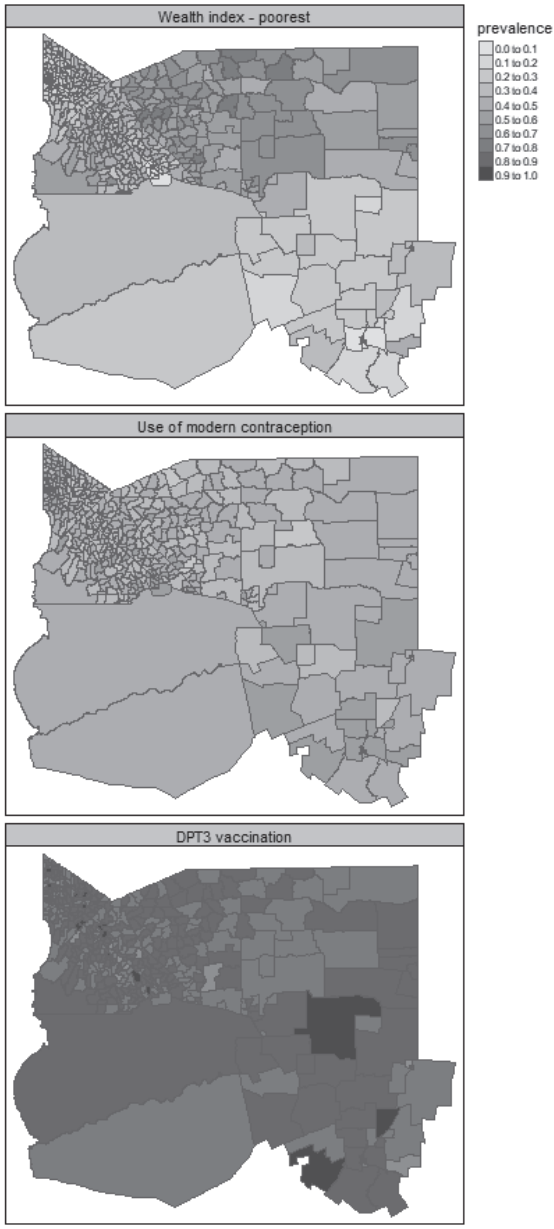


Figure 5.10: Prevalence of characteristics per EA - Synthetic population 4

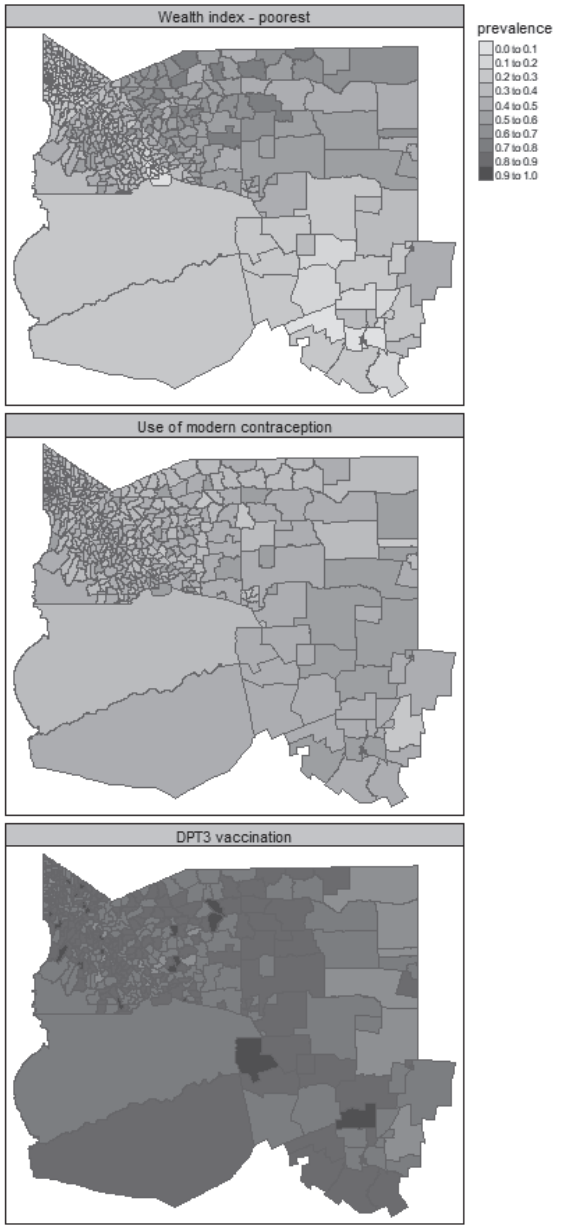


Figure 5.11: Prevalence of characteristics per EA - Synthetic population 5

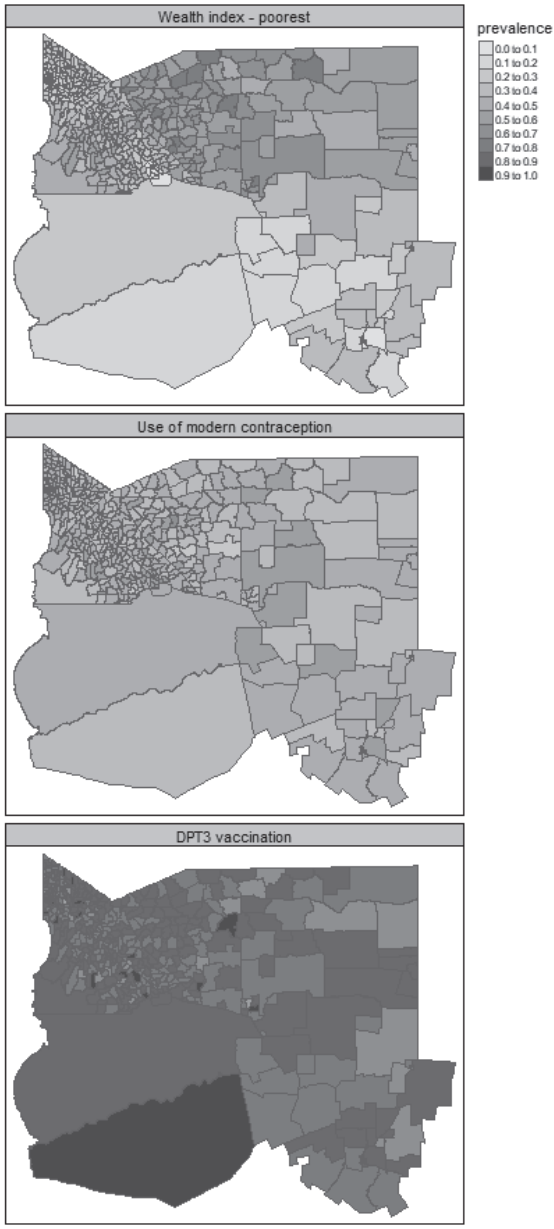


Figure 5.12: ICC (♦) and boxplots of prevalences per EA/segment - Synthetic population 2

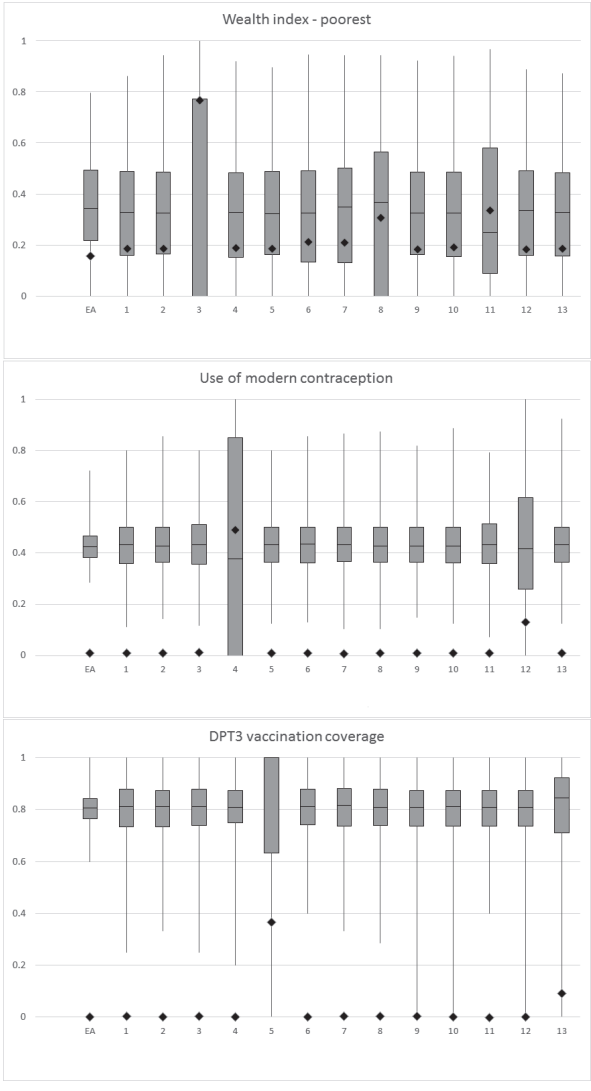


Figure 5.13: ICC (♦) and boxplots of prevalences per EA/segment -Synthetic population 3

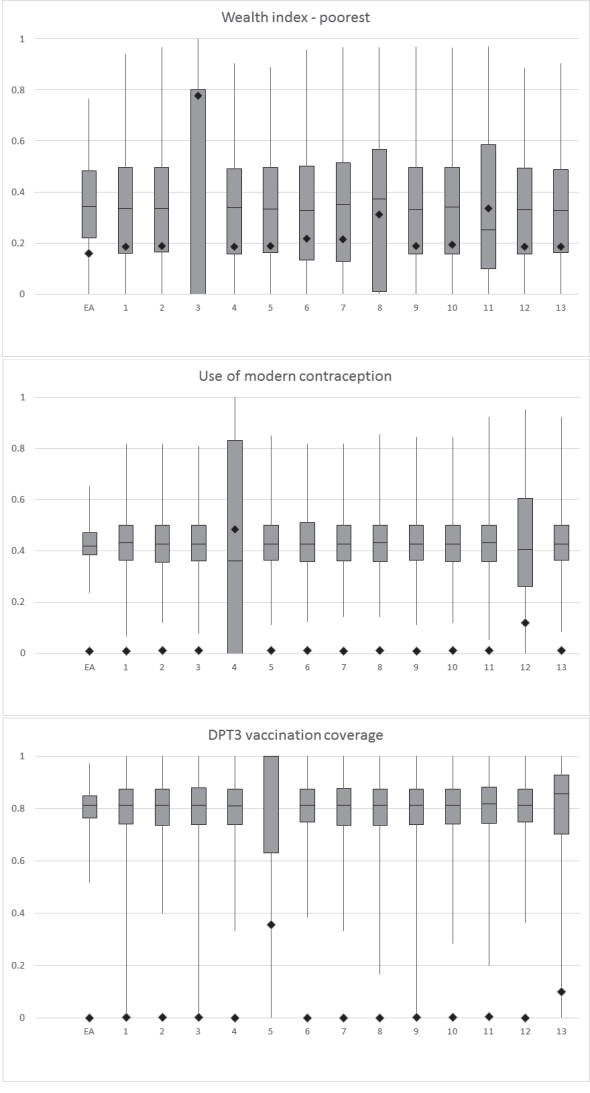


Figure 5.14: ICC (♦) and boxplots of prevalences per EA/segment - Synthetic population 4

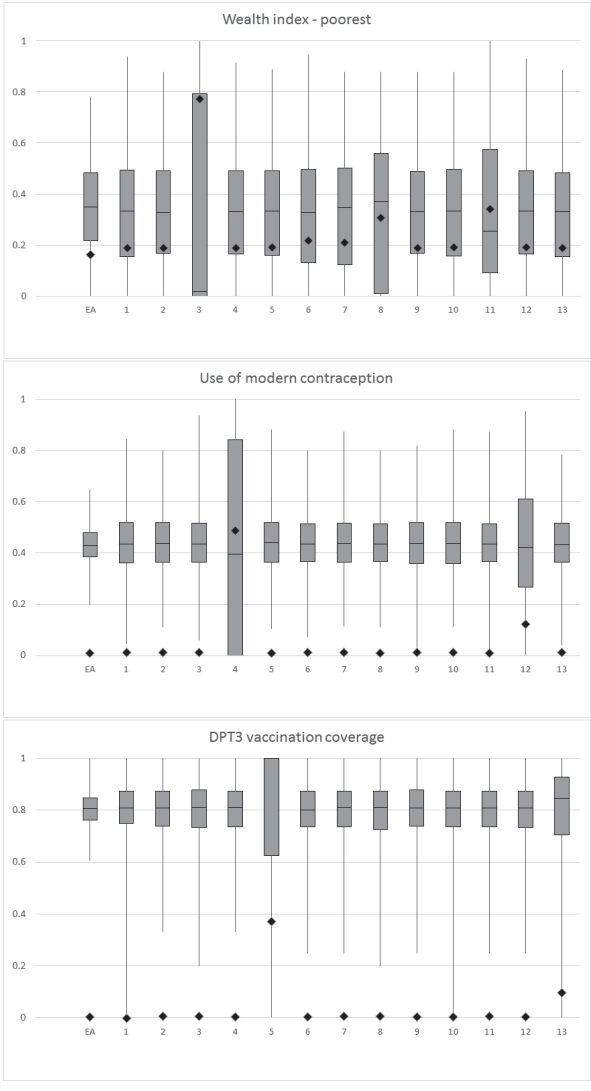
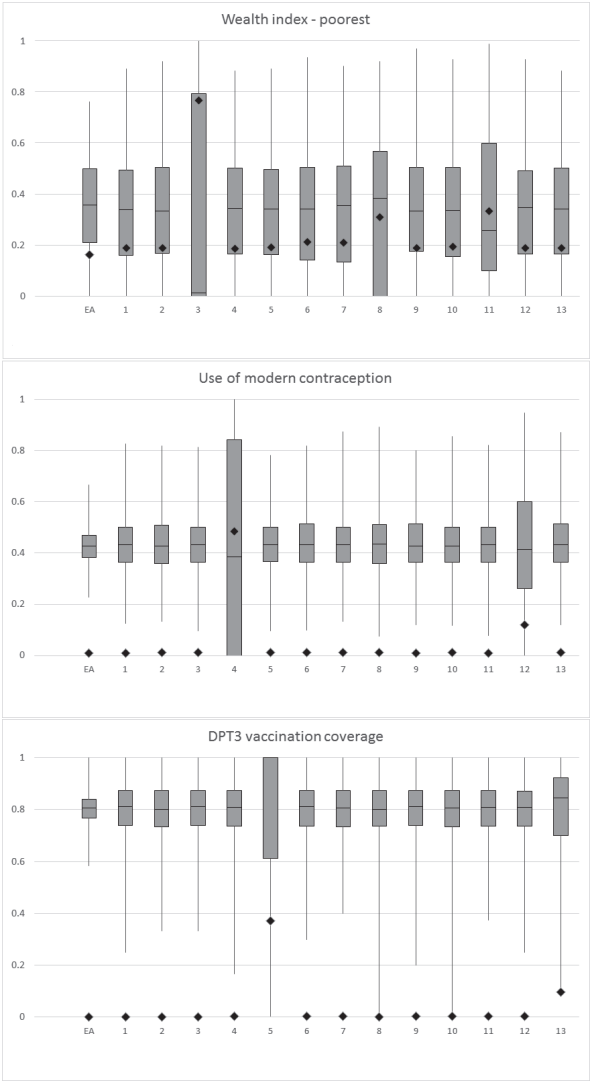


Figure 5.15: ICC (♦) and boxplots of prevalences per EA/segment - Synthetic population 5



5.C Results for each of the 5 synthetic populations

5.C.1 Baseline results

Table 5.8: Required number of clusters - Baseline - Population 1

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	70	-	15	-	15	-	70	-
1. Baseline	70	1.0	18	1.2	12	0.8	70	1.0
b. Extreme scenario's								
2. Homogeneous	70	1.0	18	1.2	18	1.2	70	1.0
3. Clustered by wealth index	282	4.0	18	1.2	18	1.2	282	4.0
4. Clustered by contraception	70	1.0	188	12.5	15	1.0	188	2.7
5. Clustered by DPT3	94	1.3	18	1.2	94	6.3	94	1.3
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	94	1.3	18	1.2	18	1.2	94	1.3
7. Clustered by improved water	94	1.3	18	1.2	15	1.0	94	1.3
8. Clustered by adequate structure	94	1.3	24	1.6	18	1.2	94	1.3
9. Clustered by adequate space	70	1.0	18	1.2	18	1.2	70	1.0
10. Clustered by non-solid fuel	70	1.0	18	1.2	15	1.0	70	1.0
d. Moderate clustering								
11. 50% replaced from 3.	141	2.0	18	1.2	18	1.2	141	2.0
12. 50% replaced from 4.	70	1.0	58	3.9	18	1.2	70	1.0
13. 50% replaced from 5.	70	1.0	18	1.2	36	2.4	70	1.0

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.9: Required number of clusters - Baseline - Population 2

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	70	-	15	-	15	-	70	-
1. Baseline	70	1.0	18	1.2	18	1.2	70	1.0
b. Extreme scenario's								
2. Homogeneous	70	1.0	18	1.2	12	0.8	70	1.0
3. Clustered by wealth index	282	4.0	18	1.2	18	1.2	282	4.0
4. Clustered by contraception	70	1.0	188	12.5	18	1.2	188	2.7
5. Clustered by DPT3	70	1.0	18	1.2	94	6.3	94	1.3
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	94	1.3	18	1.2	18	1.2	94	1.3
7. Clustered by improved water	70	1.0	18	1.2	18	1.2	70	1.0
8. Clustered by adequate structure	94	1.3	18	1.2	18	1.2	94	1.3
9. Clustered by adequate space	70	1.0	18	1.2	15	1.0	70	1.0
10. Clustered by non-solid fuel	70	1.0	18	1.2	15	1.0	70	1.0
d. Moderate clustering								
11. 50% replaced from 3.	141	2.0	18	1.2	15	1.0	141	2.0
12. 50% replaced from 4.	70	1.0	70	4.7	15	1.0	70	1.0
13. 50% replaced from 5.	70	1.0	18	1.2	47	3.1	70	1.0

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.10: Required number of clusters - Baseline - Population 3

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	47	-	15	-	15	-	47	-
1. Baseline	70	1.5	18	1.2	12	0.8	70	1.5
b. Extreme scenario's								
2. Homogeneous	70	1.5	18	1.2	12	0.8	70	1.5
3. Clustered by wealth index	282	6.0	18	1.2	15	1.0	282	6.0
4. Clustered by contraception	70	1.5	188	12.5	12	0.8	188	4.0
5. Clustered by DPT3	70	1.5	18	1.2	94	6.3	94	2.0
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	94	2.0	18	1.2	12	0.8	94	2.0
7. Clustered by improved water	70	1.5	18	1.2	12	0.8	70	1.5
8. Clustered by adequate structure	94	2.0	24	1.6	18	1.2	94	2.0
9. Clustered by adequate space	70	1.5	18	1.2	12	0.8	70	1.5
10. Clustered by non-solid fuel	70	1.5	18	1.2	15	1.0	70	1.5
d. Moderate clustering								
11. 50% replaced from 3.	141	3.0	18	1.2	15	1.0	141	3.0
12. 50% replaced from 4.	70	1.5	47	3.1	12	0.8	70	1.5
13. 50% replaced from 5.	70	1.5	18	1.2	47	3.1	70	1.5

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.11: Required number of clusters - Baseline - Population 4

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	70	-	15	-	15	-	70	-
1. Baseline	70	1.0	18	1.2	15	1.0	70	1.0
b. Extreme scenario's								
2. Homogeneous	70	1.0	18	1.2	18	1.2	70	1.0
3. Clustered by wealth index	282	4.0	18	1.2	18	1.2	282	4.0
4. Clustered by contraception	70	1.0	188	12.5	18	1.2	188	2.7
5. Clustered by DPT3	70	1.0	18	1.2	94	6.3	94	1.3
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	94	1.3	18	1.2	18	1.2	94	1.3
7. Clustered by improved water	70	1.0	18	1.2	18	1.2	70	1.0
8. Clustered by adequate structure	94	1.3	18	1.2	18	1.2	94	1.3
9. Clustered by adequate space	70	1.0	18	1.2	12	0.8	70	1.0
10. Clustered by non-solid fuel	70	1.0	18	1.2	18	1.2	70	1.0
d. Moderate clustering								
11. 50% replaced from 3.	141	2.0	18	1.2	18	1.2	141	2.0
12. 50% replaced from 4.	70	1.0	70	4.7	18	1.2	70	1.0
13. 50% replaced from 5.	70	1.0	24	1.6	47	3.1	70	1.0

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.12: Required number of clusters - Baseline - Population 5

	poorest		contraception		DPT3		all three	
Scenario	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff	nr. clusters	rel. diff
a. Baseline								
two-stage	70	-	15	-	15	-	70	-
1. Baseline	70	1.0	18	1.2	18	1.2	70	1.0
b. Extreme scenario's								
2. Homogeneous	70	1.0	18	1.2	18	1.2	70	1.0
3. Clustered by wealth index	282	4.0	24	1.6	18	1.2	282	4.0
4. Clustered by contraception	70	1.0	188	12.5	18	1.2	188	2.7
5. Clustered by DPT3	70	1.0	24	1.6	118	7.9	118	1.7
c. Clustering on underlying characteristics								
6. Clustered by improved toilet	70	1.0	18	1.2	12	0.8	70	1.0
7. Clustered by improved water	70	1.0	18	1.2	18	1.2	70	1.0
8. Clustered by adequate structure	94	1.3	18	1.2	15	1.0	94	1.3
9. Clustered by adequate space	70	1.0	18	1.2	18	1.2	70	1.0
10. Clustered by non-solid fuel	70	1.0	18	1.2	18	1.2	70	1.0
d. Moderate clustering								
11. 50% replaced from 3.	141	2.0	18	1.2	18	1.2	141	2.0
12. 50% replaced from 4.	70	1.0	47	3.1	18	1.2	70	1.0
13. 50% replaced from 5.	70	1.0	18	1.2	36	2.4	70	1.0

The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Additional results 5.C.2

Table 5.13: Required number of clusters - Increasing cluster size - Population 1

	50 households per segment			75 households per segment		
		rel. diff			rel. diff	
Scenario	nr. clusters	clusters	households[a]	nr. clusters	clusters	households[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	4	0.3	0.9	-	-	-
1. Baseline	9	0.6	1.2	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	6	0.4	0.8	6	0.4	1.2
3. Clustered by wealth index	8	0.5	1.1	5	0.3	1.0
4. Clustered by contraception	6	0.4	0.8	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	8	0.5	1.1	6	0.4	1.2
7. Clustered by improved water	6	0.4	0.8	6	0.4	1.2
8. Clustered by adequate structure	6	0.4	0.8	6	0.4	1.2
9. Clustered by adequate space	6	0.4	0.8	6	0.4	1.2
10. Clustered by non-solid fuel	6	0.4	0.8	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.8	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.8	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.14: Required number of clusters - Increasing cluster size - Population 2

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	5	0.3	1.1	-	-	-
1. Baseline	9	0.6	1.2	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	9	0.6	1.2	6	0.4	1.2
3. Clustered by wealth index	9	0.6	1.2	6	0.4	1.2
4. Clustered by contraception	6	0.4	0.8	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	9	0.6	1.2	6	0.4	1.2
7. Clustered by improved water	9	0.6	1.2	6	0.4	1.2
8. Clustered by adequate structure	8	0.5	1.1	6	0.4	1.2
9. Clustered by adequate space	8	0.5	1.1	6	0.4	1.2
10. Clustered by non-solid fuel	6	0.4	0.8	5	0.3	1.0
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.8	6	0.4	1.2
12. 50% replaced from 4.	8	0.5	1.1	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.15: Required number of clusters - Increasing cluster size - Population 3

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	6	0.4	1.3	-	-	-
1. Baseline	6	0.4	0.8	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	6	0.4	0.8	6	0.4	1.2
3. Clustered by wealth index	6	0.4	0.8	6	0.4	1.2
4. Clustered by contraception	9	0.6	1.2	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	6	0.4	0.8	6	0.4	1.2
7. Clustered by improved water	8	0.5	1.1	6	0.4	1.2
8. Clustered by adequate structure	8	0.5	1.1	6	0.4	1.2
9. Clustered by adequate space	6	0.4	0.8	5	0.3	1.0
10. Clustered by non-solid fuel	8	0.5	1.1	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	8	0.5	1.1	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.8	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.16: Required number of clusters - Increasing cluster size - Population 4

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	6	0.4	1.3	-	-	-
1. Baseline	9	0.6	1.2	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	9	0.6	1.2	6	0.4	1.2
3. Clustered by wealth index	6	0.4	0.8	6	0.4	1.2
4. Clustered by contraception	9	0.6	1.2	6	0.4	1.2
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	9	0.6	1.2	6	0.4	1.2
7. Clustered by improved water	9	0.6	1.2	6	0.4	1.2
8. Clustered by adequate structure	9	0.6	1.2	6	0.4	1.2
9. Clustered by adequate space	9	0.6	1.2	6	0.4	1.2
10. Clustered by non-solid fuel	9	0.6	1.2	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.8	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.8	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

Table 5.17: Required number of clusters - Increasing cluster size - Population 5

	50 households per segment			75 households per segment		
	nr.	rel. diff		nr.	rel. diff	
Scenario	clusters	clusters	HHs[a]	clusters	clusters	HHs[a]
a. Baseline						
two-stage (25 HHs)	15	-	-	-	-	-
one-stage (all HHs)	6	0.4	1.3	-	-	-
1. Baseline	6	0.4	0.8	6	0.4	1.2
b. Extreme scenario's						
2. Homogeneous	6	0.4	0.8	6	0.4	1.2
3. Clustered by wealth index	6	0.4	0.8	6	0.4	1.2
4. Clustered by contraception	7	0.5	0.9	5	0.3	1.0
5. Clustered by DPT3	24	1.6	3.2	9	0.6	1.8
c. Clustering on underlying characteristics						
6. Clustered by improved toilet	6	0.4	0.8	6	0.4	1.2
7. Clustered by improved water	8	0.5	1.1	6	0.4	1.2
8. Clustered by adequate structure	6	0.4	0.8	6	0.4	1.2
9. Clustered by adequate space	6	0.4	0.8	6	0.4	1.2
10. Clustered by non-solid fuel	6	0.4	0.8	6	0.4	1.2
d. Moderate clustering						
11. 50% replaced from 3.	6	0.4	0.8	6	0.4	1.2
12. 50% replaced from 4.	6	0.4	0.8	6	0.4	1.2
13. 50% replaced from 5.	12	0.8	1.6	6	0.4	1.2

[a] Calculated as the number of clusters times 50, 75 households, or the average number of households per EA (for row 2) divided by the number of clusters under a two-stage design times 25 households.
The minimum number of clusters are based on bootstrapped sample means and based on the requirement that 95% of sample means should fall within 5 percentage point from the population mean.

5.D Additional statistics

Table 5.18: Observed ICCs in DHS Namibia 2013

	Poorest	Contraception	DPT3
Namibia	0.4142	0.0396	0.0885
Caprivi	0.4015	0.0008	0.0221
Erongo	0.0302	0.0380	0.0959
Hardap	0.2799	0.0194	0.0030
Karas	0.3624	0.0191	0.1197
Kavango	0.3842	0.0379	0.0724
Khomas	0.1324	0.0053	0.0486
Kunene	0.3179	0.0683	0.2407
Ohangwena	0.2694	0.0371	0.0677
Omaheke	0.1394	-0.0055	0.0504
Omusati	0.1487	0.0457	0.0116
Oshana	0.1772	0.0541	-0.0460
Oshikoto	0.2314	-0.0062	0.0237
Otjozondjupa	0.2132	-0.0043	0.1021

Bibliography

United Nations Human Settlements Programme (UN-HSP) (2003): *The challenge of slums: Global report on human settlements 2003*, chapter 1. Development context and the millennium agenda - Revised and updated version (April 2010), pages 5–16. Earthscan Publications Ltd. Cited on page 184.

Agarwal, S., C. Liu, and N. S. Souleles (2007): The reaction of consumer spending and debt to tax rebates – evidence from consumer credit data. *Journal of Political Economy*, 115(6):986–1019. Cited on page 28.

Ando, A. and F. Modigliani (1963): The 'life cycle' hypothesis of saving: Aggregate implications and tests. *American Economic Review*, 53(1):55–84. Cited on page 10.

Andren, D. (2014): Does part-time sick leave help individuals with mental disorders recover lost work capacity? *Journal of Occupational Rehabilitation*, 24:344–360. Cited on pages 118, 122, and 149.

Andren, D. and M. Svensson (2012): Part-time sick leave as a treatment method for individuals with musculoskeletal disorders. *Journal of Occupational Rehabilitation*, 22:418–426. Cited on pages 118 and 119.

Angrisani, M., M. Hurd, and S. Rohwedder (2015): The Effect of Housing and Stock Wealth Losses on Spending in the Great Recession. CESR-Schaeffer Working Paper Series 2015-017. Cited on pages 10 and 28.

Attanatio, O., L. Blow, R. Hamilton, and A. Leicester (2009): Booms and busts: consumption, house prices and expectations. *Economica*, 76(301):20–50. Cited on page 10.

Attanatio, O., A. Leicester, and M. Wakefield (2011): Do House Prices Drive Consumption Growth? The Coincident Cycles of House Prices and Consumption in the UK. *Journal of the European Economic Association*, 9(3):399–435. Cited on page 19.

Banks, J., R. Blundell, P. Levell, and J. P. Smith (2015): Life-cycle consumption patterns at older ages in the US and the UK: can medical expenditures explain the difference? IFS Working Paper W15/12. Cited on pages 76, 85, and 104.

Bennett, S., A. Radalowicz, V. Vella, and A. M. Tomkins (1994): A computer simulation of household sampling schemes for health surveys in developing countries. *International Journal of Epidemiology*, 23:1282–1291. Cited on page 180.

Bennett, S., T. Woods, W. M. Liyanage, and D. L. Smith (1991): A simplified general method for cluster-sample surveys of health in developing countries. *World Health Statistics Quarterly*, 44:98–106. Cited on page 192.

Bernacki, E. J., J. A. Guidera, J. A. Schaefer, and S. Tsai (2000): A facilitated early return to work program at a large urban medical center. *Journal of Occupational and Environmental Medicine*, 42(12):1172–1177. Cited on page 118.

Bethge, M. (2016): Effects of graded return-to-work: a propensity-score-matched analysis. *Scandinavian Journal of Work Environment and Health*, 42(4):273–279. Cited on pages 118 and 119.

Binswanger, J., D. Schunk, and V. Toepoel (2013): Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly*, 77(3):783–797. Cited on page 20.

Börsch-Supan, A. and K. Stahl (1991): Life cycle savings and consumption constraints. *Journal of Population Economics*, 4(3):233–255. Cited on page 74.

Bovenberg, A. L. and L. Meijdam (2001): The Dutch pension system. In A. H. Borsch-Supan and M. Miegel, editors, *Pension reform in six countries*, pages 39–67. Springer, New York. Cited on page 15.

Breiman, L. (2001): Random Forests. *Machine Learning*, 45:5–32. Cited on page 185.

Browning, M. and M. Collado (2001): The response of expenditures to anticipated income changes: panel data estimates. *American Economic Review*, 91(3):681–692. Cited on pages 14 and 34.

Butrica, B. A., R. W. Joghnson, and G. B. Mermin (2009): Do health problems reduce consumption at older ages? Center for Retirement Research Working Paper 2009-9. Cited on page 76.

Campbell, J. Y. (1991): A Variance Decomposition for Stock Returns. *The Economic Journal*, 101(405):157–179. Cited on page 29.

Campbell, J. Y. and J. F. Cocco (2007): How do house prices affect consumption? Evidence from micro data. *Journal of Monetary Economics*, 54(3):591–621. Cited on pages 13 and 19.

Ferrer-i Carbonell, A. and P. Frijters (2004): How important is methodology for the estimates of the determinants of happiness? *Economic Journal*, 114(497):641–659. Cited on page 85.

Christelis, D. (2011): Imputation of missing data in waves 1 and 2 of SHARE. SHARE Working Paper. Cited on page 91.

Christelis, D., D. Georgarakos, and T. Jappelli (2015): Wealth shocks, unemployment shocks and consumption in the wake of the Great Recession. *Journal of Monetary Economics*, 72:21–41. Cited on pages 10, 28, and 34.

Coile, C. C. and P. B. Levine (2006): Bulls, bears, and retirement behavior. *ILR Review*, 59(3):408–429. Cited on page 11.

Corrigan, P. W. and S. G. McCracken (2005): Place first, then train: an alternative to the medical model of psychiatric rehabilitation. *Social Work*, 50(1):31–39. Cited on page 118.

Coulibaly, B. and G. Li (2006): Do homeowners increase consumption after the last mortgage payment? An alternative test of the permanent income hypothesis. *Review of Economics and Statistics*, 88(1):10–19. Cited on page 34.

Craik, F. I. and C. L. Grady (2002): *Principles of frontal lobe function*, chapter Aging, memory, and frontal lobe functioning, pages 528–540. Oxford University Press. Cited on page 99.

Crawford, R. (2013): The effect of the financial crisis on the retirement plans of older workers in England. *Economics Letters*, 121(2):156–159. Cited on page 11.

Cylus, J. and I. Papanicolas (2015): An analysis of perceived access to health care in Europe: How universal is universal coverage? *Health Policy*, 119(9):1133–1144. Cited on page 94.

De Bresser, J. and M. Knoef (2015): Can the Dutch meet their own retirement expenditure goals? *Labour Economics*, 34:100–117. Cited on pages 11, 21, 25, 32, 62, and 65.

De Bresser, J., M. Knoef, and L. Kools (2018): Cutting one's coat according to one's cloth. How did the Great Recession affect retirement resources and expenditure goals? Netspar Academic Series DP 05/2018-029. Cited on page 9.

De Nardi, M., E. French, and J. B. Jones (2010): Why do the elderly save? The role of medical expenses. *Journal of Political Economy*, 118(1):39–75. Cited on page 75.

Dean, D., J. Pepper, R. Schmidt, and S. Stern (2015): The effects of vocational rehabilitation for people with cognitive impairments. *International Economic Review*, 56:399–426. Cited on pages 119 and 136.

DigitalGlobe (2014): Quickbird 50cm imagery. Overview. [Online]. Available: http://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9. [Accessed: 01-Feb-2018]. Cited on page 183.

Disney, R., J. Gathergood, and A. Henley (2010): House price shocks, negative equity, and household consumption in the United Kingdom. *Journal of the European Economic Association*, 8(6):1179–1207. Cited on pages 10 and 28.

DNB (2009): A closer look at pension funds' investment policies. Dutch Central Bank Querterly Bulletin December 2009. Cited on pages xvii and 43.

——— (2014): Investment mix and interest rate hedging decisive for pension fund recovery. Dutch Central Bank Bulletin May 2014. Cited on pages 16 and 43.

——— (2016): Financiële positie Pensioenfondsen. Dutch Central Bank May 2016. Cited on page 16.

Domeij, D. and M. Johannesson (2006): Consumption and health. *Contributions in Macroeconomics*, 6(1):1–30. Cited on page 74.

Duggan, M. (2005): Do new prescription drugs pay for themselves? The case of second-generation antipsychotics. *Journal of Health Economics*, 24:1–31. Cited on page 136.

Dutch Association of Insurers (2016): Dutch Insurance Industry in Figures 2016. Cited on page 125.

Dynan, K., J. Skinner, and S. Zeldes (2004): Do the rich save more? *Journal of Political Economy*, 112:397–444. Cited on page 37.

Engen, E. M., W. G. Gale, and C. E. Uccello (2005): Effects of stock market fluctuations on the adequacy of retirement wealth accumulation. *Review of Income and Wealth*, 2005(3):397–418. Cited on page 13.

Evans, W. N. and W. K. Viscusi (1991): Estimation of state-dependent utility functions using survey data. *Review of Economics and Statistics*, 73(1):94–104. Cited on pages 77 and 78.

Everhardt, T. P. and P. R. de Jong (2011): Return to work after long term sickness. The role of employer based interventions. *De Economist*, 159:361–380. Cited on page 125.

Fink, G., I. Günther, and K. Hill (2014): Slum residence and child health in developing countries. *Demography*, 51:1175–1197. Cited on page 184.

Finkelstein, A., E. F. Luttmer, and M. J. Notowidigdo (2009): Approaches to estimating the health state dependence of the utility function. *American Economic Review*, 99(2):116–121. Cited on pages 75, 76, and 85.

——— (2013): What good is wealth without health? The effect of health on the marginal utility of consumption. *Journal of the European Economic Association*, 11(1):221–258. Cited on pages 74, 75, 76, 77, 80, 84, 90, 102, 103, 105, and 109.

Fischer, J. A. V. and A. Sousa-Poza (2008): Personality, job satisfaction and health - The mediating influence of affectivity. *Swiss Journal of Economics and Statistics*, 144:379–435. Cited on page 113.

French, E. and J. Song (2014): The effect of Disability Insurance receipt on labor supply. *American Economic Journal: Economic Policy*, 6:291–337. Cited on page 136.

Galway, L. P., N. Bell, S. A. E. A. Shatari, A. Hagopian, G. Burnham, A. Flaxman, W. M. Weiss, J. Rajaratnam, and T. K. Takaro (2012): A two-stage cluster sampling method using gridded population data, a GIS, and Google Earth(TM) imagery in a population-based mortality survey in Iraq. *International Journal of Health Geographies*, 11:12. Cited on page 179.

Glisky, E. L. (2007): *Brain aging: models, methods, and mechanisms*, chapter Changes in cognitive function in human age. CRC Press/Taylor & Francis. Cited on page 111.

Goda, G. S., J. Shoven, and S. N. Slavov (2011): What explains changes in retirement plans during the great recession. *American Economic Review*, 101:29–34. Cited on page 11.

Goda, G. S., J. B. Shoven, and S. N. Slavov (2012): Does stock market performance influence retirement intentions? *Journal of Human Resources*, 47(4):1055–1081. Cited on page 11.

Goudswaard, K. (2011): De pensioenleeftijd in beweging. *Jaarverslag Stichting Instituut GAK*, pages 5–8. Cited on page 53.

Hartigan, J. A. and M. Wong (1979): A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108. Cited on page 185.

Haveman, R., K. Holden, A. Romanov, and B. Wolfe (2007): Assessing the maintenance of savings sufficiency over the first decade of retirement. *International Tax and Public Finance*, 14(4):481–502. Cited on page 13.

Hernæs, Ø. (2017): Activation against absenteeism: Evidence from a sickness insurance reform in Norway. IZA Discussion Paper No. 10991. Cited on pages 118 and 149.

Himelein, K., S. Eckman, and S. Murray (2014): Sampling nomads: A new technique for remote, hard-to-reach, and mobile populations. *Journal of Official Statistics*, 30(2):191–213. Cited on page 179.

Høgelund, J., A. Holm, and L. F. Eplov (2012): The effect of part-time sick leave for employees with mental disorders. *The Journal of Mental Health Policy and Economics*, 15:157–170. Cited on page 119.

Høgelund, J., A. Holm, and J. McIntosh (2010): Does graded return-to-work improve sick-listed workers' chance of returning to regular working hours? *Journal of Health Economics*, 29:158–169. Cited on pages 118, 119, and 149.

Hsieh, C.-T. (2003): Do consumers react to anticipated income changes? Evidence from the Alaska permanent fund. *American Economic Review*, 93(1):397–405. Cited on page 34.

Hurd, M. D., M. Reti, and S. Rohwedder (2009): The effect of large capital gains or losses on retirement. In *Developments in the Economics of Aging*, pages 127–163. University of Chicago Press. Cited on page 11.

Johnson, D., J. Parker, and N. Souleles (2006): Household expenditure and the income tax rebates of 2001. *American Economic Review*, 96(5):1589–1610. Cited on pages 28 and 37.

de Jong, P., M. Gielen, and V. Haanstra-Veldhuis (2014): Verzekeringsgraad kleine werkgevers - eindrapportage. Research commissioned by the ministry of social affairs. Cited on page 124.

de Jong, P., M. Lindeboom, and B. van der Klaauw (2011): Screening Disability Insurance applications. *Journal of the European Economics Association*, 9(1):106–129. Cited on page 122.

Kahneman, D. and A. Deaton (2010): High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences*, 107(38):16489–16493. Cited on page 102.

Kalmijn, M. (2006): Educational Inequality and Family Relationships: Influences on Contact and Proximity. *European Sociological Review*, 22:1–16. Cited on page 84.

Kausto, J., H. Miranda, K.-P. Martimo, and E. Viikari-Juntura (2008): Partial sick leave - review of its use, effects and feasibility in the Nordic countries. *Scandinavian Journal of Work Environment and Health*, 34(4):239–249. Cited on page 118.

Kausto, J., E. Viikari-Juntura, L. J. Virta, R. gould, A. Koskinen, and S. Solovieva (2014): Effectiveness of new legislation on partial sickness benefit on work participation: A quasi-experiment in Finland. *BMJ Open*, 4:e006685. Cited on page 118.

Knoef, M., J. Been, R. Alessie, K. Caminada, K. Goudswaard, and A. Kalwij (2016a): Measuring retirement savings adequacy: developing a multi-pillar approach in the Netherlands. *Journal of Pension Economics and Finance*, 15(1):55–89. Cited on page 54.

Knoef, M., J. Rhuggenaath, J. Been, K. Caminada, and K. Goudswaard (2016b): De toereikendheid van pensioenopbouw na de crisis en pensionhervormingen. *Netspar Industry Series*, Design 68:1–98. Cited on pages xvii and 54.

Koning, P. (2017): Privatizing sick pay: Does it work? *IZA World of Labor*, 324:1–9. Cited on page 122.

Koning, P. and M. Lindeboom (2015): The rise and fall of Disability Insurance enrollment in the Netherlands. *Journal of Economic Perspectives*, 29:151–172. Cited on page 122.

Kools, L. and M. Knoef (2017): Health and the marginal utility of consumption: Estimating health state dependence using equivalence scales. Netspar Academic Series DP 04/2017-008. Cited on page 73.

Kools, L. and P. Koning (2018): Graded return-to-work as a stepping stone to full work resumption. IZA DP No. 11471. Cited on page 117.

Krause, N., L. K. Dasinger, and F. Neuhauser (1998): Modified work and return to work: A review of the literature. *Journal of Occupational Rehabilitation*, 8(2):113–139. Cited on page 118.

Lemeshow, S., A. G. Tserkovnyi, J. L. Tulloch, J. E. Dowd, S. K. Lwanga, and J. Keja (1985): A computer simulation of the EPI survey strategy. *International Journal of Epidemiology*, 14:473–481. Cited on page 180.

Lillard, L. A. and Y. Weiss (1997): Uncertain health and survival: effects on end-of-life consumption. *Journal of Business & Economic Statistics*, 15(2):254–268. Cited on pages 75 and 76.

Linard, C., M. Gilbert, R. W. Snow, A. M. Noor, and A. J. Tatem (2012): Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS ONE*, 7(2):e31743. Cited on page 209.

Lindeboom, M. and R. Montizaan (2018): Pension reform: disentangling the impact on retirement behavior and private savings. Netspar Discussion Paper 2018-012. Cited on pages 13 and 37.

Lloyd, C. T., A. Sorichetta, and A. J. Tatem (2017): Data descriptor: High resolution global gridded data for use in population studies. *Scientific Data*, 4:170001. Cited on pages .

Luman, E. T., A. Worku, Y. Berhane, and R. M. A. L. Cairns (2007): Comparison of two survey methodologies to assess vaccination coverage. *International Journal of Epidemiology*, 36:633–641. Cited on page 181.

Lumsdaine, R. and A. Exterkate (2013): How survey design affects self-assessed health responses in the Survey of Health, Ageing, and Retirement in Europe (SHARE). *European Economic Review*, 63:299–307. Cited on page 90.

Maestas, N., K. J. Mullen, and A. Strand (2013): Does Disability Insurance receipt discourage work? Using examinar assignment to estimate casual effects of SSDI receipt. *American Economic Review*, 103:1797–1829. Cited on page 136.

Markussen, S., A. Mykletun, and K. Røed (2012): The case for presenteeism - Evidence from Norway's sickness insurance program. *Journal of Public Economics*, 96:959–972. Cited on pages 118, 136, and 145.

Markussen, S. and K. Røed (2014): The impacts of vocation rehabilitation. *Labour Economics*, 31:1–13. Cited on pages 118, 119, 122, 136, 139, and 140.

Markussen, S., K. Røed, and R. C. Schreiner (2017): Can compulsory dialogues nudge sick-listed workers back to work? *The Economic Journal*, page doi:10.1111/ecoj.12468. Cited on page 136.

Mastrobuoni, G. (2009): Labor supply effects of the recent social security benefit cuts: Empirical estimates using cohort discontinuities. *Journal of Public Economics*, 93(11-12):1224–1233. Cited on page 15.

McCarthy, J. (1995): Imperfect insurance and differing propensities to consume across households. *Journal of Monetary Economics*, 36:301–327. Cited on page 37.

Mete, C. (2005): Predictors of elderly mortality: health status, socioeconomic characteristics and social determinants of health. *Health Economics*, 14:135–148. Cited on page 90.

Mian, A., K. Rao, and A. Sufi (2013): Household balance sheets, consumption, and the economic slump. *The Quarterly Journal of Economics*, 128(4):1687–1726. Cited on page 10.

Milligan, P., A. Njie, and S. Bennett (2004): Comparison of two cluster sampling methods for health surveys in developing countries. *International Journal of Epidemiology*, 33:469–476. Cited on page 180.

Mitchell, O. S. and J. F. Moore (1998): Can Americans afford to retire? New evidence on retirement saving adequacy. *The Journal of Risk and Insurance*, 65(3):371–400. Cited on page 13.

Modigliani, F. and A. Ando (1960): The permanent income and the life cycle hypothesis of saving behavior: comparisons and tests. In I. Friend and R. Jones, editors, *Proceedings of the Conference on Consumption and Savings*, pages 49–174. Philadelphia, PA: University of Pennsylvania. Cited on page 10.

Mundlak, Y. (1978): On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85. Cited on page 82.

[Namibia] National Statistics Agency (NSA) (2011a): 2011 Census EA boundaries. Digital Namibia [Online]. Available: https://digitalnamibia.nsa.org.na/. [Accessed: 19-Feb-2018]. Cited on page 183.

——— (2011b): Namibia Population and Housing Census 2011. Main report. Windhoek. Cited on page 183.

——— (2013): Nambia 2011 Population and Housing Census [PUMS dataset]. Version 1.0. Windhoek. Cited on page 183.

OECD (2010): *Sickness, disability and work: Breaking the barriers. A synthesis of findings across OECD countries*. OECD Publishing, Paris. Cited on page 118.

——— (2015): *Health at Glance 2015, OECD indicators*. OECD Publishing, Paris. Cited on page 94.

Östh, J., W. A. V. Clark, and B. Malmberg (2015): Measuring the scale of segregation using k-nearest neighbor aggregates. *Geographical analysis*, 47:34–49. Cited on page 206.

Paiella, M. (2009): The stock market, housing and consumer spending: a survey of the evidence on wealth effects. *Journal of Economic Surveys*, 23(9):947–973. Cited on page 10.

Paiella, M. and L. Pistaferri (2017): Decomposing the wealth effect on consumption. *The Review of Economics and Statistics*, 99(3):710–721. Cited on page 11.

Palumbo, M. G. (1999): Uncertain medical expenses and precautionary saving near the end of the life cycle. *Review of Economic Studies*, 66(2):395–421. Cited on page 92.

Parker, J. (1999): The reaction of household consumption to predictable changes in social security taxes. *American Economic Review*, 89(4):959–973. Cited on page 28.

Perez-Heydrich, C., J. L. Warren, C. R. Burgert, and M. E. Emch (2013): Guidelines on the use of DHS GPS data. Spatial analysis reports No. 8. Cited on page 186.

Petrović, A., M. van Ham, and D. Manley (2018): Multiscale measures of population: Within-and between-city variation in exposure to the sociospatial context. *Annals of the American Association of Geographers*, 0:1–18. Cited on page 206.

Poterba, J. M. (2000): Stock market wealth and consumption. *Journal of Economic Perspectives*, 14(2):99–118. Cited on page 11.

Pradhan, M. and M. Ravallion (2000): Measuring poverty using qualitative perceptions of consumption adequacy. *The Review of Economic and Statistics*, 82:462–471. Cited on pages 75, 81, 99, and 104.

Rehwald, K., M. Rosholm, and B. Roulande (2016): Does activating sick-listed workers work? Evidence from a randomized experiment. IZA Discussion Paper No. 9771. Cited on pages 118, 119, 122, 136, 139, 140, and 145.

Riedl, M. and I. Geishecker (2014): Keep it simple: estimation strategies for ordered response models with fixed effects. *Journal of Applied Statistics*, 41(11):2358–2374. Cited on page 99.

Roodman, D. (2011): Fitting fully observed recursive mixed-process models with CMP. *The Stata Journal*, 11(2):159–206. Cited on page 31.

Rose, A. M. C., R. F. Grais, D. Coulombier, and H. Ritter (2006): A comparison of cluster and systematic samping methods for measuring crude mortality. *Bulletin of the World Health Organization*, 84:290–296. Cited on page 181.

Running, S. W., R. R. Nemani, F. A. Heinsch, M. Zhao, M. Reeves, and H. Hashimoto (2014): A countinuous satellite-derived measure of global terrestrial primary production. *BioScience*, 54(6):547–560. Cited on page 183.

Scheil-Adlung, X. and J. Bonan (2012): Can the European elderly afford the financial burden of health and long-term care? Assessing impacts and policy implications. International Labour Office, Social Security Department. Geneva: ILO. Cited on page 94.

Scholnick, B. (2013): Consumption smoothing after the final mortgage payment: testing the magnitude hypothesis. *Review of Economics and Statistics*, 95(4):1444–1449. Cited on page 34.

Scholz, J. K., A. Seshadri, and S. Khitatrakun (2006): Are Americans saving optimally for retirement? *Journal of Political Economy*, 114(4):607–643. Cited on page 13.

Sinai, T. and N. S. Souleles (2005): Owner-occupied housing as a hedge against rent risk. *The Quarterly Journal of Economics*, 120(2):763–789. Cited on page 13.

Skinner, J. (1985): Variable lifespan and the intertemporal elasticity of consumption. *Review of Economics*, 67(4):616–623. Cited on page 92.

——— (2007): Are you sure you're saving enough for retirement? *Journal of Economic Perspectives*, 21(3):59–80. Cited on page 13.

Sloan, F. A., W. K. Viscusi, H. W. Chesson, C. J. Conover, and K. Whetten-Goldstein (1998): Alternative approaches to valuing intangible health losses, the evidence for multiple sclerosis. *Journal of Health Economics*, 17(4):475–497. Cited on page 77.

van Sonsbeek, J.-M. and R. H. J. M. Gradus (2013): Estimating the effects of recent disability reforms in the Netherlands. *Oxford Economic Papers*, 65:832–855. Cited on page 122.

Staubli, S. and J. Zweimüller (2013): Does raising the early retirement age increase employment of older workers? *Journal of Public Economics*, 108:17–32. Cited on page 15.

Templ, M., B. Meindl, A. Kowarik, and O. Dupriez (2017): Simulation of Synthetic Complex Data: The R Package simPop. *Journal of Statistical Software*, 79:10. Cited on page 187.

Thaler, R. H. (1990): Anomalies: Saving, fungibility, and mental accounts. *Journal of Economic Perspectives*, 4(1):193–205. Cited on page 10.

The Namibia Ministry of Health and Social Services (MoHSS) and ICF International (MoHSS and ICF) (2014): Namibia Demographic and Health Survey 2013 [Dataset]. Windhoek, namibia, and Rockvill, Maryland, USA: MoHSS and ICF International. Cited on page 183.

Thomson, D. R., L. Kools, and W. C. Jochem (2018): Linking synthetic populations to household geolocations: A demonstration in Namibia. *Data*, 3(3):30. Cited on pages 182, 185, and 188.

Thomson, D. R., F. R. Stevens, N. W. Ruktanonchai, A. J. Tatem, and M. C. Castro (2017): *GridSample*: An R package to generate household survey primary sampling units (PSUs) from gridded population data. *International Journal of Health Geographies*, 16:25. Cited on page 179.

Turner, A. G., R. J. Magnani, and M. Shuaib (1996): A not quite as quick but much cleaner alternative to the Expanded Programme on Immunization (EPI) cluster survey design. *International Journal of Epidemiology*, 25:198–203. Cited on page 180.

United Nations, Department of Economic and Social Affairs, Population Division (UN) (2017): World Population Prospects. The 2017 Revision, custom data acquired via website. Cited on page 195.

United Nations Office for the Coordination of Humanitarian Affairs (OCHA) Regional Office for Southern Africa

(ROSA) (UN-OCHA ROSA) (2001a): Namibia - Education Facilities. Humanitarian Data Exchange [Online]. Available: https://data.humdata.org/organization/ocha-rosa. [Accessed: 19-Feb-2017]. Cited on page 183.

——— (2001b): Namibia - Health Facilities. Humanitarian Data Exchange [Online]. Available: https://data.humdata.org/organization/ocha-rosa. [Accessed: 19-Feb-2017]. Cited on page 183.

Van der Laan, J. (2009): Representativity of the LISS panel. Statistics Netherlands discussion paper 09041. Cited on page 20.

Verbeek-Oudijk, D., I. Woittiez, E. Eggink, and L. Putman (2014): Who cares in Europe? A comparison of long-term care for the over-50s in sixteen European countries. SCP publication 2014-9. Cited on page 95.

Viikari-Juntura, E., J. Kausto, R. Shiri, L. Kaila-Kangas, E. Pekka Takala, J. Karppinen, H. Miranda, R. Luukkonen, and K. Pekka Martimo (2012): Return to work after early part-time sick leave due to musculoskeletal disorders: A randomized controlled trial. *Scandinavian Journal of Work Environment and Health*, 38(2):134–143. Cited on pages 118 and 119.

Viscusi, W. K. and W. N. Evans (1990): Utility functions that depend on health status: Estimates and economic implications. *American Economic Review*, 80(3):353–374. Cited on pages 74, 77, and 78.

Wainwright, E., D. Wainwright, E. Keogh, and C. Eccleston (2011): Fit for purpose? Using the fit note with patients with chronic pain: A qualitative study. *British Journal of General Practice*, 61(593):794–800. Cited on pages 118 and 137.

World Health Organization (2015): Vaccination Coverage Cluster Surveys: Reference Manual. Version 3 - WORKING DRAFT Update July 2015. Cited on pages 178 and 180.

Nederlandse samenvatting

Onderzoek naar vermogen, gezondheid en dataverzameling.

Sociale verzekeringen hebben als doel om individuen een financieel vangnet te bieden wanneer ze met tegenslag geconfronteerd worden. Echter, bij het vormgeven van een systeem van sociale verzekeringen wordt vaak niet alleen gekeken naar de bescherming die het systeem biedt, maar ook hoe met behulp van het systeem het welvaartsniveau kan worden verhoogd. Hiervoor is kennis nodig over zowel de individuele reacties op tegenslag, zoals ziekte of een vermogensschok, als de individuele reacties op het systeem dat men moet beschermen tegen die tegenslag. Als iemand bijvoorbeeld de hoogte van bijdragen aan sociale verzekeringen wil bepalen, zal diegene eerst moeten begrijpen hoe individuen bij voorkeur hun financiële middelen verdelen over verschillende mogelijke levensuitkomsten en hoe consumptiepatronen beïnvloed worden door negatieve schokken zoals ziekte. Daarnaast zal diegene moeten begrijpen welke negatieve effecten sociale verzekeringen met zich mee kunnen brengen in de vorm van *moral hazard* en of deze teniet gedaan kunnen worden met behulp van complementaire interventies. Om zulke gedragseffecten te kunnen onderzoeken is toegang tot data op individueel niveau, die van hoge kwaliteit is en representatief voor de onderzoekspopulatie, essentieel. Verbeteringen in dataverzamelingsmethoden is dus van groot belang voor een goed begrip van de werking van sociale verzekeringen.

Dit proefschrift bevat vier papers die betrekking hebben op de hierboven beschreven problematiek. In het eerste paper geven we antwoord op de vraag: *"Wat is het effect van de dalingen in Nederlandse pensioen- en woningvermogens gedurende de periode 2008-2014 op de minimale uitgaven na pensionering?"* In de laatste jaren is het pensioensysteem in Nederland onderhevig geweest aan verschillende veranderingen. Naar aanleiding van de toenemende grijze druk is de fiscaal gefaciliteerde pensioenopbouw versoberd en de AOW-leeftijd verhoogd. Daarnaast kwamen pensioenfondsen in de problemen. De levensverwachting, en daarmee de verplichtingen, nam sneller toe dan verwacht en de rente was gedaald. Ook waren de mogelijkheden om tegenvallende beleggingsresultaten te compenseren met premieverhogingen beperkt door de toenemende grijze druk. Tijdens de financiële crisis zagen pensioenfondsen geen andere mogelijkheid dan indexaties achterwege laten en in sommige gevallen moesten de pensioenuitkeringen zelfs in nominale termen gekort worden. Tegelijkertijd was er een scherpe daling in de huizenprijzen. Deze ontwikkelingen leidden tot ongerustheid: hebben huishoudens nog wel voldoende middelen om hun oude dag te financieren?

Een pensioen wordt toereikend bevonden als het de individu in staat stelt zijn levensstandaard van voor pensionering voort te zetten. Dit wordt doorgaans op een nogal pragmatische manier beoordeeld: het bruto inkomen na pensionering wordt voldoende bevonden als het tenminste gelijk is aan 70% van het gemiddelde bruto inkomen voor pensionering. Het idee achter deze 70%-maatstaf is dat individuen na pensionering niet langer hoeven te sparen, geen werk gerelateerde uitgaven meer hebben en meer tijd hebben om zelf klusjes in en om het huis te doen om zo hun uitgaven te verlagen. Als door de veranderingen in het pensioenstelsel het pensioeninkomen van iemand onder deze 70% duikt zouden we ons volgens deze maatstaf dus zorgen moeten gaan maken. Echter, optimale vervangingsratio's kunnen door de tijd heen veranderen. Het optimale levenscyclusmodel voorspelt dat individuen die tegen een onverwachte daling van het vermogen aanlopen, zowel vandaag als in de toekomst hun consumptie verminderen. Individuen creëren zo als het ware hun eigen

vangnet door schokken in toekomstig pensioeninkomen te spreiden over een langere periode, met lagere optimale vervangingsratio's als gevolg.

Om een inschatting te maken van de omvang van deze individuele gedragsreacties maken wij gebruik van een enquête waarin mensen zowel voor als na de crisis gevraagd zijn naar hun minimale uitgaven na pensionering. We koppelen deze informatie aan administratieve gegevens over (pensioen)vermogen. Aan de hand van deze data schatten we wat het effect is van een schok in het pensioenvermogen op de minimale uitgaven na pensionering. Dankzij het verplichte karakter van de Nederlandse pensioenen kunnen we het effect van de individuele vermogensschok onderscheiden van meer algemene effecten zoals pessimisme. Daarnaast kijken we aan de hand van simulaties hoe de toereikendheid van pensioenen is veranderd tussen 2008 en 2014, en wat de rol van veranderingen in minimale uitgaven hierbij is geweest.

De resultaten laten zien dat een daling van 100 euro in pensioenannuiteiten zich vertaalt in een daling van 23-33 euro in minimale uitgaven na pensionering. Echter hebben meer algemene veranderingen in sentiment ook tot dalingen in gewenste uitgaven geleid. Jongeren reageren vooral op dalingen in het woonvermogen, waar ouderen sterker reageren op dalingen in het pensioenvermogen. Daarnaast lijken individuen met hogere inkomens hun toekomstige uitgaven meer aan te passen dan individuen met lagere inkomens. De simulaties laten zien dat het percentage mensen met te weinig pensioen, gedefinieerd als een netto pensioenannuiteit die kleiner is dan de gewenste uitgaven, licht gestegen is tijdens de crisisjaren. Echter, als mensen hun gewenste uitgaven niet hadden aangepast, was dit percentage bijna verdubbeld.

In het tweede paper keren we terug naar het optimale levenscyclusmodel, maar verschuiven we onze aandacht naar de rol van gezondheid in het vormen van consumptievoorkeuren. Het levenscyclusmodel is een nuttig middel om welvaartseffecten van bijvoorbeeld ziektekostenverzekeringen of het pensioenstelsel te evalueren. Volgens dit model is het totale nut tijdens iemands leven het hoogst als het verwachte marginale nut constant blijft over de levenscyclus, waarbij rekening gehouden wordt

met factoren zoals ongeduld en risico-aversie. Het verwachte marginale nut hangt af van de kans op gebeurtenissen zoals het verliezen van een baan of het vormen van een gezin, welke óf het toekomstig inkomen beïnvloeden (in het geval van baanverlies) óf het nut dat ontleend wordt aan een extra euro consumptie (in het geval van gezinsformatie). In beide gevallen wordt het optimale niveau van bijdragen en uitkeringen beïnvloed. Ook gezondheid zou een rol in dit model kunnen spelen en niet alleen omdat een verslechtering van de gezondheid iemands potentiële verdiencapaciteit kan beïnvloeden. Hoe gezond iemand is kan namelijk ook invloed hebben op het marginale nut van consumptie, bijvoorbeeld omdat iemand minder plezier ontleent aan een avontuurlijke vakantie in tijden van slechte gezondheid, maar bijvoorbeeld meer nut ontleent aan de uitgaven aan een schoonmaker. Het empirisch onderzoek gericht op het meten van het effect van gezondheid op het marginale nut van consumptie geeft gemengde resultaten en is voornamelijk gebaseerd op data uit de Verenigde Staten. Daarom proberen we in dit paper de volgende vraag te beantwoorden: *"Wat is het effect van gezondheid op het marginale nut van consumptie in Europa?"*

Om deze vraag te beantwoorden ontwikkelen we een methodologisch kader waarbinnen een relatie wordt gelegd tussen subjectieve uitspraken over inkomenstevredenheid en het levenscyclusmodel. Het voordeel van dit methodologisch kader ten opzichte van andere methodes is dat er gebruik gemaakt wordt van een vraag die in veel verschillende representatieve nationale enquêtes gesteld wordt en dat met deze methode ook precieze resultaten geschat kunnen worden als de panel data maar relatief weinig jaargangen bevat. Dit is in het bijzonder relevant binnen de Europese context, waar geharmoniseerde panel data pas relatief recent zijn geïntroduceerd. Wij passen de methode toe op data van SHARE, een enquête gericht op 50+'ers in verschillende Europese landen.

De resultaten laten zien dat een verslechtering van de gezondheid leidt tot een stijging van het marginale nut van consumptie voor de gemiddelde Europeaan. Dit betekent dat de welvaart verhoogd kan worden door inkomen te verschuiven van periodes van goede gezondheid, naar periodes van slechte gezondheid (dat wil zeggen, hogere premies en uitke-

ringen). Echter, een verslechtering van de cognitieve gezondheid leidt tot een daling van het marginale nut van consumptie, waarschijnlijk omdat het moeilijker wordt om te plannen en initiatief te tonen.

Ook in het derde paper kijken we naar economisch gedrag na ziekte. Tot 2004 waren alle werknemers in Nederland verzekerd tegen arbeidsongeschiktheid onder de Wet op Arbeidsongeschiktheidsverzekering (WAO). Via deze verzekering had men recht op een uitkering ter hoogte van 70% van hun inkomen wanneer zij arbeidsongeschikt werden. Dat dit type verzekering ook negatieve effecten met zich mee kan brengen, bleek uit de hoge instroom in dit programma tijdens de jaren tachtig en negentig van de vorige eeuw, ook wel bekend als *'the Dutch disease'*. Het systeem bleek een aantrekkelijk alternatief te bieden voor regulier ontslag, zodat veel van de uitkeringsgerechtigden eigenlijk niet langdurig arbeidsongeschikt waren. Om het stijgende verzuim terug te dringen, is onder meer de Wet verbetering Poortwachter aangenomen, met een sterkere controle voor instroom en grotere verantwoordelijkheden voor de werkgever. Mocht een werknemer ziek worden, dan is de werkgever nu verplicht het loon twee jaar door te betalen. Tegelijkertijd moeten zowel werkgever als werknemer zich actief inzetten voor de re-integratie van de zieke werknemer. Pas als de werknemer na twee jaar nog steeds niet aan het werk kan, komt de verzekering tegen inkomensverlies door langdurige arbeidsongeschiktheid in beeld.

Een voorbeeld van hoe werkgevers en werknemers actief re-integratie kunnen bevorderen, is het werken onder aangepaste omstandigheden tijdens het ziekteverlof. Zo zou iemand eerst slechts een aantal uren kunnen werken en dat elke week een klein beetje op kunnen bouwen, *graded return-to-work*. Deelnemen aan (aangepast) werk tijdens het ziekteverlof zou kunnen helpen het verlies van menselijk kapitaal tegen te gaan en in sommige gevallen kan het zelfs helpen bij een sneller herstel van fysieke klachten. Echter, er bestaat het risico dat wanneer men te snel opbouwt het lichaam overbelast wordt zodat het herstelproces juist langzamer zal verlopen. De huidige academische literatuur laat zien dat deeltijd en/of aangepast werk tijdens het ziekteverlof een effectieve manier is om de

duur van afwezigheid door ziekte te verkorten en de kans op permanente arbeidsongeschiktheid te verminderen. Echter, er is nog maar weinig bekend over hoe zulke trajecten het best opgezet kunnen worden. In het vierde paper beantwoorden we daarom de vraag *"Hangt de effectiviteit van deeltijd werkhervatting tijdens het ziekteverlof af van (1) het moment dat het traject is gestart; (2) het aantal uren dat iemand werkt bij de start; (3) het ziektebeeld?"*

We beantwoorden deze vraag op basis van het cliëntenbestand van een private onderneming die casemanagement verzorgt bij ziektegevallen. Deze partij helpt bij het uitvoeren van de verplichtingen van de Wet Verbetering Poortwachter en bij het opstellen van een plan van aanpak voor re-integratie. Of een zieke werknemer aan deeltijd werkhervatting deelneemt hangt samen met de verwachte herstelkans van de werknemer, zodat een simpel regressiemodel onjuiste schattingen zal geven. Dit lossen we op door een instrumentele variabele te genereren die weergeeft welke voorkeuren de casemanager van de cliënt heeft met betrekking tot het starten van een deeltijd werkhervattingstraject. De ene casemanager zal geneigd zijn dit type traject vaker, vroeger of met een grotere deeltijdfactor in te starten dan een andere casemanager, wat invloed kan hebben op het traject dat de individuele cliënt zal ondergaan.

De resultaten laten zien dat deeltijd werkhervatting tijdens het ziekteverlof nog effectiever is als het snel en intensief gestart wordt. Waarschijnlijk biedt dit type start iemand meer kans om als volwaardige werknemer deel te nemen aan werkprocessen. Dit geldt echter niet voor werknemers die last hebben psychologische of psychiatrische problemen. In die gevallen kan er beter wat langer gewacht worden tot het traject gestart wordt. In tegenstelling tot eerdere literatuur, laten de resultaten zien dat ondanks dat de deeltijd werkhervatting wel leidt tot kortere ziekteduren, het geen invloed heeft op de kans dat iemand langdurig arbeidsongeschikt raakt. Dit verschil kan verklaard worden door de omstandigheden waaronder de individuen in de 'controle groep' in Nederland verkeren. Ondanks dat zij niet deelnemen aan het traject van deeltijd werkhervatting tijdens het ziekteverlof, blijven zij via de Wet verbetering Poortwachter wel in contact met hun werkgever en moeten ze op andere manieren aan hun

terugkeer werken. Als er ook maar een kleine kans is dat iemand herstelt, zal dit waarschijnlijk ook bereikt worden zonder de inzet van deeltijd werkhervatting.

Het laatste paper heeft betrekking op dataverzamelingsmethodes. Om de sociaal-economische bescherming van huishoudens te vergelijken tussen landen en over de tijd heen, is vergelijkbare en nauwkeurige informatie nodig over zaken als armoede, inkomensongelijkheid en ziekte. Bij voorkeur wordt deze informatie regelmatig geactualiseerd. Binnen Nederland kunnen we gebruik maken van een grote hoeveelheid administratieve data en jaarlijkse grootschalige enquêtes, maar het is een stuk lastiger om deze informatie te vergaren in veel landen met een laag- of middeninkomen. Daar is men afhankelijk van enquêtes die eens in de zoveel jaar bij huishoudens thuis worden afgenomen. Het uitzetten van deze enquêtes kost veel tijd en geld. Onderzoek naar efficiëntere en makkelijkere manieren om enquêtes uit te zetten zijn dus belangrijk om (internationaal) onderzoek naar sociale zekerheid te bevorderen.

De meest gangbare methode om huishoudens te selecteren voor dit type enquêtes is *two-stage cluster sampling*. Deze methode houdt in dat men eerst op willekeurige basis een aantal kleine regio's selecteert, om vervolgens een willekeurige selectie van huishoudens binnen deze regio's te maken. De tweede stap is noodzakelijk, omdat de regio's vaak te groot zijn om in een dag alle huishoudens te kunnen interviewen. Met behulp van deze methode kan het veldwerk geconcentreerd worden in slechts enkele regio's, maar deze regio's moeten wel in ieder geval twee keer bezocht worden. Dit herhaaldelijk bezoeken leidt tot hoge kosten en het risico dat mobiele huishoudens, bijvoorbeeld huishoudens met seizoenswerkers, uitgesloten worden van de enquête. Nieuwe methodes om huishoudens te selecteren zoals *gridded sampling* geven de mogelijkheid om kleinere regio's te definiëren, zodat het mogelijk is om alle huishoudens binnen die regio te enquêteren. Deze methode, ook wel *one-stage cluster sampling* genoemd, zou tot substantieel lagere kosten kunnen leiden omdat de identificatie- en interviewfase gecombineerd kunnen worden op een dag en het werkgebied kleiner is. Daarnaast is de kans dat mobiele

huishoudens meegenomen worden groter. Echter, als huishoudens die erg op elkaar lijken ook de neiging hebben dicht bij elkaar in de buurt wonen, zal het met deze methode nodig zijn om meer huishoudens te interviewen, wat weer leidt tot hogere kosten.

Daarom gaan wij in dit paper op zoek naar het antwoord op de vraag: *"Hoeveel extra clusters moeten er getrokken worden onder one-stage cluster sampling om steekproefschattingen te krijgen met de precisie van een two-stage cluster sample?"* Dit doen we door eerst een synthetische geo-gecodeerde microdataset te genereren die alle huishoudens in Oshikoto (Namibië) bevat. Hiervoor maken we gebruik van informatie uit recente enquêtes, een census en ruimtelijke covariaten. Deze informatie combineren we met behulp van verschillende voorspel- en clusteringsmethodes. De resulterende data hebben dezelfde statistische eigenschappen als de echte populatie. Echter, een vergelijkbare dataset van de echte populatie zou niet publiek beschikbaar kunnen worden gemaakt vanwege de privacygevoeligheid van de informatie. Op basis van gesimuleerde uitkomsten van de twee *sampling* methodes toegepast op de synthetische populatie, stellen wij het aantal clusters vast dat nodig is om een accurate schatting van populatiegemiddeldes te verkrijgen. Hierbij nemen we verschillende scenario's van clustering van huishoudenstypes aan.

De resultaten laten zien dat een *one-stage cluster sample* niet perse tot grotere steekproeven hoeft te leiden, tenzij er perfecte socio-economische segregatie is op basis van een van de karakteristieken die de survey meet. In zo'n extreme situatie kan de ideale steekproefgrootte bijna dertien keer groter zijn dan in het geval van een *two-stage cluster sample*. Echter, in bijna alle andere situaties, stijgt de ideale steekproefgrootte met hoogstens 30%, zodat *one-stage cluster sampling* in de praktijk een haalbaar alternatief kan zijn voor *two-stage cluster sampling*.

Curriculum Vitae

Lieke Kools was born in Boarnsterhim, the Netherlands on June 7, 1990. In 2012 she obtained a BSc degree in Econometrics, Operations Research and Actuarial Studies from the University of Groningen and in 2014 she received a MSc degree in Econometrics and Operations Research from the same institution. During her studies she taught several bachelor level courses on mathematics and statistics. Moreover, she spent a semester at Ewha Womans University[19] in Seoul, South Korea. Her MSc thesis, written during an internship at TNO (Netherlands Organisation for Applied Scientific Research), got awarded best MSc thesis of the Faculty of Economics and Business 2014 and was published in *Energy*.

In 2014 she became a PhD candidate at the department of Economics at Leiden University and Junior Research Fellow at Netspar. During her PhD she participated in several international PhD workshops organized by i.a. *CESifo*, *LISER*, and *IIPF* and presented her work at various international conferences such as the annual meetings of *ESPE*, *EEA-ESEM*, and *EALE*. Furthermore, she presented policy relevant work at *CPB* (Netherlands Bureau for Economic Policy Analysis), the Dutch Ministry of Social Affairs and at several Netspar events. In 2018, she was a visiting PhD at WorldPop, University of Southampton.

Besides her research, Lieke organized seminars for the PhD candidates of the Leiden Law School. Moreover, as a member of SMO Promovendi, a do-thank aimed at promoting valorisation under young researchers, she developed workshops on valorisation and entrepreneurship and edited a popular-scientific book on the concept of self-reliance in health care.

[19] Womans is a neologism to emphasize each individual woman who receives education.

In de boekenreeks van het E.M. Meijers Instituut van de Faculteit der Rechtsgeleerdheid, Universiteit Leiden, zijn in 2017 en 2018 verschenen:

MI-274 E.J.M. Vergeer, *Regeldruk vanuit een ander perspectief. Onderzoek naar de beleving van deregulering bij ondernemers*, (diss. Leiden)

MI-275 J.J. Oerlemans, *Investigating Cybercrime*, (diss. Leiden), Amsterdam: Amsterdam University Press 2017, ISBN 978 90 8555 109 6

MI-276 E.A.C. Raaijmakers, *The Subjectively Experienced Severity of Imprisonment: Determinants and Consequences*, (diss. Leiden), Amsterdam: Ipskamp Printing, 2016, ISBN 978 94 0280 455 3

MI-277 M.R. Bruning, T. Liefaard, M.M.C. Limbeek, B.T.M. Bahlmann, *Verplichte (na)zorg voor kwetsbare jongvolwassenen?*, Nijmegen: Wolf Legal Publishers 2016, ISBN 978 94 624 0351 2

MI-278 A.Q. Bosma, *Targeting recidivism. An evaluation study into the functioning and effectiveness of a prison-based treatment program*, (diss. Leiden), Zutphen: Wöhrmann 2016

MI-279 B.J.G. Leeuw, F.P. Ölçer & J.M. Ten Voorde (red.), *Leidse gedachten voor een modern straf(procesrecht*, Den Haag: Boom Juridisch 2017, ISBN 978 94 6290 392 0

MI-280 J. Tegelaar, *Exit Peter Paul? Divergente toezichthoudersaansprakelijkheid in de Europese Unie voor falend financieel toezicht, bezien vanuit het Europeesrechtelijke beginsel van effectieve rechtsbescherming*, (Jongbloed scriptieprijs 2016), Den Haag: Jongbloed 2017, ISBN 978 90 8959 129 6

MI-281 P. van Berlo et al. (red.), *Over de grenzen van de discipline. Interactions between and with-in criminal law and criminology*, Den Haag: Boom Juridisch 2017, ISBN 978 94 6290 390 6

MI-282 J. Mačkić, *Proving Discriminatory Violence at the European Court of Human Rights*, (diss. Leiden), Amsterdam: Ipskamp Printing 2017

MI-283 D.V. Dimov, *Crowdsourced Online Dispute Resolution*, (diss. Leiden), Amsterdam: Ipskamp Printing 2017, ISBN 978 94 0280 578 9

MI-284 T. de Jong, *Procedurele waarborgen in materiële EVRM-rechten*, (diss. Leiden), Deventer: Kluwer 2017, ISBN 978 90 1314 413 0

MI-285 A. Tonutti, *The Role of Modern International Commissions of Inquiry. A First Step to Ensure Accountability for International Law Violations?*, (diss. Leiden), Amsterdam: Ipskamp Printing 2017

MI-286 W. de Heer, *Gelijkheid troef in het Nederlandse basisonderwijs*, (diss. Leiden), Amsterdam: Ipskamp Printing 2017, ISBN 978 94 0280 697 7

MI-287 J. Wieland, *De bescherming van concurrentiebelangen in het bestuursrecht*, (diss. Leiden), Den Haag: Boom Juridisch 2017, ISBN 978 94 6290 427 9, e-ISBN 978 94 6274 772 2

MI-288 D.M. Broekhuijsen, *A Multilateral Tax Treaty. Designing an instrument to modernize international tax law*, (diss. Leiden), Amsterdam: Ipskamp Printing 2017

MI-289 L. Kovudhikulrungsri, *The right to travel by air for persons with disabilities* (diss. Leiden), Amsterdam: Ipskamp Printing 2017

MI-290 R. Hage, *Handhaving van privaatrecht door toezichthouders*, (diss. Leiden), Deventer: Kluwer 2017

MI-291 M. Diamant, *Het budgetrecht van het Nederlandse parlement in het licht van het Europees economisch bestuur*, (diss. Leiden), Deventer: Kluwer 2017, ISBN 978 90 1314 555 7

MI-292 R. Passchier, *Informal constitutional change: Constitutional change without formal constitutional amendment in comparative perspective*, (diss. Leiden), Amsterdam: Ipskamp Printing 2017

MI-293 T. Leclerc, *Les mesures correctives aux émissions aériennes de gaz à effet de serre. Contribution à l'étude des interactions entre les ordres juridiques en droit international public*, Amsterdam: Ipskamp Printing 2017

MI-294 M. Fink, *Frontex and Human Rights. Responsibility in 'Multi-Actor Situations' under the ECHR and EU Public Liability Law*, (diss. Leiden), Amsterdam: Ipskamp Printing 2017

MI-295 B.A. Kuiper-Slendebroek, *Rechter over Grenzen. De toepassing en interpretatie van internationaal recht in het Nederlands privaatrecht*, (diss. Leiden), Ipskamp Printing 2017

MI-296 Y.N. van den Brink, *Voorlopige hechtenis in het Nederlandse jeugdstrafrecht. Wet en praktijk in het licht van internationale en Europese kinder- en mensenrechten*, (diss. Leiden), Deventer: Kluwer 2017, ISBN 978 90 1314 683 7, e-ISBN 978 90 1314 684 4

MI-297 M.L. Diekhuis-Kuiper, *Het woord en de daad. Kenmerken van dreigbrieven en de intenties waarmee ze geschreven werden*, (diss. Leiden), Den Haag: Boom Criminologie 2017, ISBN 978 94 6236 795 1

MI-298 Y.N. van den Brink et al., *Voorlopige hechtenis van jeugdigen in uitvoering. Een exploratief kwantitatief onderzoek naar rechterlijke beslissingen en populatiekenmerken*, Nijmegen: Wolf Legal Publishers 2017, ISBN 978 94 6240 455 7

MI-299 V. Borger, *The Transformation of the Euro: Law, Contract, Solidarity*, (diss. Leiden), Amsterdam: Ipskamp Printing 2017

MI-300 N.N. Koster, *Crime victims and the police: Crime victims' evaluations of police behaviour, legitimacy, and cooperation: A multi-method study*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018

MI-301 Jingshu Zhu, *Straightjacket: Same-Sex Orientation under Chinese Family Law – Marriage, Parenthood, Eldercare*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018

MI-302 Xiang Li, *Collective Labour Rights and Collective Labour Relations of China*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018, ISBN 978 94 0280 924 4

MI-303 F. de Paula, *Legislative Policy in Brazil: Limits and Possibilities*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018, ISBN 978 94 028 0957 2

MI-304 C. Achmad, *Children's Rights in International Commercial Surrogacy. Exploring the challenges from a child rights, public international human rights law perspective*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018

MI-305 E.B. Beenakker, *The implementation of international law in the national legal order – A legislative perspective*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018

MI-306 Linlin Sun, *International Environmental Obligations and Liabilities in Deep Seabed Mining*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018

MI-307 Qiulin Hu, *Perspectives on the Regulation of Working Conditions in Times of Globalization – Challenges & Obstacles Facing Regulatory Intervention*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018

MI-308 L.M. de Hoog, *De prioriteitsregel in het vermogensrecht*, (diss. Leiden), Vianen: Proefschriftmaken.nl 2018

MI-309 E.S. Daalder, *De rechtspraakverzamelingen van Julius Paulus. Recht en rechtvaardigheid in de rechterlijke uitspraken van keizer Septimius Severus*, (diss. Leiden), Den Haag: Boom Juridisch 2018, ISBN 978 94 6290 556 6, ISBN 978 94 6274 946 7 (e-book)

MI-310 T.H. Sikkema, *Beginsel en begrip van verdeling*, (diss. Leiden), Vianen: Proefschriftmaken.nl 2018

MI-311 L. Kools, *Essays on wealth, health and data collection*, (diss. Leiden), Amsterdam: Ipskamp Printing 2018, ISBN 978 94 028 1168 1

E.M. MEIJERS INSTITUUT
INSTITUUT VOOR RECHTSWETENSCHAPPELIJK ONDERZOEK

One of the aims of social insurance programs is to provide a financial safety net to households when encountering adverse circumstances. However, apart from offering mere protection, a system of social insurance can also be designed with the aim to increase overall welfare. In order to make the appropriate design decisions one needs to understand how individuals react to both negative shocks, such as health and wealth shocks, and the system put in place to protect them from these shocks. For example, in order to determine appropriate levels of contributions and benefits in social insurance contracts, one needs to understand how individuals prefer to move resources between different potential life outcomes and how consumption patterns are affected by negative shocks such as illness. Moreover, one needs to understand which (negative) behavior can be provoked by income protection and how such moral hazard can be counteracted by complementary efforts to income support. To gain understanding on such behavioral effects, access to high quality microdata is crucial.

This thesis contains four essays aiming to generate empirical insights to facilitate the design of social insurance program. It focusses on changes in consumption preferences in reaction to health and wealth shocks, the effectivity of complementary efforts to sickness benefits programs, and methods to construct the surveys needed to gain these empirical insights.

This is a volume in the series of the Meijers Research Institute and Graduate School of the Leiden Law School of Leiden University. This study is part of the Law School's research program on 'Reform of Social Legislation'.

ISBN 978-94-0281-168-1