The Economic Costs of Life-Course Transitions

Inaugural-Dissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaft am Fachbereich Wirtschaftswissenschaft der Freien Universität Berlin

vorgelegt von Julian Johannes Schmied (M.Sc.)

aus Würzburg

Berlin
Dezember 2022

Dekan

Professor Dr. Dr. **Giacomo Corneo**
Freie Universität Berlin

Erstgutacher

Professor Dr. **Timm Bönke**
Freie Universität Berlin

Zweitgutacher

Professor Dr. **Carsten Schröder**
Freie Universität Berlin

Datum der Disputation: 05.05.2023

Erklärung über Zusammenarbeit mit Ko–AutorInnen und Veröffentlichungen

Kapitel 1

- in Zusammenarbeit mit Christian Dudel und Jan Marvin Garbuszus
- In 2020 erschienen als „Assessing differences in household needs: A comparison of approaches for the estimation of equivalence scales using German expenditure data“ in *Empirical Economics*; 60, pp. 1629–1659 DOI: 10.1007/s00181-020-01822-6

Kapitel 2

- in Zusammenarbeit mit Christian Dudel
- Veröffentlicht als „Pension adequacy standards: An empirical estimation strategy and results for the US and Germany“ MPIDR Working Paper 2019/3 DOI: 10.4054/MPIDR-WP-2019-003;
- Revise and Resubmit bei *Fiscal Studies*

Kapitel 3

- Das Kapitel wurde selbstständig verfasst
- Revise and Resubmit bei *The Journal of the Economics of Aging*

Kapitel 4

- in Zusammenarbeit mit Peter Eibich, Ricky Kanabar und Alexander Plum
- Veröffentlicht als „In and out of unemployment – labour market dynamics and the role of testosterone“ AUT Economics Working Paper Series 2020/13 und MPIDR Working Paper 2020/33. DOI: 10.4054/MPIDR-WP-2020-03
- In 2022 erschienen als „In and out of unemployment – Labour market transitions and the role of testosterone “ in *Economics and Human Biology* 46 (101103), p. 1-11, DOI: 10.1016/j.ehb.2022.101123

Weitere Fachartikel in Ko-Autorenschaft

mit Dudel, C., Werding, M. (2021): Lebensstandardsicherung im Alter: Wie hoch muss die Rente sein? DAV-Kompass 2021

mit Dudel C., Werding M. (2020). Sicherungsziele der Rente: empirische Messung und Ergebnisse. Wirtschaftsdienst 100(3), 185-193

mit Mazur A., Kanabar R., Plum A (2022): Testosterone correlates change as men age, *The Aging Male*, 25(1), 29-40

mit Freytag A. (2019): Debt relief and good governance: New evidence of developing countries cor the period 1990-2013, *Applied Econometrics and International Development* 19 (1), 15-32

mit A. Marr: (2017): Financial inclusion and poverty: The case of Peru, *Regional and Structural Economics* 16(2), 29-40.

Acknowledgments

I prepared this dissertation while I was a doctoral student at the Max Planck Institute for Demographic Research (MPIDR) in Rostock, Germany. The first year of the PhD was funded by the Deutsche Forschungsgemeinschaft (DFG) for the project "Einkommensabhängigkeit von Äquivalenzskalen und Nichtlinearität von Engel-Kurven," headed by Christian Dudel. Since October 2016, I have been part of the Research Group Labor Demography, headed by Peter Eibich and Mikko Myrskylä. At the same time, I have been enrolled as a PhD candidate at the Free University Berlin.

First of all, I want to thank Christian Dudel, who not only initiated the PhD project and brought me to the MPIDR, but has also been a mentor throughout the PhD. I am also deeply grateful to my first supervisor Timm Bönke who guided me well through the PhD, and was a crucial advisor for my research. I also benefited enormously from the advice and the experience of Carsten Schröder and Martin Werding.

I also want to thank a number of people who inspired me with their teaching and/or gave me encouraging feedback for my research and/or motivated me to pursue a PhD. They include Adam Scott and Hamid Mohtadi from the University of Wisconsin, Milwaukee; Alexander Kritikos and Malcolm Dunn from the University of Potsdam; and Oliver Farhauer from the University of Passau.

Next, my gratitude goes to my fellow researchers from the Labor Demography group, including Christian Dudel, Peter Eibich, Jo Hale, Chia Liu, Angelo Lorenti, Lydia Palumbo, Mikko Myrskylä, Hector Pifaree i Arolas, and Joshua Wilde. Thank you all for the fruitful seminars and the peer review groups where I received so much good feedback on my papers. Also, thanks to my office mate Robert Stelter who also gave a lot of good advices.

Thanks also to my co-authors for their great ideas and their dedication. They include Ana Marr, Andreas Freytag, Alexander Plum, Ricky Kanabar, Marvin Garbuszus, Christian Dudel, Peter Eibich, and Martin Werding.

In addition, I am grateful to the MPIDR PhD group. Thank you all for the great common lunches. It is so important to share thoughts and concerns with people facing the same challenges. Special thanks go to Andreas Höhn, Angela Carollo, and Marie Böhnstedt, with whom I shared a lot of time and thoughts. I would also like to thank Rosie Seaman and Kieron Barclay, the poor native speakers I so often asked for help with editing.

Most importantly, however, I want to thank my family, including my wife Sarah and my sons Tilmann and Samuel who gave me grounding during this emotional rollercoaster called the PhD. Last but not least, I wish to thank my parents and my brother, who have supported me throughout my life.

Contents

List of Tables

List of Figures

Preface

Throughout the life course, individuals experience a number of transitions, such as partnering, becoming a parent, beginning a job, leaving a job, or retiring. These transitions are accompanied by significant changes in individual needs, well-being, and consumption. It is, however, difficult to quantify precisely the effects of these changes. Some of the existing methods for doing so are easy to apply, but they rest on strong assumptions. While there are more realistic approaches, they require the use of complex econometric methods, and have high data demands. This dissertation tackles these methodological challenges, and extends existing methods for which the assumptions appear problematic. Moreover, for more disruptive transitions such as unemployment, the dissertation explores causal pathways and novel sources of data as potential explanatory factors.

The first chapter compares different methods for estimating equivalence scales. The second chapter uses some of these methods to identify benchmark replacement rates. The third chapter extends this idea by estimating income-dependent benchmarks. Finally, the fourth chapter examines the transitions into and out of unemployment, and the degree to which these transitions can be explained by certain biomarkers.

In the first chapter, which is co-authored with Christian Dudel and Marvin Garbuszus, we focus on changes in household composition: e.g., when singles become couples, when couples become parents, or when parents have another child. People's consumption patterns and monetary needs change as they experience these transitions. For example, while a two-person household needs more income than a one-person household, the former needs less income than two separate households due to economies of scale. Exactly how much less income the household needs is assessed using equivalence scales.

Equivalence scales are used to make the incomes of households of different sizes and compositions comparable, and to provide the basis for calculating inequality and poverty measures (e.g., Buhmann et al., 1988). There have been numerous contributions to improving the methodology. Starting with parametric single equation models based on Engel (1895), scholars have applied multi-equation modeling and demand systems, semiparametric estimation, partial identification, and matching. All of these studies generated largely varying equivalence scales. However, as the authors used different datasets, countries, and/or periods, whether the methodology actually drove the results remained unclear.

Therefore, we conducted a broad comparison of these methods using a single dataset, the German Income and Expenditure Survey. We found that the methods indeed produce different equivalence scales. However, while the methods based on unrealistic assumptions resulted in implausible equivalence scales, the methods based on mild assumptions resulted in very similar scales. The latter methods were also close to the modified OECD scale, an approximation that is often used in applied research.

Equivalence scales assess how much income a reference household needs to achieve the same welfare level as a comparison household (Deaton and Muellbauer, 1986). In the second chapter of the dissertation, which is based on joint work with Christian Dudel, we apply this concept to the retirement transition. In particular, we assess how much income a retired individual needs to achieve the same welfare level as a counterfactual working person. To find the appropriate estimation strategy, we take advantage of the comparison of methods conducted in the previous chapter.

When individuals retire, their monetary needs change. As work-related costs, such as commuting expenses, typically decline after retirement, the existing benchmarks suggest that less income is needed in retirement than while working. However, since expenses for leisure or health typically increase with age, retirees may also need more income. The framework we suggest allows us to assess what percentage of a retiree's working income is needed to maintain his or her standard of living. Essentially, our metric is a net replacement rate; i.e., the ratio of after-tax retirement income to after-tax end-of-career working income.

We demonstrate the approach with data from the German Income and Expenditure Survey and the Health and Retirement Study. We find that after taxes in the US and Germany, nearly 100% of a retiree's working income is needed to maintain his or her standard of living. When the standard of living is assessed based on subjective economic well-being, instead of consumption-based welfare indicators, a smaller percentage of the retiree's working income is needed. We also find that parametric, semi-parametric, and non-parametric approaches yield similar results.

While existing pension adequacy standards are mostly based on ad hoc rules (Haveman et al., 2007), this paper contributes to the literature by establishing a general empirical framework with mild data requirements. It gives a population-wide yardstick for saving decisions, which is particularly relevant for Germany. Due to severe population aging, benefits from the statutory pension system have declined significantly. New savings tools have been implemented, but they are used on a voluntary basis. Thus, it is increasingly up to the individual to determine what constitutes an adequate level of saving. We hope that our standards can help individuals achieve that goal.

One issue remains unresolved in the second chapter: namely, whether the adequacy standard can be equally applied to the whole income distribution. Survey evidence suggests that in order to have an adequate standard of living, people with lower incomes need to have a different share of their working income than people with high incomes (Binswanger and Schunk, 2012). We applied econometric tests of income dependence in the second chapter, but the rejection of the Null was dependent on the type of test and the country that was investigated.
Therefore, the third chapter of the dissertation again focuses on the retirement transition, but introduces an income-dependent benchmark. While income-dependent benchmark replacement rates have been used in public policy, the justification for doing so has been flawed. For example, the UK Pension Commission (2004) suggested a benchmark replacement rate of 80% for top earners, and a rate of 50% for low earners. This advice is based on the observation that in the UK, high earners realize lower replacement rates, and vice versa. However, these benchmarks are supposed to reflect the replacement rates that individuals, or a pension system, should aim for. The benchmark replacement rates of the Commission do not take into account whether retirees are satisfied with their financial situation after retirement (Crawford and O'Dea, 2012).
Instead of these rates, I use a replacement rate that maintains income satisfaction through retirement (Dudel et al., 2016). I apply the Generalised Absolute Equivalence Scale Exactness (GAESE) framework, which was originally derived for income-dependent equivalence scales (Donaldson and Pendakur, 2004; Biewen and Juhasz, 2017), and longitudinal data from the German Socio-Economic Panel (SOEP). Applying fixed-effects ordered response models and the blow-up-and cluster strategy, as discussed in Baetschmann et al. (2015), I find that in the UK and Germany, the benchmark decreases significantly with income, from 75% for incomes around 1000 to 64% in Germany.
Lower-income households have to make an extra effort to maintain themselves financially at the welfare level they had before retirement, Extra effort implies, that while they are working, these households have to consume less of their net income to accumulate wealth, as they have replace proportionally more of their earnings than higher earners do. However, given that they usually have a lower capacity to save than high earners, because they have less wealth and higher fixed costs, they might not be able to do so on their own (Crossley and O'Dea, 2010). Supporting them in their efforts might well help to increase their retirement satisfaction.

In the preceding chapters, transitions from one state to another are considered as given, and the effects are measured independently from the cause. For more disruptive transitions, such as unemployment, it is essential to understand why these transitions occur. Unemployment has been shown to be detrimental to health in general, and to mental health in particular, as well as to economic well-being before and after retirement (e.g., Bijlsma et al., 2017; Marcus, 2014; Arulampalam, 2001). Therefore, the fourth chapter, which is a collaboration with Peter Eibich, Ricky Kanabar, and Alexander Plum, estimates employment transition models, and explores biomarkers as potential drivers of unemployment.

According to job search theory, unemployment spells are not only a simple problem of the demand and supply of labor, but are, rather, the result of the employee's rational decision. It is further assumed that these decisions are, in turn, affected by individual characteristics, such as time discounting or risk aversion. However, new sources of data that can be used to explore these issues are now available to social researchers. Biomarkers that measure latent biological processes have become crucial sources of information in the health literature (e.g. Sumner et al., 2020). One biomarker, testosterone, has received special attention in the economic literature, and has been linked to personality traits such as high risk tolerance and aggression, but also to non-cognitive skills such as motivation, pro-social behavior, and persistence (e.g. Gielen et al., 2016; Hughes and Kumari, 2019; Apicella et al., 2008; Carré and McCormick, 2008).

Therefore, we test to what degree the variation in testosterone levels in the population can explain the probabilities of transitioning into and out of unemployment. Using data from the UK Household Study Understanding Society, we follow the individual employment histories for samples of initially employed and initially unemployed British men. Applying dynamic random-effects models to account for unobserved heterogeneity, we find that individuals with high testosterone levels are more likely to become unemployed, but they are also more likely to exit unemployment. Based on previous studies and descriptive evidence, we argue that these effects are likely driven by the personality traits and the occupational sorting of men with high testosterone levels.

Our findings suggest that latent biological processes that are not related to illness and disability can affect job search behavior. Moreover, we find evidence of individual heterogeneity in labor market outcomes that is not explicitly taken into account by conventional job search theory. Ignoring this kind of heterogeneity might lead to misleading evaluations of public measures aimed at increasing the employability of individuals (Uysal and Pohlmeier, 2011).

Chapter 1

Assessing differences in household needs: A comparison of approaches for the estimation of equivalence scales using German expenditure data[1]

1.1 Introduction

Equivalence scales are used to make the incomes of households of different sizes and compositions comparable. They provide the basis for calculating inequality and poverty measures (e.g., Buhmann et al., 1988; Szelky et al., 2004). It has, however, been pointed out that these measures are sensitive to the specific equivalence scale used, and there has so far been no consensus on which equivalence scale should be applied (e.g., Lewbel, 1989a; Blundell and Lewbel, 1991).

A well-known example of an equivalence scale is the so-called modified OECD scale (Hagenaars et al., 1994). The household of an adult living alone is used as a reference, and is assigned a value of one. Adding individuals aged 14 and older to the household increases this value by 0.5 per person, and adding children below age 14 increases it by 0.3 per child. Thus, for instance, a household of two adults with one child has an equivalence scale value of 1.8. Dividing the income of such households by 1.8 yields equivalence income, which is standardized relative to the reference household, and can be directly compared across household types. Another commonly applied equivalence scale is the square root scale, which has been in use at least as long as the modified OECD scale (e.g., Atkinson et al., 1995), and has been applied by the OECD in some of their more recent publications (e.g., OECD, 2008). In this approach, incomes are divided by the square root of the household size. Because they are easy to apply, the modified OECD scale and the square root scale are widely used in applied research.

Apart from these so-called expert scales, a broad range of empirical methods have been proposed for estimating equivalence scales (Phipps and Garner, 1994; Muellbauer

[1] This is a post-peer-review, pre-copyedit version of an article published in *Empirical Economics*.The final authenticated version is available online and open acceess at: 10.1007/s00181-020-01822-6

and van de Ven, 2004). Comparisons of those methods are surprisingly scarce in the literature. Existing studies have focused on subjective approaches (Bellemare et al., 2002; Schwarze, 2003), or have covered expenditure-based approaches that are mostly no longer in use (e.g., Nicholson, 1976; Lancaster and Ray, 1998).

In this paper, we conduct a direct comparison of several different methods for the estimation of equivalence scales using the same dataset, the German Sample Survey of Income and Expenditure (*Einkommens- und Verbrauchsstichprobe*; EVS). We focus on approaches that use expenditure data to estimate a single equivalence scale value per household type that does not vary by household income. Using the classic approach of Engel (1895) as a starting point, we cover the modern methodological developments in the field. These include extensions of the Linear Expenditure System (Lluch, 1973; Howe et al., 1979), which have often been applied to German expenditure data; the quadratic extension (QAI) (Banks et al., 1997) of the influential Almost Ideal Demand System (AI) (Deaton and Muellbauer, 1980b), which is now the standard approach for modeling household demand; semiparametric approaches (Pendakur, 1999; Stengos et al., 2006); and nonparametric approaches based on the counterfactual framework (Szulc, 2009; Dudel, 2015). These methods roughly span a continuum in terms of model complexity, data requirements, and the restrictiveness of the underlying assumptions.

To compare the different approaches for estimating equivalence scales, we apply several parametric, semiparametric, and nonparametric tests that enable us to assess the underlying identifying assumptions of the approaches. We also apply a set of theoretically and empirically grounded criteria that allow us to judge the plausibility of the equivalence scale estimates. These two sets of criteria (identification assumptions; plausibility criteria) can be consistently applied to all methods. To demonstrate the practical relevance of our research, we complement the analysis by using the resulting equivalence scales to calculate indices of inequality and poverty.

We find that a set of approaches lead to results that can be deemed more plausible than the results of other approaches, even though all of these approaches violate at least one of the plausibility criteria. The more plausible estimates are based on demand systems or newer semi- and nonparametric approaches. It appears that equivalence scales based on the more plausible estimates are also similar to the modified OECD scale, at least for households with fewer than two children. For larger families, they are closer to the square root scale.

Our paper contributes to the literature in several ways. To the best of our knowledge, we are conducting the first comparison of methods for the estimation of expenditure-based equivalence scales that covers more recent methodological developments from the literature, and that uses recent data. Our comparison study is motivated by the observation that existing overviews of equivalence scales tend to obscure the differences between the methods applied because the countries, the datasets, and the time periods used in conjunction with these methods vary. For instance, equivalence scale estimates for several different countries are often shown next to each other (e.g., Buhmann et al., 1988). While some countries have similar scales (Phipps and Garner, 1994; Burkhauser et al., 1996), this is not always the case, and discrepancies are possible (Lancaster et al., 1999). Similar issues might arise for equivalence scales based on different datasets because, for example, of differences in the variables used or in the preparation of the data (Dudel et al., 2017a); and for equivalence scales estimated for different points in time because, for example, the prices may have changed (Pendakur, 2002). In our analysis, we try to avoid these issues. Our findings show that while equivalence scales differ considerably, a subset of the approaches in our application leads to more plausible equivalence scales and to consistent results with respect to inequality and poverty measurements.

The remainder of this paper is structured as follows. In section 1.2, we introduce the basic assumptions of equivalence scales, as well as criteria for the assessment of equivalence scales. The approaches we apply to estimate equivalence scales, along with their underlying assumptions, are explained in section 1.3. The dataset we use and the subset selection process are described in section 1.4. In section 1.5, we present results for the tests of the assumptions of the different approaches, and for equivalence estimates. We also compare our estimates with results from earlier literature. Section 1.6 concludes.

1.2 Equivalence scales

1.2.1 Preliminaries and basic definition of equivalence scales

Let $\mathbf{z} = (z_1, \dots, z_k)$ denote a vector of k household characteristics, such as household size, number of children, or age of household members. All households can choose between m goods with prices captured in a vector $\mathbf{p} = (p_1, \dots, p_m)$. Household demand is given by the demand function $D(p, y, z) = \mathbf{q} = (q_1, \dots, q_m)$, where q_i is the demand for good i and y is household income. Household utility is given by $U(\mathbf{q}, z)$. The expenditure function can be defined by $E(u, \mathbf{p}, \mathbf{z}) = \min_q [\mathbf{p}'\mathbf{q} | U(\mathbf{q}, z) = u]$.[2] Using these

[2] Note that household utility functions typically ignore the distribution of resources within the household, and may thus be hard to defend, as it is individual household members who derive utility from consumption (Phipps and Burton, 1995). Still, household utility functions are the theoretical foundation of equivalence scales, and including individual needs and preferences and intrafamily bargaining in the derivation of equivalence scales is beyond the scope of this paper.

preliminaries, household equivalence scales are defined as

$$S(u, \mathbf{p}, \mathbf{z}_h, \mathbf{z}_r) = \frac{E(u, \mathbf{p}, \mathbf{z}_h)}{E(u, \mathbf{p}, \mathbf{z}_r)}, \tag{1.1}$$

where $\mathbf{z}_h$ and $\mathbf{z}_r$ are the household characteristics of two different households h and r. Thus, an equivalence scale is a function that returns the ratio of the expenditures of two households of different compositions with the same level of utility and facing the same prices. The reference household $\mathbf{z}_r$ is usually fixed as the household of a single adult, but any other household type could also be chosen. Throughout our analysis, we will often assume the former type, and will then write $S(u, \mathbf{p}, \mathbf{z}_h)$, thus dropping $\mathbf{z}_r$.

1.2.2 Assessing equivalence scales: Identification, income independence, and Engel curves

Equivalence scales as defined by equation (1.1) are not identified if ordinal utility is assumed (Pollak and Wales, 1979; Lewbel, 1989b; Blundell and Lewbel, 1991; Pollak, 1991). This is because equivalence scales require interpersonal comparisons of utility that are not possible under the assumption of ordinal utility. Any approach for estimating equivalence scales has to deal with this issue of identification. Three main approaches for obtaining equivalence scales are used in the literature. The first approach is based on experts' more or less heuristic assessments of equivalence scales (see Fisher, 2007, for a review). The second approach is based on individuals' subjective evaluations of utility drawn from income (see Schröder, 2004, for a review). This approach has, for example, been applied to survey data on income satisfaction (e.g., Schwarze, 2003; Biewen and Juhasz, 2017; Borah et al., 2019), and to customized survey data that directly relate specific income levels to specific welfare levels (Koulovatianos et al., 2005). The third main approach is based on consumption and expenditure data; this approach will be the focus of our study.

In expenditure-based approaches, a common solution to the identification problem is to employ (indirect) utility functions of a certain structure. For instance, if we assume that equivalence scales do not depend on the welfare level – i.e., $S(u, \mathbf{p}, \mathbf{z}_h) = S(\mathbf{p}, \mathbf{z}_h)$ – they can be identified (e.g., Blundell and Lewbel, 1991). This assumption is called, or is related to, *independence of base* (Lewbel, 1989b) and *equivalence scale exactness* (Blackorby and Donaldson, 1993) (IB/ESE). For practical purposes, this assumption often – but not always – implies that equivalence scales do not depend on the income levels (or expenditure levels) of the households under consideration. More specifically, equivalence scales are considered income-independent if the same value is applied to all households of a certain type.[3]

In practice, the independence of base is connected to assumptions about the functional form of Engel curves. Depending on the approach used for estimating equivalence scales, assumptions of varying levels of generality are applied. These assumptions can be tested empirically, which allows us to judge whether the corresponding approaches yield trustworthy estimates. In section 3, we will discuss approaches that require (1) linear or quadratic Engel curves, which are only shifted by a constant for different household types; (2) arbitrarily shaped Engel curves, but which are only shifted by a constant for different household types, and thus have the same shape for all household types; (3) and arbitrarily shaped Engel curves with no restrictions across household types, which also implies that unlike for the first and second types of Engel curves, income independence does not hold.

1.2.3 Assessing equivalence scales: Plausibility

In addition to applying the identification assumptions discussed above, we assess approaches for equivalence scale estimation by the resulting scale values; i.e., the values $S(u, \mathbf{p}, \mathbf{z}_h)$ attains for different values of $\mathbf{z}_h$. In the literature, several criteria have been discussed based on economic theory and empirical regularities. While some of these criteria can be seen as properties that equivalence scales have to exhibit to be deemed plausible, other criteria are more debatable. None of the approaches we apply leads to estimates that satisfy any of the criteria by design, and all of the approaches could lead to estimates that violate one or several of the criteria.

To describe the criteria formally, we assume that the equivalence scales only depend on household size n, such that they can be written as $S(u, \mathbf{p}, n)$; or, alternatively, that equivalence scales depend on the number of adults n_a and the number of children n_c,

[3] More recently, approaches have been proposed that relax the *independence of base* assumption (e.g., Donaldson and Pendakur, 2004, 2006; Garbuszus, 2018), and several studies – often based on subjective approaches to equivalence scales – have supported the idea of equivalence scales decreasing in income (e.g., Koulovatianos et al., 2005; Biewen and Juhasz, 2017). Another strand of the literature has focused on the estimation of indifference scales (e.g., Chiappori, 2016), which are designed to measure individual welfare within households. We did not implement these approaches in this paper because they require data that are not provided in our dataset. Moreover, these scales have not been broadly adopted in applied welfare analysis and poverty research, in which equivalence scales based on the independence of base assumption remain the standard approaches used.

$S(u, \mathbf{p}, n_a, n_c)$. Using this notation, we discuss the following criteria:

$$\begin{aligned} S(u, \mathbf{p}, n+1) &> S(u, \mathbf{p}, n), & (1.2) \\ S(u, \mathbf{p}, n+1) &\leq S(u, \mathbf{p}, n) + 1, & (1.3) \\ S(u, \mathbf{p}, n+i+1) - S(u, \mathbf{p}, n+i) &\leq S(u, \mathbf{p}, n+i) - S(u, \mathbf{p}, n+i-1), & (1.4) \\ S(u, \mathbf{p}, n_a+1, n_c) &> S(u, \mathbf{p}, n_a, n_c+1), & (1.5) \end{aligned}$$

The criterion stated in equation (1.2) has been referred to as the “household size effect” (Stengos et al., 2006), and indicates that equivalence scales have to be strictly increasing functions of household size. Using the household of a single person as a reference with $n = 1$ thus implies that for $n > 1$, the equivalence scale has to be larger than one. The assumption underlying this criterion is that every additional household member generates costs; i.e., $E(u, \mathbf{p}, n+1) > E(u, \mathbf{p}, n)$. As this criterion is generally accepted in the literature, many studies have used it to evaluate the plausibility of equivalence scales (e.g., Deaton and Muellbauer, 1986; Wilke, 2006; Stengos et al., 2006).

Criterion (1.3) states that the effect of the household size must be no more than one, due to economies of scale. Larger values would indicate, for example, that a couple needs more than two singles. This is unlikely, because of economies of scale in consumption. Two adults can reduce their costs when, for example, they cook together; children often share rooms (see Deaton and Muellbauer, 1980a, for more examples). These observations also motivate criterion (1.4), which states that the scale increase diminishes with household size or at least remains constant. In other words, every additional household member adds less – or at least does not add more – to the scale than the previous one. There might be some constellations in which (1.4) does not hold. For example, a couple might have enough space in their current home for a first child, but if having a second child compels them to move into a larger dwelling. Therefore, adding the second child would be more expensive than adding the first, which demonstrates that there could be exceptions to criterion (1.4).

The fourth criterion in equation (1.5) states that an additional adult adds more to the equivalence scale than a child. This is based on the assumption that children generate lower costs than adults, because, for instance, they consume less food. The extent to which this criterion holds might depend on the age threshold used to distinguish between adults and children.

1.3 Expenditure-based methods for the estimation of equivalence scales

1.3.1 Engel's approach

The idea of using household expenditures to assess household welfare is usually attributed to Engel (1895), and is based on the observation that the share of household expenditures spent on food depends on household type, and declines as income rises. Assuming that two households achieve the same level of welfare if the shares of their expenditures allocated to food are equal, the equivalence scales can be identified by comparing the incomes of different types of households that allocate the same share of their expenditures to food.

This approach can be implemented as follows (Deaton and Muellbauer, 1986). Letting w_f denote the share of expenditures on food, the following regression equation, as proposed by Working (1943), can be estimated based on demand data (also see Leser, 1963):

$$w_f = \alpha + \beta_x \log(x/n) + \beta_a n_a + \beta_c n_c + \gamma' \mathbf{z}, \tag{1.6}$$

where x is total expenditure, x/n is per capita expenditure, n_a and n_c denote the number of adults and children in the household, respectively; and $\mathbf{z}$ captures socio-demographic variables other than household type. Now let us consider two households that allocate the same share of their expenditures to food as given by equation (1.6), but that are of different types. Equating expenditure shares and solving for the ratio of incomes x_h and x_r that the households need to achieve the share spent on food gives

$$S = \frac{x_h}{x_r} = \left(\frac{n_h}{n_r}\right) \exp\left(\frac{\beta_a}{\beta_x}(n_{a,h} - n_{a,r}) + \frac{\beta_c}{\beta_x}(n_{c,h} - n_{c,r})\right), \tag{1.7}$$

where n_r is the size of the reference household, and $n_{a,r}$ and $n_{c,r}$ capture the number of adults and children in the reference household. n_h, $n_{a,h}$, and $n_{c,h}$ are defined in a similar way for the comparison household.

This approach assumes that equivalence scales do not depend on income or expenditure levels. Moreover, prices are usually not included, even though it would be possible to do so. Thus, this approach has low data requirements, and is easy to apply. Engel curves, as defined by (1.6), are linear. While linear Engel curves are not necessary for applying this approach (Leser, 1963), empirical applications typically use linear Engel curves.

One popular variant of the Engel approach was suggested by Rothbarth (1943). His idea was to assess the utility of adults by considering goods that are exclusively consumed by adults, such as tobacco, alcohol, and adult clothes. Compared to a couple without children, a couple with children needs to be compensated to the extent that the household resets its expenditures on those adult goods to the level of the reference household (Lancaster and Ray, 1998).

1.3.2 Linear expenditure system and extensions

The Linear Expenditure System (LES) proposed by Stone (1954) is the earliest full expenditure system; meaning that it is based not on a single equation, but on a system of equations, each of which covers expenditures for one of the m goods. It also takes into account price changes, which makes it possible to impose and test restrictions of economic utility theory. [4]

Starting from a Stone-Geary utility function, the following set of m expenditure functions can be derived:

$$x_i \quad = \quad p_i a_i + b_i \left(x - \sum_{j=1}^{m} p_j a_j \right) \tag{1.8}$$

with x denoting total expenditures and $x_i = p_i q_i$, i.e., expenditure on good i; $p_i a_i$ being interpreted as the minimum expenditure on good i; and b_i being the marginal budget share of good i, with the restriction that $\sum b_i = 1$.

This set of equations can be estimated separately for each household type (for an estimation of the LES, see, e.g., Deaton, 1975). Given these parameter estimates, a pragmatic way to calculate the equivalence scales is based on a comparison of the minimum expenditures by household type (e.g., Kohn and Missong, 2003), while p_i is set to one.

$$S = \frac{\sum_{i=1}^{m} a_i^h}{\sum_{i=1}^{m} a_i^r}, \tag{1.9}$$

[4] The LES imposes the restrictions of adding-up, homogeneity, and symmetry. Adding up requires the total value of the demand functions to equal total expenditure; homogeneity requires the demand function to be homogeneous of degree zero in prices; symmetry requires the cross-price derivatives of the demand to be symmetric (for a more complete discussion, see Deaton and Muellbauer, 1980b). In principle, those criteria could also be used for evaluating the applied approaches, and demand systems have been regularly tested for their consistency with utility theory (e.g., Haag et al., 2009). In this comparison of approaches, however, we refrain from applying these criteria, as most of the applied approaches are independent of prices, and the (cross-) price elasticities that are needed for testing cannot be derived or tested in these approaches. In both the expenditure systems and the demand systems, adding up is automatically satisfied when estimating with ordinary least squares. Homogeneity and Slutsky symmetry have been rejected in all of the demand systems but the QAI demand system (Deaton and Muellbauer, 1980b; Blundell et al., 1998).

where a_i^r is the reference household's minimum expenditure on good i facing prices p for good i; a_i^h is the comparison household's minimum expenditure on good i facing prices p for good i. The LES has inspired several extensions, of which we cover two variants: the Extended Linear Expenditure System (ELES; Lluch, 1973) and the Quadratic Expenditure System (QES; Howe et al., 1979). Essentially, the ELES expands the LES by introducing saving, which is treated as an additional commodity. In contrast to the linear Engel curves of the LES, the QES assumes a quadratic relationship between expenditure and (marginal) total expenditure. For both variants, the equivalence scale can be calculated in the same way as in the basic LES.

In terms of data demands, the LES and its extensions fall somewhere in the middle: expenditure data are needed for several expenditure categories; whereas data on prices can be included, but are not needed, as p_i can be set to one. Equivalence scales based on linear expenditure systems are income-independent; although the QES uses quadratic Engel curves instead of linear curves.

1.3.3 Almost ideal demand system and extensions

The AI system arose from the search for a model that provides a good fit for empirical demand data, while having properties deemed desirable for demand systems.[5] Starting from the price-independent generalized logarithmic (PIGLOG) class of preferences, the expenditure share for good i, w_i can be derived to equal:

$$w_i = \alpha_i + \sum_{j=1}^{m} \gamma_{ij} \log p_j + \beta_i \log\left(\frac{x}{P}\right), \tag{1.10}$$

with

$$\log P = \alpha_0 + \sum_{i=1}^{m} \alpha_i \log p_i + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \gamma_{ij} \log p_i \log p_j.$$

with γ_{ij} capturing the effect of the price of good j on the share of expenditures on good i, β_i being the marginal effect of log income, and α_i being a parameter. P is a price deflator for income. As P makes the model nonlinear, in empirical applications linear approximations are often used (see, e.g., Barnett and Serletis, 2008). Here, we will use the (nonlinear) translog price index, as proposed by Deaton and Muellbauer (1980b).

To estimate equivalence scales, some parameters have to be added to the AI demand system. We follow a general approach suggested by Ray (1983) for introducing equiva-

[5] The AI demand system is included because of its importance for empirical work throughout the years. Although it is now known that PIGLOG equivalence scales lack identification (Pendakur, 1999), the model was widely used for a long period of time.

lence scales in demand systems. If we want to compare the reference household to one other household type only, this approach is implemented by using:

$$w_i = \alpha_i + \sum_j \gamma_{ij} \log p_j + \beta_i^* \log\left(\frac{x}{SP}\right), \tag{1.11}$$

where

$$\begin{aligned} S &= 1 + \rho d_h \\ \beta_i^* &= \beta_i + \eta_i d_h \end{aligned}$$

while assuming that the comparison household needs more resources than the reference household. S denotes the equivalence scale value. d_h is a dummy for the respective household comparison type while ρ captures the needs of the comparison households relative to the needs of the reference household. η_i plus β_i gives the income elasticity for the comparison household. Given P, the parameters can be found using nonlinear, seemingly unrelated regressions (Greene, 2012).

The AI demand system essentially assumes that the relationship between log income and expenditure shares is linear. But for some commodities, this relationship has been found to be nonlinear. To account for the nonlinearity, and to provide a better fit for the demand data, Banks et al. (1997) introduced the Quadratic AI demand system. The QAI demand system essentially includes an additional quadratic term of (deflated) log income. Equivalence scales are estimated by expanding the approach of Ray (1983) to cover this term.

While the AI and the QAI demand systems are rather flexible models that can fit many patterns of household demand, they also require data on prices. Thus, unlike in Engel's approach, at least two cross-sections of demand data are required in these systems. Equivalence scale exactness is also required.

1.3.4 Semiparametric approaches

The approaches presented so far all rely on the assumption that the relationship between log (deflated) income and expenditure or expenditure shares is linear or quadratic. While this assumption might be appropriate for some commodities, it might not hold for others (Banks et al., 1997). In an effort to address this problem, Pendakur (1999) developed a semiparametric approach to estimating equivalence scales that avoids strong assumptions regarding the relationship between income and expenditure shares by estimating nonlinear Engel curves. Writing the expenditure share for food, w_f, as a function of income y, prices p, and household type d_h, the approach assumes that

$$w_f(p, \log(y), d_h = 0) = w_f(p, \log(y) + \phi, d_h = 1) + \mu(p). \tag{1.12}$$

Here, the relationship between log income and the expenditure share for food as captured by $w_f(p, \log(y), d_h)$ can be of any functional form. It is, however, assumed that this functional form is equal across household types ("shape invariance"), and is only shifted vertically by price elasticity, $\mu(p)$, and horizontally by the log equivalence scale ϕ. Equivalence scales can be calculated as

$$S = \exp(\phi). \tag{1.13}$$

Estimation proceeds by using nonparametric methods to estimate the shape of $w_f(p, \log(y), d_h = 0)$ and of $w_f(p, \log(y), d_h = 1)$. In a second step, assuming constant prices, the log equivalence scale ϕ is found via a grid search, whereby the difference between the two sides of equation (1.12) is minimized (Pendakur, 1999). Stengos et al. (2006) proposed a variant of this method, which we also include in the set of methods we apply. They modified the second step of the approach, penalizing high or low values of ϕ. This yields more plausible estimates than the original method of Pendakur (1999), particularly for comparisons in which the income distributions of the reference and the comparison household types overlap slightly, as the loss function used by Pendakur (1999) is deficient in this case.

While the semiparametric approach is flexible regarding the functional form of Engel curves, it requires the independence of base assumption (Pendakur, 1999). The data requirements are relatively low, as a single cross-section of data suffices. In principle, the share of expenditures on food can be replaced with the share of expenditures on other commodities. For instance, it would be possible to implement the ideas of Rothbarth (1943) in a semiparametric way (see Section 3.1). A drawback of the semiparametric approach is that including covariates in the first estimation step is not straightforward. Moreover, the approach relies to some extent on the selection of homogenous subsets of households.

1.3.5 Counterfactual approaches

The counterfactual approach rephrases equivalence scales in the potential outcomes framework (e.g., Holland, 1986). Let us assume that in theory, every household can be considered to belong to the reference household type (e.g., single-adult household) and the comparison household type (e.g., couple with one child). $y^0(u)$ is the income needed to achieve utility u when the household is of the reference type, and $y^1(u)$ is the income needed to achieve utility u when the household is of the comparison type.

Assuming that a household achieves utility level u^0 when it is of the reference type, equivalence scales are given by $\mathrm{E}[y^1(u^0)/y^0(u^0)]$ (Szulc, 2009; Dudel, 2015). Note that this definition differs from the common definition of average treatment effects, where a difference is used instead of a ratio. Because of the ratio, $\mathrm{E}[y^1(u^0)/y^0(u^0)]$ is not point-identified using standard assumptions.

More specifically, either $y^0(u)$ or $y^1(u)$ is observed; never both. That is, at any point in time, some households are observed as being of the reference type, but not of the comparison type, and vice versa. Still, under some assumptions, the marginal distributions of $y^0(u)$ or $y^1(u)$ can be estimated (e.g., Imbens, 2004). However, this strategy is not sufficient for estimating equivalence scales. Based on these expectations and after applying some simple algebra, the identification problem becomes clearer in (1.14).

$$\mathrm{E}\left[\frac{y^1(u^0)}{y^0(u^0)}\right] = \frac{\mathrm{E}\left[y^1(u^0)\right]}{\mathrm{E}\left[y^0(u^0)\right]} - \frac{1}{\mathrm{E}\left[y^0(u^0)\right]}\mathrm{Cov}\left[\frac{y^1(u^0)}{y^0(u^0)}, y^0(u^0)\right]. \tag{1.14}$$

The covariance term on the right-hand side requires the joint distribution of $y^0(u)$ and $y^1(u)$, which is not point-identified (Abbring and Heckman, 2007). Szulc (2009) avoided this problem by estimating the geometric mean of $y^1(u^0)/y^0(u^0)$ instead of (1.14), while Dudel (2015) has proposed the use of lower and upper bounds on (1.14). That is, the equivalence scales are not point-identified. For the comparison of, say, childless couples and couples with one child, the equivalence scales do not take on one specific value S, but can only be shown to be in an interval $[S^-, S^+]$. Here, we adopt this partially identified approach, as well as the approach of Szulc (2009). In the partially identified approach, estimation proceeds using a nonparametric method suggested by Fan et al. (2017). The approach of Szulc (2009) follows Abadie and Imbens (2006), and applies the Mahalanobis distance for the pair-matching of households.

In contrast to previous approaches, this identification strategy does not rely on the assumption that equivalence scales are independent of the welfare level. Furthermore, it does not rely on any specific Engel curve shape. While the partially identified approach requires few assumptions, it does not allow us to produce any point estimates. Moreover, the interval estimates generated using this approach might not be informative if they are too wide. The method proposed by Szulc (2009) avoids this issue by estimating the geometric mean; but the geometric mean will always be lower than arithmetic mean, and an increase in the variance of $y^1(u^0)/y^0(u^0)$ will push the geometric mean further away from the arithmetic mean (Cartwright and Field, 1978), leading to potentially biased estimates.

1.3.6 Testing linearity of Engel curves, shape invariance, and income independence

Most of the methods described above rely on one of three assumptions (See Table 1.1). These are, ordered by increasing generality: linearity of Engel curves, shape invariance, and income independence. Linearity of Engel curves implies shape invariance and income independence; and shape invariance implies income independence. On the other hand, income independence does not imply linearity or shape invariance. That is, both linearity and shape invariance are sufficient, but not necessary, for income independence.[6] In the literature, several tests have been proposed to assess these assumptions.

To test whether Engel curves are linear, we use two approaches. First, as suggested by Lancaster and Ray (1998), we include a quadratic term for log income in the Engel approach; i.e., a quadratic term $\beta_{x2} \log(x)^2$ is added to equation (1.6). If this term is statistically significant, then linearity of Engel curves can be rejected. Second, in a similar vein, we check the statistical significance of the coefficients of the quadratic income terms in the QAI demand system (Banks et al., 1997). In line with the previous literature, we call those coefficients λ-parameter. For each expenditure category, there is one such coefficient; in our case, there are 12 coefficients.

For testing shape invariance, we apply three approaches. First, we add a new term to the main equation of the Engel approach, interacting household type and log income, as proposed by Pendakur (1999). If the coefficient is significant, then the regression line for the comparison household is not only shifted relative to the reference household, but is rotated, and shape invariance can be rejected. Second, we calculate a correlation between the reference Engel curve and the shifted Engel curve. Hacing values close to one can be regarded as a necessary, but not a sufficient condition of shape invariance (Stengos et al., 2006). Third, we use simulations to calculate the probability that the empirical goodness-of-fit of the semiparametric approach is observed given shape invariance. If this probability is below the conventional thresholds, shape invariance is rejected. For details on the implementation, see Pendakur (1999). Here, we use the loss function proposed by Stengos et al. (2006).

In addition to these parametric and semiparametric tests, we apply two nonparametric approaches. The first approach allows us to check both linearity of Engel curves and shape invariance, and relies on the visual inspection of nonparametrically estimated Engel curves (Banks et al., 1997). The second method is based on the nonparametric,

[6] In some rare cases, shape invariance is not sufficient for income independence (Lewbel, 2010).

Table 1.1: Empirical approaches to estimate equivalence scales and their underlying assumption and properties.

Approach	Linearity of Engel curves	Shape invariance	Income independence	Price variation	Covered commodity groups
Engel's approach					
Engel (food)	✓	i	i	No	Food
Rothbarth (adult goods)	✓	i	i	No	Alcohol
Expenditure systems					
Extended linear (ELES)	✓	i	i	No	All + savings
Quadratic (QES)	×	×	✓	No	All
Demand systems					
Almost ideal (AI)	✓	i	i	Yes	All
Quadratic almost ideal (QAI)	×	×	✓	Yes	All
Semiparametric					
Original loss function	×	✓	i	No	Food
Modified loss function	×	✓	i	No	Food
Counterfactual					
Matching	×	×	×	No	Food
Partial identification	×	×	×	No	Food

Note: × not required for the approach; ✓required for the approach; *i* implied by the more general assumption on the left.

partially identified approach. A confidence interval on the bounds of the covariance term on the right-hand side of equation (1.14), $\mathrm{Cov}\,[y^1(u^0)/y^0(u^0), y^0(u^0)]$, is estimated. If this confidence interval does not include zero, which is the value of the covariance that implies income independence, then income independence can be rejected.

All of the tests described above are applied for each household type; e.g., couples without children or couples with one child. Thus, it is possible that an assumption might be rejected for one household type, but not for other types.

1.4 Data and implementation

1.4.1 Data and sample selection

We applied the methods described in the previous section to data of the German Sample Survey of Income and Expenditure (*Einkommens- und Verbrauchsstichprobe*; EVS). The EVS is a quinquennial survey conducted by the German Federal Statistical Office that covers about 0.2% of households in Germany. We used data from the years 2003, 2008, and 2013. The three cross-sections of the EVS contain nearly 130,000 households in total. For each household, detailed information on the household's income, expenditures, and savings is collected for one quarter of the year.

To reduce the heterogeneity of the sample and to ease the interpretation of the equivalence scale estimates, we selected a certain subset of households. We dropped about 34,000 households in which at least one of the adults was over age 65. Pensioners are not of major interest when calculating equivalence scales for children, as it may be expected that in most cases, their children have left the household. Based on a similar reasoning, we excluded another 14,000 households in which the children were over age 18. Next, we restricted the set of households to those residing in Western Germany, as there are large economic differences between Eastern and Western Germany (Brenke and Zimmermann, 2009). This reduced the sample by another 12,000 observations.

For some household types, there were not enough observations to produce precisely estimated equivalence scales. This led us to exclude a few hundred families with more than three children and about 3,000 single-parent families.[7] We also excluded about 20,000 households that were dependent on welfare benefits, because otherwise our equivalence scales might be influenced by the equivalence scales implied by the welfare benefits received by different household types. In Germany, for example, welfare benefit levels are partly set using equivalence scales. A couple is assumed to need 1.8 times

[7] As the overall sample size for single-parent households was small, it was not possible to further distinguish these households by the number of children. We also decided against including single-parent households as one group, as it would have been rather heterogeneous.

as much income as a single adult, and the welfare benefits the couple receives are set accordingly. Including low-income households then runs the risk of replicating this equivalence scale, which was created by policy-makers based not on differences in the behavior of households, but on assumptions made by politicians. For the same reason, we dropped about 300 households with a net income below the approximated welfare benefit level (excluding housing costs).[8]

Finally, we have tried to make the incomes and the expenditures of different households as comparable as possible. For example, when a family's housing is paid for by an employer, the household's income is not comparable to that of a household paying rent. Thus, we dropped 1,200 cases in which an employer was covering these costs. Furthermore, in line with a common practice in the literature (e.g., Donaldson and Pendakur, 2004), we removed 600 households that reported extreme income values and 6,800 households that reported extreme expenditure values. These values were considered extreme if they exceeded the sample median plus two and a half standard deviations (Banks et al., 1997). Spending above this threshold is usually attributable to highly irregular expenses (e.g., buying a car, a health shock), which can have large effects on demand system estimates. Levels of extreme spending were not highly correlated across the 12 categories, and most outliers only counted as outliers for one of the categories. Households with zero expenditures on food were also dismissed (10 households). The final sample consisted of about 32,000 households (about 11,000 households in the EVS 2013). The descriptive statistics are reported in Tables 1.2 and 1.3.

1.4.2 Main variables

Expenditure information in the EVS is collected based on a German equivalent of the United Nations' Classification of Individual Consumption According to Purpose (COICOP). Total expenditures are broken down into 12 commodity groups: (1) food and non-alcoholic beverages; (2) alcoholic beverages and tobacco; (3) clothing and footwear; (4) housing, water, electricity, and heating; (5) furniture, household equipment, and routine household maintenance; (6) health; (7) transportation; (8) communication; (9) recreation and culture; (10) education; (11) restaurants and hotels; and (12) miscellaneous goods and services. While these expenditure categories are, in turn, based on more detailed expenditure information, for our estimation, we used only these 12 categories. Price information for each of the 12 expenditure categories was provided by the German Federal Statistical Office. Monthly prices were aggregated into quarterly prices by calculating the average. We thus included annual price variation between the years 2003,

[8] In 2013, Germany granted benefits of EUR 382 to a single adult, EUR 690 to a childless couple, and EUR 224 for additional children.

Table 1.2: Descriptive statistics

Monthly* total expenditures (in EURO), mean	2,440
Monthly* net income (in EURO), mean	3,703
Monthly* net income (in EURO), min	382
Monthly* net income (in EURO), max	12,752
Age of the household head (in YEARS), mean	43
Share of single households (A, in PERCENT)	36
Share of couple households (AA in PERCENT)	29
Share of couple households with one child (AAC, in PERCENT)	14
Share of couple households with two children (AACC, in PERCENT)	16
Share of couple households with two children (AACCC, in PERCENT)	4
Share of tenures (in PERCENT)	48
Share of low educated (in PERCENT)**	5
Share of higher educated (in PERCENT)***	42
Share of people from low density areas (in PERCENT)	10
Share of dual earners (in PERCENT)	26

Note: * The reporting period of the EVS denotes three months: the values shown are divided by three, and can therefore be regarded as approximately monthly. ** Including individuals with no degree or a degree from a ”Hauptschule”. *** Including individuals with ”Fachabitur” or ”Abitur”.

Data: German Sample Survey of Income and Expenditure 2003, 2008, 2013.

2008, and 2013; as well as seasonal variation within these years.[9]

The socio-demographic variables we used included the number of adults and the number of children under age 18 in each household. The household type was assigned based on these two variables. We distinguished between households made up of a single adult (A), a childless couple (AA), a couple with one child (AAC), a couple with two children (AACC), and a couple with three children (AACCC) (see Table 1.2 for the sample composition with respect to the household type). Single-adult households were used as the reference household type for all equivalence scales.

Additional control variables were dummy variables indicating whether both partners in a couple were full-time employed; as well as variables capturing the quarter of the year (spring, summer, autumn, winter), the age and the level of education (1 = no education, 2 = vocational training, 3 = foreman, 4 = college, 5 = university degree) of the household head, the type of region (ranging from one for rural areas to seven for densely populated areas in cities), and a dummy variable for homeownership. We included full-time employment of both partners as a dummy, because these couples likely differed from other couples in the time they had available for home production, and, thus,

[9] As the German Federal Statistical Office from which we obtained our price indices for Germany does not provide regional price indices, we could not control for regional variation in prices.

Table 1.3: Expenditure categories

	Monthly* expenditures in EURO				Expenditure shares in %				Change of prices (log)			
	Min	Mean	Sd	Max	Min	Mean	Sd	Max	Min	Mean	Sd	Max
Food	8.3	312.8	170.1	1,128.0	0.4	13.1	5.2	55.0	4.6	4.7	0.0	4.7
Alcohol & tobacco	0.0	38.2	48.7	302.0	0.0	1.7	2.3	21.2	4.6	4.6	0.1	4.7
Clothing	0.0	135.6	116.8	802.7	0.0	5.4	3.8	32.1	4.6	4.6	0.0	4.7
Housing	40.0	829.8	357.3	2,465.0	5.5	35.9	10.8	89.5	4.6	4.6	0.0	4.7
Furniture	0.0	115.3	171.4	1,529.0	0.0	4.2	5.3	51.5	4.6	4.6	0.0	4.6
Health	0.0	75.7	130.3	1,216.7	0.0	2.8	4.2	41.9	4.6	4.6	0.0	4.6
Transportation	0.0	319.5	360.4	5,046.7	0.0	12.2	8.8	83.0	4.6	4.6	0.0	4.7
Communication	0.0	69.9	37.3	266.0	0.0	3.2	1.9	22.4	4.5	4.6	0.1	4.7
Recreation	0.0	278.6	255.1	1,907.7	0.0	10.8	7.4	56.3	4.6	4.6	0.0	4.7
Education	0.0	24.3	56.5	602.0	0.0	0.9	2.3	36.1	4.5	4.6	0.0	4.7
Restaurants	0.0	141.4	141.6	968.7	0.0	5.6	4.8	48.4	4.6	4.6	0.0	4.7
Miscellaneous	0.0	98.9	92.9	818.7	0.0	4.0	3.2	34.4	4.6	4.6	0.0	4.7

Note: * The reporting period of the EVS denotes three months: the values shown are divided by three, and can therefore be regarded as approximately monthly.
Data: German Sample Survey of Income and Expenditure 2003, 2008, 2013.

in their expenditures. Including the quarter of the year allowed us to control for seasonal spending (e.g., vacations); including the type of region allowed us to indirectly capture price differences affecting behavior, like higher rents in cities; including homeownership enabled us to determine whether households had rent expenditures, which could represent a sizable proportion of household expenditures; and age and education allowed us to control for further heterogeneity in household spending.

1.4.3 Implementation

In this section, we briefly provide some details concerning the implementation of the approaches (see section 3 for the theoretical concept of the approaches, or, for further details, see the studies that introduced the methods shown in Table 1.6).

First, to ease the comparison between the methods, we used total expenditures instead of income in all of the approaches but the ELES. While in the single parametric, the semiparamteric, and the counterfactual approaches, it was feasible to use either income or total expenditures, the ELES was explicitly designed to use income. Second, all of the single-equation models were estimated without price information, and were based on the 2013 EVS sample. The demand systems, on the other hand, included price information for 2003, 2008, and 2013. Third, for the approach of Rothbarth (1943), we used alcohol as the adult good. In order to obtain reasonable results, we excluded families with zero expenditures on these commodities (García and Labeaga, 1996). As a large number of the families in our sample had zero expenditures (about 2,400), this sample restriction was applied only to this approach. Fourth, for the semiparametric approaches, we sought to find the values of ϕ and μ that minimize equation (1.12) by inserting start intervals that increase with household size – that is, 0.9 and 2.0 for AA, 0.9 and 2.2 for AAC, 0.9 and 2.5 for AACC, and 0.9 and 3.5 for AACCC for ϕ – and used increments of 0.01.

The ability of the applied approaches to consider control variables was limited in some cases. For example, as the estimation of the Engel curves in the semiparametric approach was pursued nonparametrically, it did not allow for the consideration of control variables. In some of the other approaches, the control variables were not used in a conventional way. For example, in the matching approach by Szulc (2009) the control variables were used as matching variables. Moreover, in the nonparamteric approach by Dudel (2015) nonparamateric densities were calculated conditional on the control variables.

Depending on the specific approach applied, estimation was carried out using OLS as implemented in base R; nonlinear, seemingly unrelated regression as implemented in the R package nlsur (Garbuszus, 2017); nonparametric kernel methods as implemented in the R package np (Hayfield and Racine, 2008); and pair-matching as implemented in the R package Matching (Sekhon, 2011).

To make standard errors between methods as comparable as possible, we calculated bootstrapped standard errors for every approach. However, for the QAI demand system and the QES, bootstrapping was computationally out of reach. For the QAI demand system, we used analytic standard errors (Ray, 1983).[10] For the rest of the approaches, we applied the resampling bootstrap and used 500 replications. The confidence intervals were based on percentiles of the bootstrap replications. Constructing confidence intervals for the nonparametric bounds by Dudel (2015) was not straightforward. Our general aim was to construct an interval that covered the complete identification region with a fixed probability (95%). Further details are provided in the supplementary materials.

1.5 Results

1.5.1 Testing identifying assumptions: Linearity, shape invariance, income independence

Table 1.4: λ-parameters of the QAI demand system.

λ	coefficient
Food	-0.011***
Alcohol, Tobacco	-0.005***
Clothing	-0.019***
Housing	0.001
Furniture	0.008***
Health	0.002
Transportation	0.065***
Communication	-0.004***
Recreation	-0.011***
Education	-0.001.
Restaurants	-0.015***
Miscellaneous	-0.010***

Note: Significant coefficients indicate that a quadratic specification provides the better fit of the data; p ¡ 0.1, * p ¡ 0.05, ** p ¡ 0.01, *** p ¡ 0.001.
Data: German Sample Survey of Income and Expenditure 2003, 2008, 2013.

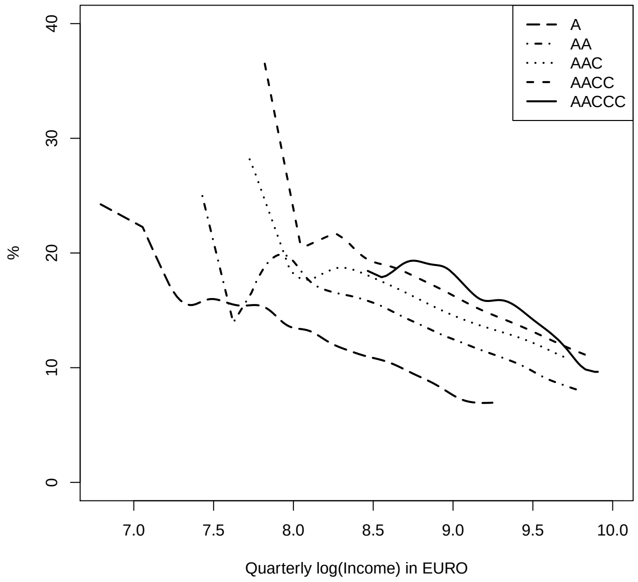
Before we present the equivalence scale estimates, we discuss the results of the econometric tests regarding the identifying assumptions of the different approaches: namely,

[10] For the AI, we calculated both analytic and bootstrapped standard errors, and found them to be almost identical.

linearity of Engel curves, shape invariance, and income independence (see also section 3.6).

The results for the linearity of Engel curves depended on the test, the commodity, and the household type used; but, overall, they indicate that linearity can be rejected. Estimating Engel's approach as in equation (1.6) with an additional quadratic term of log per capita income gives a p-value of 0.065 for the resulting coefficient. It is therefore significant at the 10-% level. The results for Rothbarth's approach are similar (p=0.062). Table 1.4 shows the λ-parameters of the QAI demand system. Most coefficients are highly statistically significant, except housing, health, and expenditures on education. In Figure 1.1, nonparametric regression estimates of log income on the share of expenditures allocated to food are displayed, stratified by household type (see the supplementary materials for the other commodity groups). For the food share, the curves are mostly approximately linear, except at lower income levels, which likely explains the results for the QAI demand system. Visually inspecting the rest of the commodity groups, we notice that most cases are well fitted by a quadratic specification, while in a few cases, a nonparametric regression is needed (for example, clothing in families with three children; see Figure 5 in the supplementary materials); but those cases appear to be exceptions.

Figure 1.1: Nonparametric estimation of Engel curves across household types.



Data: German Sample Survey of Income and Expenditure 2013.

Turning to shape invariance, the results mostly indicate that shape invariance seems to hold. Judging from the results shown in Figure 1.1, Engel curves for the food expenditure share are approximately shape invariant. The exception might be families with three children. The parametric test of shape invariance confirms this, as in the comparison between singles and families with three children (Column AACCC in Table 1.5), the interaction term is significant at the 5% level. By contrast, the results of the semiparametric tests of shape invariance generally do not reject shape invariance (See Table 1.5). With the loss function of Pendakur (1999), the correlation coefficients are larger than they are with the loss function suggested by Stengos et al. (2006). Neither is low enough to lead us to reject shape invariance.

The outcomes of the nonparametric test, displayed in the last row of Table 1.5, indicate that income independence likely does not hold, even though shape variance is not rejected. This means that the different tests do not give a consistent picture. In the literature, tests of shape invariance have also led to mixed results, depending on the type of test, the expenditure category, and the household type (see Banks et al., 1997; Stengos et al., 2006; Pendakur, 1999). On the other hand, the rejection of income independence is consistent with earlier findings (Koulovatianos et al., 2005; Biewen and Juhasz, 2017). A potential explanation for this finding is that income independence only holds for middle and high incomes; while at low income levels, equivalence scales are income-dependent, as suggested by Figure 1.1. Irrespective of why this might be the case, the results presented here make it hard to judge the approaches exclusively by their assumption; except for the approaches that assume linearity of Engel curves. It thus appears that the plausibility criteria laid out in section 2 and applied below are crucial when attempting to decide between the non-linear methods.

1.5.2 Equivalence scale estimates

Equivalence scale estimates for all methods are presented in Table 1.6. More specifically, using the household of a single adult (A) as the reference, estimates are shown for childless couples (AA), couples with one child (AAC), couples with two children (AACC), and couples with three children (AACCC). Below the point estimates and in brackets, we show 95%-confidence intervals based on bootstrapping (see section 4.3). Unless it is otherwise stated, it may be assumed that we rely on these intervals when discussing similarities between the methods. In addition, we calculated the equivalence scale elasticity, which is defined through $S = h^{\alpha}$, where S is the equivalence scale value, h is household size, and α is elasticity (Buhmann et al., 1988). Generally, α lies between zero and one, with a value of zero implying that additional household members do not generate any additional costs, and a value of one implying that there are no economies of scale. Scale elasticity might hide some more subtle differences between equivalence scales, but it allows for a simple comparison across methods. Here, we discuss these

Table 1.5: Parametric and semiparametric tests of shape invariance and income independence.

A vs.	AA	AAC	AACC	AACCC
Parametric (p-value)	0.453	0.954	0.810	0.021
Semiparametric* (p-value)	0.366	0.364	0.219	0.405
Semiparametric (correlation coefficient)				
Pendakur (1999)	0.963	0.911	0.967	0.736
Stengos et al. (2006)	0.798	0.821	0.723	0.736
Nonparametric covariance test (covariance interval)	[-866; -14]	[-859; -105]	[-966; -113]	[-795; -230]

Note: For the parametric and semiparametric tests, the Null hypothesis indicates that Engel curves are shape invariant or income-independent; If the nonparametric covariance interval does not include zero, which is the value of the covariance that implies income independence, then income independence can be rejected; A correlation coefficient close to unity can be regarded as necessary but not sufficient condition of shape invariance (Stengos et al., 2006).

* based on the objective function of Stengos et al. (2006).

Data: German Sample Survey of Income and Expenditure 2013.

more nuanced differences, while also presenting a broad overview based on elasticities. The last rows of Table 1.6 display the expert scales often used by researchers: namely, the modified OECD scale and the square root scale. In the last column, we show which plausibility criteria – as discussed in section 2.3 – the respective equivalence scale point estimates violate. The modified OECD scale and the square root scale do not violate any of these criteria. For additional comparisons, Table 1.7 shows examples of equivalence scale estimates based on older waves of the EVS taken from the literature. For the methods that have not yet been applied to the EVS, Table 1.8 shows equivalence scales for different countries and datasets.

Compared to the other approaches we applied , the single-equation approach by Engel yields the highest scale values. The economies of scale are small for the second adult (A to AA), and are non-existent for children. The equivalence scale elasticity is around 0.94, which is close to the estimate reported by Merz and Faik (1995) based on an older version of the EVS (see Table 1.7). A possible explanation for these high scale values was provided by Deaton and Muellbauer (1986), who argued that using expenditures on food, as Engel's approach does, overestimates the costs of raising children. The reasoning is that most expenditures related to children will be expenditures on food; thus, even if after the birth of a child the consumption of the parents remains the same, the share of the household's expenditures on food will increase. Thus, keeping the relative expenditures on food constant, as Engel's approach does, will lead to overcompensation. In addition, the results of this approach are questionable, given that the linearity of Engel curves is rejected, and it is not consistent with most of the plausibility criteria.

For the Rothbarth approach, which replaces the food share in Engel's approach with expenditures on an adult good, we use the household of two adults without a child as a reference. This approach is not suitable for estimating the equivalence scale value of a childless couple relative to that of the household of a single adult. The Rothbarth approach results in scale values that are considerably lower than those of the Engel approach, and its scale elasticity is rather small, especially compared to that of all of the other approaches. For instance, according to the Rothbarth estimates, a couple with one child needs roughly 30% more income to be as well-off as a childless couple; while according to the Engel approach the estimated additional income needed is around 50% (calculated as 2.66 divided by 1.72). This observation is in line with Deaton and Muellbauer (1986), who argued that the Rothbarth approach underestimates the costs of having children, and should therefore lead to lower equivalence scale values. Deaton and Muellbauer (1986) also reported findings based on the Rothbarth approach that are close to our estimates, although they based their analysis on data for Sri Lanka. However, as in the Engel approach, linearity of Engel curves is a questionable assumption. Moreover,

Table 1.6: Income-independent equivalence scales.

Approach	A	AA	AAC	AACC	AACCC	Elasticity*	Plausibility criteria
Engel's approach							
Engel (food)	1.00	1.72	2.66	3.63	4.67	0.94	(1.3), (1.4), (1.5)
		(1.65–1.80)	(2.53–2.77)	(3.43–3.83)	(4.33–5.01)		
Rothbarth (adult goods)		1.00	1.29	1.23	1.22	0.14	(1.2), (1.4)
			(1.17–1.40)	(1.12–1.34)	(1.01–1.42)		
Expenditure systems							
Extended linear	1.00	1.71	1.79	1.99	2.24	0.52	(1.4)
		(1.68–1.74)	(1.75–1.83)	(1.95 - 2.03)	(2.16–2.31)		
Quadratic	1.00	1.90	2.03	1.97	2.17	0.53	(1.2), (1.4)
		—	—	—	—		
Demand systems							
Almost ideal	1.00	1.20	1.25	1.31	1.39	0.21	(1.4)
		(1.16–1.24)	(1.20–1.30)	(1.26–1.37)	(1.29–1.49)		
Quadratic almost ideal	1.00	1.58	1.76	2.11	2.26	0.52	(1.4)
		(1.51–1.66)	(1.65–1.86)	(1.97–2.25)	(2.05–2.47)		

Table 1.6: continued

Approach	A	AA	AAC	AACC	AACCC	Elasticity*	Plausibility criteria
Semiparametric							
Original loss function	1.00	1.22 (1.16–1.27)	2.40 (2.34–2.45)	2.10 (2.05–2.15)	1.16 (1.11–1.21)	0.40	(1.2), (1.3), (1.4), (1.5)
Modified loss function	1.00	1.76 (1.75–1.77)	1.62 (1.60–1.64)	1.89 (1.86–1.91)	2.15 (2.13–2.17)	0.48	(1.2), (1.4)
Counterfactual							
Matching	1.00	1.66 (1.64–1.68)	1.90 (1.87–1.93)	2.12 (2.08–2.16)	2.38 (2.29–2.47)	0.55	(1.4)
Partial identification	1.00	[1.54;1.72] (1.54–1.72)	[1.67;1.84] (1.67–1.84)	[1.90;2.08] (1.89–2.09)	[1.76;1.88] (1.75–1.89)	0.45**	(1.2), (1.4)
Expert scales							
Square root scale	1.00	1.41	1.73	2.00	2.23	0.50	none
Modified OECD scale	1.00	1.50	1.80	2.10	2.40	0.54	none

Note: bootstrapped 95% confidence intervals in parentheses; for the quadratic almost ideal demand system, standard errors are analytic; for the quadratic expenditure system, confidence intervals have not been calculated.

* This is the average of the k household comparisons $\alpha_k = \frac{lnS_k}{lnn_k}$ where $k = \{A, AA; A, AAC; A, AACC; A, AACCC\}$.

** Denotes the elasticity of the interval means.

Data: German Sample Survey of Income and Expenditure 2003, 2008, 2013.

Table 1.7: Equivalence scales estimated with the German Sample Survey of Income and Expenditure (EVS)

Approach	Source	Period	A	AA	AAC	AACC	AACCC	Elasticity
Engel (food)	Merz and Faik (1995)[a]	1983	1.00	1.97	2.37	2.85	3.41	0.77
Linear expenditure system	Scheffter (1991)[b]	1983	1.00	1.35	1.68	1.84	2.03	0.44
Extendended linear	Faik (2011)[c]	2003	1.00	1.65	1.78	1.92	2.13	0.49
Quadractic expenditures system	Kohn and Missong (2003)[d]	1988-93	1.00	1.73	2.05	1.94	2.28	0.53
Semiparametric (mod)	Wilke (2006)[e,*]	1998	1.00	1.48	1.56	1.74	1.95	0.44

p.435, Table 2; basic food, married couples 18-64 years, share specification.

p.118, only married couples, unscaled results.

p.310, Table 2.

p.442, Table 9; for AAC and AACC households, the scales are given depending on the age of the child (see Table 8). Using the average of respective numbers here.

p.794-796, Table 8-10; Nadaraya-Watson Estimator, employed sample.

Because different reference household were used some of the values were not directly reported in the table, but were recalculated.

Table 1.8: Equivalence scales based on expenditure data for other datasets and countries.

Approach	Source	Country & Period	A	AA	AAC	AACC	AACCC	Elasticity
Rothbarth	Deaton & Muellbauer (1986)[a]	Sri Lanka 1969		1.00	1.12	1.21		0.11
QAI demand system	Balli and Tiezzi (2010)[b]	Italy 97-04	1.00	1.20	1.31	1.94		0.39
QAI demand system	Michelini (2001)[c]	New Zealand 94-95	1.00	1.53	1.98	2.35	2.66	0.61
QAI demand system	Blacklow et al. (2010)[d]	Australia 88-03	1.00	1.36	1.53	1.69		0.39
Semiparametric	Pendakur (1999)[e]	Canada 1990	1.00	1.97	2.39	2.75		0.76
Semiparametric (mod)	Stengos et al. (2006)[f]	Canada 1996	1.00	1.65	1.90	2.31	2.94	0.64
Matching	Szulc (2009)[g]	Poland 2003	1.00	1.66	1.77	2.40		0.61

p.736, Table 2.

p.770, Table 5, Using the sample mean.

p.391, Table 4; Column EPS-Q(α, β, ϕ_h) The author mentioned that this is the preferred structure. It is a quadratic AI demand system with shifted demands. For AAC households, children aged 0-3 are considered.For AACC households children aged 3-8 for the first child and aged 0-3 for the second child are considered. For AACCC households, the corresponding age ranges are 0-3, 3-8, and 8-14 years.

p.175, Table 5, Column PS-QUAIDS.

p.20, Table 3b, Column 2-4.

p.636, Table 4.

p.83, Table 1, Column 2, 1 Match.

the Rothbarth approach is also not consistent with two plausibility criteria, as it leads to equivalence scale values that are not strictly increasing with household size; and the increases in the scale values by household size are not decreasing.[11]

The ELES yields a scale value for couples without children (AA) that is roughly similar to the value reported by the Engel approach. However, for larger households, the scale values of the ELES are lower than those of the Engel approach, and are closer to the square root scale. The results are very similar to the findings of Faik (2011) based on the EVS 2003. For families with more than one child, our scale values are slightly higher. Compared to the scale values of the ELES, the QES has higher values for smaller households, but lower values for larger households. The equivalence scale elasticity is very similar in both cases, and between the elasticity of the square root scale and the modified OECD scale. In the QES, no confidence intervals are reported, as the estimation procedure did not converge for many of the bootstrap samples, and the inference conditional on convergence could be biased. Apart from this, using the QES might seem more appropriate than using the ELES, as it does not rely on linearity of Engel curves. On the other hand, the QES violates the "household size effect" criterion in equation (1.2), as couples with two children have a lower scale value than couples with one child. That is also the case for the QES estimated by Kohn and Missong (2003) with an older version of the EVS. As this criterion is generally considered essential for equivalence scales, the QES estimates can be seen as implausible.

The AI demand system leads to comparatively low equivalence scale values, and it has a rather low elasticity of 0.2, which indicates that additional household members add very little to the equivalence scales. As is the case for other methods that require linearity of Engel curves, the AI demand system might not lead to reliable estimates because one of its key identifying assumptions is violated. Thus, using the QAI demand system should be more appropriate. Apart from the scale value for couples without children, the estimates of the QAI demand system are between the square root scale and the modified OECD scale, and the confidence intervals of its scale values include the values of both of these expert scales. Correspondingly, its equivalence scale elasticity is also between the elasticities of the expert scales. While the QAI demand system has not been estimated with German data before, estimates for other countries are available (see Table 1.8). Our results fall somewhere in the middle; the estimates reported by Michelini (2001) are generally higher, while Balli and Tiezzi (2010) and Blacklow et al. (2010) reported lower estimates. Like the estimates provided by Balli and Tiezzi (2010), our results violate the plausibility criterion stating that the increase of scale values with household size should become smaller with increasing household size. However, for our

[11] Note that criterion (5) cannot be checked as there is no comparison of A and AA.

estimates as well as for the estimates of Balli and Tiezzi (2010), this violation occurs for households with several children, for which there might be exceptions to this criterion, as we argued in section 1.2.3.

Looking at the semiparametric methods, we can see that the approach by Pendakur (1999) leads to scale values that are rather spread out. For instance, the scale value of 1.2 for couples without children is low, while the scale value of 2.4 for a couple with one child is rather high. While shape invariance cannot be rejected, the estimates violate all four plausibility criteria. This might be due to the deficient loss function. While modifying the loss function according to Stengos et al. (2006) leads to more plausible results, it is still the case that not all criteria are satisfied; e.g., the scale values are not strictly increasing with household size. Compared with the estimates reported by Stengos et al. (2006) for Canada, our estimates are relatively low, and are closer to the results of Wilke (2006) using the EVS of 1998. Equivalence scales reported by Wilke (2006) do not violate the household size effect (See Table 1.7). A possible explanation for this finding, is that in contrast to Pendakur (1999), Stengos et al. (2006), and our application, he used a model based on multiple expenditure categories.

Although it relies on a very different identification strategy, the approach by Szulc (2009) leads to estimates that are close to those of the QAI demand system. For most household types, the confidence intervals for the point estimates of the two methods overlap, and the scale elasticities are also very close. The latter is also the case when compared to the square root scale and the modified OECD scale. Compared to Szulc (2009), who calculated equivalence scales for Poland, we observe that the scale value for couples without children is similar (A to AA), while the scales for the other comparisons are lower. Of the four plausibility criteria, one is violated: the scale value increases by 0.22 for the second child, but by 0.26 for the third child. But as we argued previously this might be realistic. Moreover, this approach does not require linearity of Engel curves, shape invariance, or income independence. The identification bounds provided by the completely nonparametric approach of Dudel (2015) are generally lower than the estimates of the matching method. However, they do not strictly increase with household size (AACC to AACCC), even though the confidence intervals overlap.

To summarize, the approaches that assume linearity of Engel curves (Engel, Rothbarth, ELES, AI), and the semiparametric approach by Pendakur (1999) and its variant (Stengos et al., 2006) are either based on identifying assumptions that can be rejected, or they contradict one or several of the plausibility criteria. The matching approach by Szulc (2009) and the QAI demand system (Banks et al., 1997) violate only criterion (1.4) for which exceptions seem realistic, especially for households with several children. The nonparametric approach by Dudel (2015) might violate criterion (1.2), and, thus, criterion

(1.4), although this violation is not statistically significant based on a comparison of confidence intervals.

These finding indicate that, overall, there is no approach that does not violate at least one of the plausibility criteria. The approaches that are shown to have fewer or less serious violations are either based on the counterfactual framework and do not require strong identifying assumptions (matching, nonparametric); or make use of all expenditure categories, and thus more data than most methods, combined with a flexible specification (QAI demand system). At the same time, applying the QAI demand system to different institutional contexts (Germany, Italy, Australia, New Zealand) can also lead to different results (Table 8). Studies using the same dataset and methods, but for different periods, have found roughly similar equivalence scales elasticities, compared to our results (Table 7). Finally, when using the more plausible equivalence scales to calculate common inequality and poverty indicators, we find that the resulting measures are very similar to each other (see supplementary materials).

1.6 Conclusion

In this paper, we compared 10 different empirical approaches for the estimation of equivalence scales, covering parametric, semiparametric, and fully nonparametric methods. Applying these approaches to German expenditure data from the Sample Survey of Income and Expenditure (waves 2003, 2008, 2013), we found that only a subset of methods produce plausible equivalence scales. These plausible equivalence scales are, however, similar to each other when applying them in the calculation of inequality and poverty indices. Our findings regarding income-independence are somewhat mixed, but indicate that income-independent scales might be appropriate for many questions, especially when studying all income levels. If, on the other hand, the focus is on low or high incomes, then income-independent scales might not be a good choice.

While we covered several very different approaches, our conclusions are restricted to a limited set of methods only; and many methods have been proposed in the literature that we were not able to include here. For example, the approach suggested by Pendakur and Sperlich (2010) was not applied, as it requires long time-series of price variation. While the EVS dates back to 1962, there have been a number of structural breaks in the collection of the expenditure data that would complicate the analysis. Moreover, specifications other than the Working-Leser specification have been proposed, some of which make the resulting equivalence scales income-dependent (Donaldson and Pendakur, 2004). Another potential restriction of our findings is their validity for other contexts; while our findings are promising, we cannot be certain that applying the methods to other countries and datasets would produce consistent sets of equivalence scales.

For researchers applying single exact equivalence scales, using the modified OECD scale can be seen as a reasonable choice if an income-independent scale is desired, at least for Germany. Our results further suggest that the square root scale should be used in estimates for large families.

Chapter 2

Pension level benchmarks: Empirical estimation and results for the United States and Germany[1]

2.1 Introduction

In many high-income countries the population is aging, and the share of the population aged 65 or older is expected to increase substantially (United Nations, 2015). For instance, the U.S. Census Bureau predicts that the share of the population aged 65 or older will increase from 15% in 2014 to 24% in 2060 (Colby and Ortman, 2015). Population aging puts a strain on public finances, as spending on pensions increases (Attanasio et al., 2007). In response to this, the U.S. Social Security retirement age has been increased from 65 to 66 for the cohorts born between 1943 and 1954, and it will increase further for the cohorts born later (Behagel and Blau, 2012). In Germany, the statutory retirement age is also increasing (OECD, 2015). These pension reforms lead to increasing importance of individual retirement savings, in particular in Germany, where previously the public pension was the most important source of income for retirees (Börsch-Supan and Wilke, 2004).

In this context of population aging and pension reform, concerns have been raised about the financial security of retirees. One major concern is whether the pension incomes they receive are adequate. Here, pension income is defined to capture income from all sources, including public pensions, occupational pensions, and income drawn from individual savings. One indicator that is commonly used to measure pension adequacy is the replacement rate, which is defined as post-retirement income relative to pre-retirement income (Boskin and Shoven, 1984). For example, a replacement rate of 80% would imply that an individual's post-retirement income is equivalent to 80% of her pre-retirement income. Surprisingly, no clear benchmark exists for assessing what level of the replacement rate can be considered adequate. In the literature, a wide range of values of between roughly 60% and 100% can be found (Love et al., 2008).

[1] This is a version before revise and resubmit at *Fiscal Studies*

In this paper, we present an empirically-driven approach to derive the replacement rate an individual would need to maintain the living standard achieved by the end of working life. Specifically, our approach is based on keeping the individual's welfare level constant shortly before and after retirement. Pre- and post-retirement income as used in the calculation of the replacement rate are adjusted to account for pension savings and wealth. We propose using this replacement rate as a benchmark for pension adequacy, and call it the adequate replacement rate. We discuss econometric identification of the population average of the adequate replacement rate, and we show several methods for estimating it. These methods are applied to data from the U.S. and Germany. Heterogeneity in replacement rates and potential limitations of our approach are examined through sensitivity checks.

The adequate replacement rate is an attractive measure of pension adequacy for several reasons. First, our benchmark is easily interpretable, as a constant living standard is easy to understand. This makes it a useful policy indicator, and it can provide orientation for policies and individual savings decisions. Second, and related to the first point, the pension systems in some countries have been expected to provide a constant living standard, and policy debates in these countries are sometimes still framed with this goal in mind (e.g., Wilke, 2014). Third, while the conceptual framework we present is rather general, it is easy to apply, and it has modest data demands. Fourth, while other methods have been used to estimate replacement rates, they do not imply adequacy (e.g., Crawford and O'Dea, 2012).

From a methodological perspective, we present a general identification strategy for determining adequate replacement rates based on the potential outcomes framework (e.g., Imbens, 2004), and that is compatible with any indicator of welfare. We focus on indicators that have been used in the life-cycle literature (e.g. Battistin et al., 2009) and in the equivalence scales literature (e.g. Biewen and Juhasz, 2017), and use expenditure data as well as subjective measures of welfare. We apply parametric, semiparametric, and nonparametric methods for estimation. The parametric method and the semiparametric method were originally devised to estimate equivalence scales (Deaton and Muellbauer, 1986; Pendakur, 1999), while the nonparametric method is based on recent results on partial identification by Fan et al. (2017).

Applying all of the procedures we discuss in this paper provides us with easily interpretable results from the simpler approaches, while enabling us to test the assumptions and the robustness of these results using the more sophisticated methods. A requirement of the parametric estimation approach and the semiparametric estimation approach is that the replacement rates do not depend on the pre-retirement income. It is thus assumed that an individual with a low income during working life needs the same replacement rate as an individual with a high income to maintain a constant living standard. The nonparametric approach is not based on this assumption, and yields set estimates only; i.e., partially identified estimates, as described in Manski (2003). In our analyses, we use a parametric test and a semiparametric test drawn from the literature on equivalence scales to assess income independence, and we propose and apply a simple nonparametric test. We also conduct additional checks to assess how sensitive our results are with respect to endogeneity and other potential issues. The outcomes of these checks indicate that our main findings are rather robust.

We study two countries: the United States and Germany. For the U.S., our analysis is based on data from the Health and Retirement Study for 2014. For Germany, we use the most recent wave of the Income and Expenditure Survey (Einkommens- und Verbrauchsstichprobe), which was conducted in 2013. The U.S. and Germany have very different pension systems. The German retirement system is usually considered to be the archetype of the Bismarckian model, as it relies heavily on social security contributions. Until recently, the role of private savings in the German system was small. By contrast, the U.S. retirement system is a Beveridge system based on taxes and private savings play a large role. While we do not expect that the adequate replacement rates differ by country, finding the same replacement rate would have different implications in each of these two countries, as the absolute pension levels differ between the U.S. and Germany. Thus, even if the two countries had the same adequate replacement rate, there could be a gap between the actual and the required pension levels in one country, but not in the other.

In summary, we contribute to the literature in several ways. First, in this paper we provide a blueprint for the empirical estimation of pension standards by establishing a conceptual framework for the estimation of adequate replacement rates, and by discussing and comparing estimation approaches with different levels of econometric sophistication and different underlying assumptions. Second, we investigate the question of whether adequate replacement rates depend on pre-retirement income, and we apply econometric tests, including a new nonparametric test. Third, we provide comparable benchmarks for pension adequacy in the U.S. and Germany. Fourth, our benchmark can be used to help individuals make informed decisions about saving for retirement, which are becoming more important as longevity increases and public pension benefits decline. In addition, our benchmark can help policy-makers assess the well-being of the retiree population, and can help pension providers and insurance companies ensure that individuals have access to the pension plans they need. Fifth, together with this paper, we provide functions for the statistical software R that readily implement our methods, making them easily applicable to other data sets and countries.

The remainder of this paper is structured as follows. In section 2.2, we discuss the related literature, focusing on how adequacy of replacement rates is determined and what levels of replacement rates have been found empirically. Our economic framework is described in section 2.3, and our identification strategy is explained in section 2.4. The data we use is described in section 2.5. We present our main findings in section 3.5, and additional findings and sensitivity checks in section 2.7. Section 2.8 concludes.

2.2 Related work

Intuitively, it might appear that a replacement rate of 100% would allow a retiree to maintain a constant living standard, at least if the time shortly before and after retirement is considered. It may therefore be assumed that if having a certain income level enabled an individual to achieve a certain living standard before retirement, then having the same income should be sufficient to maintain this living standard after retiring.

In the literature, however, several reasons for why a replacement rate of 100% may be either above or below the adequate level have been put forward. On the one hand, values below 100% may be considered adequate given that retired individuals have no work-related expenses (e.g., commuting), are unlikely to have children living in the household, no longer have to save for retirement, and have more time for household production (Aguiar and Hurst, 2005; Love et al., 2008). On the other hand, replacement rate values above 100% may be required because retired individuals could find that their health-related expenses are increasing with age, and that precautionary saving is therefore necessary (Blundell et al., 2016). In addition, because retirees have more free time, they may wish to spend more money on leisure activities (Crawford and O'Dea, 2012).

Taxes might also play a role. To what extent differences in the taxation levels of retirees and non-retirees affect replacement rates depends on whether gross replacement rates or net replacement rates are considered; i.e., whether replacement rates are based on gross income before and after retirement, or on net income before and after retirement. If gross replacement rates are considered, it may be argued that replacement rates below 100% are adequate, given that retirees are usually taxed at lower rates than income earners. If net replacement rates are considered, the differences in the taxation levels of retirees and non-retirees should not matter.

In this paper, we will focus on the estimation of net replacement rates, as these rates are more readily comparable across countries. In the literature, both gross and net replacement rates can be found. In our supplementary materials, we also supply estimates of gross replacement rates to make it easier to compare our results with those of other studies.

In the literature on pension savings and incomes, heuristic benchmarks for pension adequacy are often used. For instance, Haveman et al. (2007) assumed for the U.S. that a net replacement rate of 70% is adequate, while Schulz and Carrin (1972) used a value of 80%. According to Love et al. (2008), replacement rates of between 70% and 100% are common. Similar values can be found in the literature for Germany. In many of these studies, the authors do not justify the chosen value, other than by declaring that the replacement rate is in the established range. In research on Germany, the authors sometimes justify using a net replacement rate value of 70% by arguing that 70% is the highest value that has ever been provided by the German public pension system (Schnabel, 2003).

A data-based approach can be used to derive the minimum replacement rate needed to avoid living in poverty (e.g., Love et al., 2008). While this approach does not establish an adequate replacement rate, it sets a lower bound. To do so, a poverty threshold is calculated. This can, for instance, be the threshold suggested by the OECD, which is calculated as 50% of the median equivalized disposable income (Knoef et al., 2016). Based on this threshold, it is possible to calculate the replacement rate required to avoid living in poverty. A related approach was suggested by VanDerhei and Copeland (2010), who used expenditure data to determine the minimum income needed to reach a certain expenditure level. Again, only a lower bound for pension adequacy was established.

A theoretically grounded approach is based on the life-cycle model, which was introduced by Modigliani and Brumberg (1954) and Friedman (1957). The life-cycle model assumes that the marginal utility of consumption is smoothed over the life course, and that – at least in simple variants – consumption itself is also smoothed. This implies that for an individual to maintain a consistent consumption level, pre-retirement income and post-retirement income should not differ too much, except perhaps after taking changes in work-related expenses or taxation levels into account (Wolfson, 2011). A similar reasoning without recourse to the life-cycle model was proposed by Henle (1972), who argued that equal levels of disposable pre-retirement income and post-retirement income are needed.

While much of the literature on life-cycle models focuses on optimal saving behavior, life-cycle model estimates also imply replacement rates. For the U.S., these rates have often been found to be between 80% and 90% (Hamermesh, 1984; Bernheim, 1992; Mitchell and Moore, 1998), but other values also have been reported. For instance, results by Scholz et al. (2006) imply a replacement rate of around 66%. These results showed that the life-cycle model does not necessarily imply constant consumption, and thus no replacement rates around 100%.

In contrast to the approaches that rely on expenditure data, Binswanger and Schunk (2012) used empirical data on subjective assessments to estimate pension adequacy. They asked individuals in the U.S. and in the Netherlands about their preferred retirement incomes. More specifically, based on the respondents' current income levels, they presented several pairs of pre-retirement income and post-retirement income levels, each of which represented different retirement saving choices and resulting replacement rates. For example, a respondent was given the choice between having a high disposable pre-retirement income with a low savings level and a correspondingly low retirement income; or a low disposable pre-retirement income and a high post-retirement income and replacement rate. Their results showed that both the American and the Dutch respondents preferred net replacement rates of between 80% and 100%. As many of the surveyed individuals were below retirement age (with a median age of between 51 and 52), the results of Binswanger and Schunk (2012) are partly based on the expectations individuals have about their needs in retirement.

Dudel et al. (2016) also looked at subjective assessments using an approach close to the one presented later in this paper. Based on data on individual satisfaction with household income and applying the equivalence scale framework, they calculated the replacement rate needed to keep an individual's standard of living constant. Using German panel data, they estimated that a replacement rate of between 82% and 90% is adequate.

2.3 Conceptual framework

Our approach is based on keeping constant the welfare of individuals around the time of retirement. It requires us to compare individuals shortly before and after retirement; to assess how much welfare changes due to retirement; and to calculate how much less or more income is needed to compensate for the change in welfare. This is similar in structure to what equivalence scales aim to achieve (Lewbel, 2010). Equivalence scales are used to compare households of different compositions, like couples with children and couples without children, and how their needs differ. Similarly, we estimate how needs change when retiring. In this section, we built on the extensive literature on equivalence scales to formally define adequate replacement rates, and to identify potential issues for their estimation.

To formally define the adequate replacement rate, let $V(\mathbf{z}, y, \mathbf{p})$ be the indirect utility function of an individual with characteristics $\mathbf{z}$, net income y, and facing prices $\mathbf{p}$. Using $V(\cdot)$, we can define an income function $I(\mathbf{z}, u, \mathbf{p}) = \min_y[y|V(\mathbf{z}, y, \mathbf{p}) = u]$, which gives the minimum income an individual with characteristics $\mathbf{z}$ and facing prices $\mathbf{p}$ needs to achieve welfare level u. Moreover, let d be a binary variable capturing whether an individual is retired ($d = 1$) or not ($d = 0$). The vector $\mathbf{z}_d = (d, \mathbf{z})$ consists of all characteristics including retirement status.

Using this notation, we define the replacement rate that keeps the living standard constant as

$$R(\mathbf{z}'_0, \mathbf{z}''_1, u, \mathbf{p}) = I(\mathbf{z}''_1, u, \mathbf{p})/I(\mathbf{z}'_0, u, \mathbf{p}), \tag{2.1}$$

where $\mathbf{z}'_0$ captures the covariate values before retirement; and, similarly, $\mathbf{z}''_1$ includes the values after retirement. Thus, $R(\mathbf{z}'_0, \mathbf{z}''_1, u, \mathbf{p})$ is the income a retired individual needs to attain welfare level u relative to the income a non-retired individual needs to achieve u. Except in terms of their retirement status, we will mostly assume that the retirees and the non-retirees are similar, unless otherwise indicated; i.e., $\mathbf{z}' = \mathbf{z}''$. This is, however, not a requirement of our approach, and we will also present some results for which $\mathbf{z}'$ and $\mathbf{z}''$ differ; e.g., results by age. As we use cross-sectional data, we also assume that prices are fixed and the same for all households; i.e., $R(\mathbf{z}'_0, \mathbf{z}''_1, u, \mathbf{p}) = R(\mathbf{z}'_0, \mathbf{z}''_1, u)$.

In the literature on equivalence scales, the retirement indicator d is replaced with an indicator of household composition. Whereas the literature on equivalence scales often starts from household utility functions (this is also the case for the life-cycle literature; see Attanasio and Weber, 2010), we consider the (indirect) utility functions of individuals. Household utility functions require that strong assumptions are met, and ignore decision-making processes and the allocation of resources within the household (Chiappori, 2016). This is also the case for the elderly, as was shown empirically by Lundberg et al. (2003) and Cherchye et al. (2012). As the data we use is at the household level only, and therefore does not allow us to examine the welfare of each individual in the household, we have chosen to restrict ourselves to studying single-person households. We also conduct sensitivity checks in which two-person households are included in the analysis.

Even after restricting our analysis to single-person households, equation (2.1) is not easily identified (Blundell and Lewbel, 1991). The approach we will follow is related to the reasoning of Engel, as discussed by Deaton and Muellbauer (1986); and on the reasoning outlined in van Praag (1991). Essentially, we assume that an indicator variable is available that measures the welfare level u. We focus on expenditure data and use several different indicators. We conduct robustness checks with respect to the choice of indicator, including subjective measures of welfare (see section 2.5.2). Our approach only requires that the welfare indicator is comparable across individuals in the sense that a specific value u' has the same meaning for all individuals; i.e., that all individuals with a specific value u' have the same welfare level, and that higher (lower) values than u' mean that individuals are better (worse) off.

An important issue that arises in this context is income independence, or independence of base (Lewbel, 1989a). Income independence means that the replacement rate needed to maintain a constant living standard does not depend on income y, and, in turn, on the welfare level u; i.e., $R(\mathbf{z}'_0, \mathbf{z}''_1, u) = R(\mathbf{z}'_0, \mathbf{z}''_1)$. This means, for instance, that the adequate replacement rate is the same for a person with a high income during working-life as it is for a person with a low income during working life. For equivalence scales, independence of base is usually rejected in empirical studies (Donaldson and Pendakur, 2006; Biewen and Juhasz, 2017).

While there are, as yet, no similar empirical results for replacement rates, the findings on expected replacement rates from Binswanger and Schunk (2012) suggest that income independence might be violated. In their survey on preferred replacement rates, they found that individuals with low incomes tend to prefer higher replacement rates than individuals with higher incomes. In the next section, we discuss several tests that can be used to assess income independence, including a new one.

The definition of the adequate replacement rate in equation (2.1) does not include several potentially endogenous variables, most notably savings and leisure; and it ignores that some of the demographic characteristics captured by $\mathbf{z}$ might be endogenous, including the retirement decision as measured by d. We will deal with these challenges in various ways in the estimation step. Savings, or more specifically income generated from savings and annuitized wealth, will be included as part of the income variable, while pension savings pre-retirement will be excluded (see section 2.5.2); to explore the potential effect of differences in leisure before and after retirement, we conduct sensitivity checks using satisfaction-based measures of welfare instead of consumption-based measures, which have been argued to at least partly capture effects of leisure (see section 2.5.2); and the potential endogeneity of the retirement decision is tackled using a regression discontinuity design (see section 2.7.3).

2.4 Identification strategy

2.4.1 Basic estimation problem

To discuss the identification and the estimation of replacement rates as defined through equation (2.1), we make use of the potential outcomes framework (e.g., Imbens, 2004). As above, let d denote an indicator variable that captures whether individuals are retired or not. Let W_d denote the welfare level, measured through a welfare indicator based on, for example, expenditure data or subjective measures. $Y_d(w)$ is the income individuals need to achieve welfare level w given their retirement status. In the following, we assume that observed income is equal to $Y_d(W_d)$, and is thus equal to the income function $I(\mathbf{z}_d, u, \mathbf{p})$ defined in the previous section. As we mentioned in section 2.2, we will assume that income is the net income after taxes and transfers, but all of the calculations described here work with gross income as well.

Each individual is observed as being either retired or not retired, and never as being both at the same time. Thus, we observe $(W_0, Y_0(W_0))$ for the non-retired, and $(W_1, Y_1(W_1))$ for the retired. As it turns out, this is not enough to estimate equation (2.1) without strong assumptions. Essentially, replacement rates as defined by equation (2.1) are given by $Y_1(W_0)/Y_0(W_0)$. Taking expectations we have

$$\mathrm{E}\left[\mathrm{E}\left(\frac{Y_1(W_0)}{Y_0(W_0)}\middle|\,\mathbf{z}\right)\right], \tag{2.2}$$

where $\mathbf{z}$ is a vector of covariates, assuming that the covariate values are the same before and after retirement. With some simple algebra, equation (2.2) can be shown to equal

$$\mathrm{E}\left[\frac{E(Y_1(W_0|\mathbf{z})}{E(Y_0(W_0|\mathbf{z})}\right] - \mathrm{E}\left[\frac{1}{\mathrm{E}(Y_0(W_0|\mathbf{z})}\mathrm{Cov}\left(\frac{Y_1(W_0)}{Y_0(W_0)}, Y_0(W_0)\middle|\,\mathbf{z}\right)\right]. \tag{2.3}$$

As we explain in more detail below, the first term in this equation can be identified using standard assumptions, while the second term cannot. Essentially, the first term only requires the marginal distributions of Y_1 and Y_0 to be identified while the second term requires the joint distribution of Y_1 and Y_0 (Abbring and Heckman, 2007).

Assuming that the pair (Y_0, Y_1) is independent from d conditional on W and $\mathbf{z}$, the first term in equation (2.3) can be identified. This is the so-called unconfoundedness assumption (Imbens, 2004). Essentially, unconfoundedness means that there is no unobserved selection into retirement. This allows differences between retirees and non-retirees, as long as they can be captured by $\mathbf{z}$; estimation in case of violation of the unconfoundedness assumption is discussed in section 2.7.3. In addition to the unconfoundedness assumption a further assumption is needed. The overlap condition requires that $0 < \Pr(d = 1|\mathbf{z}) < 1$. Somewhat simplified, this means that for all values of $\mathbf{z}$, there are both retirees and non-retirees. Moreover, the stable unit treatment value assignment is invoked, which implies that whether one individual is or is not retired does not depend on the retirement status of other individuals.

The second term in equation (2.3) is not identified using these assumptions (Abbring and Heckman, 2007). Specifically, this is because of the covariance of $Y_1(W_0)/Y_0(W_0)$ and $Y_0(W_0)$, which requires knowledge of the joint distribution of Y_1 and Y_0. This covariance term captures to what extent the replacement rate depends on the income before retirement. It thus captures the income independence of the adequate replacement rate: the replacement rate does not depend on the baseline income if the covariance is zero, but otherwise it does. If the covariance is negative, then higher baseline incomes are accompanied by lower replacement rates; and if the covariance is positive, then higher incomes are accompanied by higher replacement rates.

In the following, we outline three different approaches for estimating equation (2.3): a parametric approach and a semiparametric approach based on the equivalence scale literature, and a nonparametric method. Each of these approaches go along with different tests to empirically assess income independence, and they differ in terms of their complexity and underlying assumptions. Applying all of the estimation procedures and tests we discuss enables us to provide easily interpretable results for the simpler approaches, while testing the assumptions and the robustness of these results using the more sophisticated approaches.

We discuss a fully parametric approach that can be implemented very easily via standard linear regression, but that requires the strong assumption that the relationship between income and the welfare indicator is log-linear. It also assumes income independence. While the semiparametric approach we borrow from Pendakur (1999) and Stengos et al. (2006) does not require log-linearity, it still requires income independence. The third approach we study is based on a general approach to identification developed by Fan et al. (2017), which we apply to the estimation problem presented here. While it requires neither log-linearity nor income independence, it yields only identification bounds on replacement rates, and no point estimates.

For all of these approaches and tests, we provide functions for the statistical software R (R Core Team, 2017) that make using our framework easy (see the supplementary materials). For the implementation of the semiparametric and nonparametric approaches, we use the `np` package provided by Hayfield and Racine (2008).

2.4.2 Parametric approach

A classical approach for the estimation of equivalence scales attributed to Engel can be adapted to our estimation problem as follows (Deaton and Muellbauer, 1986). W is considered to be a function of income Y, retirement status d, and some other covariates $\mathbf{z}$; i.e., $W(Y, d, \mathbf{z})$. This Engel curve can be estimated empirically, allowing us to use its inverse, $W^{-1}(w, d, \mathbf{z})$. Given a welfare level w', the adequate replacement rate can then be calculated as $W^{-1}(w', 1, \mathbf{z})/W^{-1}(w', 0, \mathbf{z})$.

A common implementation of the Engel approach builds on the model specification proposed by Working (1943) and Leser (1963),

$$W_d = a + \log Y b_Y + db_d + \mathbf{z}'\mathbf{b}_z + \epsilon, \tag{2.4}$$

where a, b_Y, b_d, and $\mathbf{b}_z$ are regression coefficients, and ϵ is a well-behaved error term. Given parameter estimates, which can easily be calculated using least squares, consider equation (2.4) for a retiree and a non-retiree who are similar with respect to $\mathbf{z}$; assume that they have the same welfare level; and equate both variants of the equation and solve for Y_1/Y_0. This yields

$$\mathrm{E}\left(\frac{Y_1}{Y_0}\right) = \exp\left(-\frac{\hat{b}_d}{\hat{b}_Y}\right). \tag{2.5}$$

Thus, in a sense, the regression of Y on W is 'reversed', and this is used as W^{-1}. From the perspective of the potential outcome framework, this amounts to imputing Y_1 and Y_0 through the parameter estimates. Note that (2.5) does not depend on Y_0 or on W, and it implies that the covariance term in equation (2.3) is zero.

The fully parametric framework of equation (2.4) allows us to assess income independence by adding nonlinear terms. For instance, a quadratic term of log income can be added. As was discussed for equivalence scales by Lancaster and Ray (1998), this yields a formula for the income ratio that is more complicated than equation (2.5), and that depends on the baseline income. If the quadratic term is statistically significant, income independence can be rejected. We use this as a test for income independence.

2.4.3 Semiparametric approach

The log-linear functional form of the Working-Leser model might not hold empirically, depending on the welfare indicator used (Banks et al. 1997; for Germany see Garbuszus 2018). Pendakur (1999) proposed a semiparametric approach for the estimation of equivalence scales that does not require strong assumptions about the functional form of the relationship between the log income and the welfare indicator, except for some smoothness restrictions. The welfare indicator is assumed to be a nonlinear function of income, which depends on retirement status: $W(\log Y - d\alpha, d) + d\mu$, where α is the adequate replacement rate, and μ is an additional elasticity parameter. Note that no other covariates $\mathbf{z}$ are included. Given a welfare level w' and inverting W, this setup leads to $W^{-1}(w', 1)/W^{-1}(w', 0) = \alpha$, irrespective of the level of w'.

Practically, $W(\log Y, 0)$ and $W(\log Y, 1)$ are estimated separately using nonparametric regression techniques. Specifically, as suggested by Pendakur (1999), we use kernel regression, with

$$W(\log Y, D) = \frac{\sum\limits_{i:d_i=D} K_h(\log Y - \log y_i) w_i}{\sum\limits_{i:d_i=D} K_h(\log Y - \log y_i)},$$

where K_h is a kernel function with bandwidth h. The kernel function $K(\cdot)$ gives observations with values of y_i close to Y a high weight, while values of y_i far from Y have a low weight. Thus, the estimate of W at Y is based not only on observations with $y_i = Y$, but also on observations close to Y. In a next step, we follow Stengos et al. (2006) and estimate α and μ such that

$$L(\alpha, \mu) = \frac{1}{n} \sum_{i=1}^{n} \Big(\hat{W}(\log y_i - (1-d)\alpha, 1) + \mu - \hat{W}(\log y_i + d\alpha, 0) \Big)^2 \qquad (2.6)$$

is minimized, where n is the sample size. Put simply, the nonparametric estimates $\hat{W}(\log Y, 0)$ and $\hat{W}(\log Y, 1)$ are shifted by α and μ in such a way that they are as close as possible. See the supplementary materials for details of our implementation.

Rephrasing the approach in terms of the potential outcomes framework, for each individual i we observe $y_i = y_i^0(1-d) + y_i^1 d$ and $w_i = w_i^0(1-d) + w_i^1 d$. If $d_i = 0$, y_i^1 is imputed as $y_i^0 / \exp(\alpha)$. If $d_i = 1$, then $\hat{y}_i^0 = y_i^1 \exp(\alpha)$. Given estimates of $E(W_1(Y))$ and $E(W_0(Y))$, imputed values of y imply values for w. Equation (2.6) means that α is chosen so that the imputed welfare level, $\hat{w}_i$, is as close as possible to the observed welfare level; i.e., $\hat{w}_i \approx w_i$.

The semiparametric approach requires that the covariance between the replacement rate and baseline income is zero. It further requires shape invariance of the Engel curves. Shape invariance means that the relationship between the log income and the welfare indicator has the same shape for both retirees and non-retirees, except for shifts of the curve by α and μ (Pendakur, 1999). Shape invariance is thus more general than the assumption of log-linearity employed for the parametric approach. It is usually seen as a sufficient condition for income independence, although in some rare cases this rule might be violated (Lewbel, 2010).

The testing of shape invariance with the semiparametric approach proceeds using simulations based on a comparison of the value of $L(\alpha, \mu)$ that is calculated from empirical data to simulated values that are generated assuming shape invariance (see the supplementary materials for implementation). The distribution arising from the simulated values is used to assess the probability of the empirically observed value given shape invariance. If this probability is below the conventional 5%, threshold shape invariance is rejected.

2.4.4 Nonparametric approach

The parametric approach and the semiparametric approach rely on the strong assumption that the covariance term in equation (2.3) is zero; i.e., that the adequate replacement rate is income independent. They also require shape invariance or linearity of Engel curves.

Dropping the assumptions of income independence, shape invariance, and linearity, we use a nonparametric approach, building on recent results by Fan et al. (2017). This approach does not point identify the replacement rate, and gives set estimates in the sense of Manski (2003); i.e., it can only be established that the replacement rate is in a specific closed interval for which the endpoints can be estimated from the data.

Fan et al. (2017) built on work by Cambanis et al. (1976) on Frechét-Hoeffding bounds to show how bounds on expectations of the form $\mathrm{E}(k(Y_1, Y_0)|\mathbf{X})$ can be derived, where $k(\cdot)$ is a strictly subadditive function. The lower bound can be calculated as

$$\mathrm{E}^L(k(Y_1, Y_0)|\mathbf{X}) = \int_0^1 k\left(F_1^{-1}(t|\mathbf{X}), F_0^{-1}(t|\mathbf{X})\right) dt, \tag{2.7}$$

and the upper bound is given by

$$\mathrm{E}^U(k(Y_1, Y_0)|\mathbf{X}) = \int_0^1 k\left(F_1^{-1}(t|\mathbf{X}), F_0^{-1}(1-t|\mathbf{X})\right) dt, \tag{2.8}$$

where $F_1^{-1}(u|\mathbf{X})$ and $F_0^{-1}(u|\mathbf{X})$ are the quantile functions of the conditional marginal distributions of Y_1 and Y_0, respectively. These quantile functions can be estimated fully nonparametrically (see the supplementary materials).

Assuming that Y_1 and Y_0 will always be strictly positive, i.e., $Y_1 > 0$ and $Y_0 > 0$, $k(Y_1, Y_0) = Y_1/Y_0$ is strictly subadditive, and the approach of Fan et al. (2017) allows us to calculate bounds for $\mathrm{E}(Y_1/Y_0|\mathbf{X})$. This can be used to derive bounds on the covariance term in equation (2.3), which is bounded by

$$\begin{aligned}\mathrm{E}(Y_1|\mathbf{X}) - \mathrm{E}^U(Y_1/Y_0|\mathbf{X})\mathrm{E}(Y_0|\mathbf{X}) \leq& \\ &\mathrm{Cov}(Y_1/Y_0, Y_0|\mathbf{X}) \leq \mathrm{E}(Y_1|\mathbf{X}) - \mathrm{E}^L(Y_1/Y_0|\mathbf{X})\mathrm{E}(Y_0|\mathbf{X}),\end{aligned} \tag{2.9}$$

where $\mathrm{E}^U(Y_1/Y_0|\mathbf{X})$ is the upper bound for $\mathrm{E}(Y_1/Y_0|\mathbf{X})$ and $\mathrm{E}^L(Y_1/Y_0|\mathbf{X})$ is the lower bound (also see Dudel, 2015). As it follows from Fan et al. (2017) that the bounds on $\mathrm{E}(Y_1/Y_0|\mathbf{X})$ are sharp, it also follows that the bounds on the covariance given by equation (2.9) are also sharp.

The bounds of the covariance can be used to test income independence. Assuming that income independence holds, the identification bounds of the covariance as given by equation (2.9) include zero. We propose using this to test income independence. If the confidence interval of the identification region of the covariance includes zero, then the data is consistent with income independence, and income independence cannot be rejected. Specifically, we construct confidence intervals that cover the entire identification interval with a fixed probability; roughly in line with the approach used inHorowitz and Manski (2000), albeit with with some modifications (see the supplementary materials). We implement this approach using the bootstrap, and we choose the smallest interval that covers the identification intervals of 95% of the bootstrap resamples.

2.5 Data

2.5.1 Data sets: HRS and EVS

We use two different data sets for our analysis. For the U.S., we employ data of the Health and Retirement Study (HRS) from 2014, while for Germany we use data of the Income and Expenditure Survey (Einkommens- und Verbrauchsstichprobe; EVS) from 2013.

The HRS is a panel study focusing on Americans aged 50 or older that has been running since 1992 (Juster and Suzman, 1995). It is conducted by the Survey Research Center of the Institute for Social Research of the University of Michigan, and is supported by the National Institute on Aging (NIA) and the Social Security Administration (SSA). The HRS covers a broad range of questions, including questions on employment, pensions, housing, and assets. Respondents are interviewed every two years. As our identification strategy does not rely on longitudinal data, we focus on data of the 2014 wave.

The EVS is a cross-sectional household survey conducted every five years by the German Federal Statistical Office. The last available wave was conducted in 2013. For each household, detailed information on the household's income, expenditures, and savings is collected for one quarter of the year. In addition, socio-demographic variables are included, like the educational attainment of household members.

For our analysis, we restricted both the HRS sample and the EVS sample in two ways. The first restriction is with respect to household size, as we include only single-person households. This was done to avoid the need to introduce household utility functions (see section 2.3). Moreover, the resulting samples are more homogeneous, as it is, for instance, likely that the bequest motives of single adults differ from those of couples (De Nardi et al., 2010; Haan and Prowse, 2014). Moreover, it is not always clear whether households of couples should count as households of retirees, as one partner might be retired while the other is not, which would complicate the analysis (Moreau and

Stancanelli, 2015). Finally, we also excluded individuals living alone who reported being married, but the number of such cases was very small number.

The second restriction is of the age range: we restricted our samples to respondents aged 60 to 69. Restricting the samples to these ages allows us to compare non-retirees who were relatively close to retirement and retirees who had retired relatively recently; as for most of the EVS respondents in this age range the statutory retirement age was 65 or slightly older, while for the HRS 2014 respondents in this age range the lowest retirement age was 62 and the full retirement age was around 66. This restriction helps to guarantee that for all of the ages under consideration there are both retirees and non-retirees in the samples, as is required for identification in the overlap assumption (see section 2.4). If the samples had covered older ages, they might have consisted mainly of retirees or very few non-retirees; whereas if the samples had covered younger ages, they might have consisted mainly of non-retirees.

For the HRS, these restrictions leave us with 878 observations for which the main variables are available, and 798 observations for which we have complete information on the main variables and all control variables (see below). For the EVS, there are 2310 observations with complete information on the main variables and the control variables.

2.5.2 Main variables: Welfare indicator, income, and retirement status

Three ingredients are required in applying our approach: a welfare indicator, income, and retirement status.

In the literature on equivalence scales and in the literature on life cycle models, several different welfare indicators have been used, all with strengths and weaknesses. There is no consensus which indicator is the most appropriate. Because of this, we use not one but several different welfare indicators. To keep the discussion of results straightforward, we present findings based on the income share of expenditures for food at home in our main results in section 3.5, and discuss results based on other expenditure indicators and subjective indicators in section 2.7.2 and in the supplementary materials. The results of most other welfare indicators are very similar.

The expenditure share for food at home can be calculated for both the HRS and the EVS data. It has long been used for the estimation of equivalence scales, building on Engel's observation that the share of income devoted to food expenditures decreases as (log) income increases (Deaton and Muellbauer, 1986). Spending on food is commonly used as an indicator of household consumption in the life-cycle literature (e.g., Browning and Crossley, 2001; Smith, 2006). Moreover, this indicator is also available in many other data sources, and not just in the EVS and the HRS. Still, we are aware of the potential issues that can arise when using this indicator (e.g., Aguiar and Hurst, 2005), which we discuss in section 2.7.2 in addition to results for the other welfare indicators.

As the main income concept we use net household income plus annuitized wealth including housing. This means that the replacement rates we calculate are net replacement rates. While the EVS measures net income, the HRS only captures gross income. To calculate net income for the U.S., we make use of the tax simulations provided by RAND, and follow the methodology of Pantoja et al. (2017) and Blundell et al. (2016). For both the HRS and EVS we annuitize non-housing wealth with 2.5% annually, and housing wealth with 1.25% annually (similar to Crawford and O'Dea, 2012). For non-retired individuals, we exclude all components of wealth that are related to pension savings; for Germany, this includes life insurance. For some asset types, such as stocks or cash savings, it is not clear whether they are intended as retirement savings. These assets are always included in wealth.

To remove outliers with respect to income, we drop households with net incomes below zero. For households in the HRS, this can occur because taxes are only simulated. In a next step, the income distribution is trimmed and the top and bottom 2.5% incomes are dropped from our analysis. Among the households in the EVS, incomes below zero are rare, but are possible because of the way the Federal Statistical Office calculates household income. By design, no households with incomes above $12,000$ euros per month are included in the EVS. Thus, there is no need to drop the high-income outliers. As the threshold below which we drop households from our analysis, we use the German welfare benefit level to which all individuals with incomes below this threshold are entitled.

Determining whether an individual is retired is rather straightforward for the EVS sample, as retirement is a comparatively clear-cut transition in Germany. Specifically, we use an indicator readily included in the data that captures the labor force state of the respondent, and whether she was retired. In addition, we make use of information on working hours, and assume that all individuals who were working 20 or more hours per week were not retired. There are, however, very few individuals whose labor force status was retired and who reported working more than 20 hours per week.

For the HRS sample, assigning the retirement status is more complicated, as several definitions of retirement status can be used (e.g., Behagel and Blau, 2012). Our main analysis uses the labor force status provided by RAND, and we conduct several sensitivity checks using varying definitions of the retirement status. The labor force status provided by RAND is based on several variables, including the number of hours worked and whether the respondent considers herself retired. As in case of Germany, we assume that individuals who were working more than 20 hours per week were not retired.

For all three main variables (welfare, income, retirement) both the HRS and EVS offer alternatives. For instance, instead of data on expenditures on food at home, we could use data on nondurable expenditures, which are available in both the HRS and EVS. Another alternative indicator is satisfaction with income, which has been argued to be a broader measure of welfare also influenced by leisure (Dudel et al., 2016), but is available in the HRS only. We have we conducted extensive sensitivity checks using these and other alternative indicator. An overview of these checks is given in section 2.7, while a detailed description is given in the supplementary materials. Moreover, the supplementary materials also include results using gross income.

2.5.3 Other variables

Both the parametric approach and the nonparametric approach allow for the inclusion of covariates, controlling for potential heterogeneity between retirees and non-retirees. For instance, there are obvious differences with respect to age. While there is considerable overlap in this variable between the two groups, the retirees in our samples were, on average, older than the non-retirees (see supplementary materials for descriptive results). In addition to age, we use the following control variables: education; gender; whether the individual was divorced, as this might indicate alimony payments; whether the individual reported owning the home she was living in; and whether the county where the individual was living was in a rural area, a metropolitan area, or an area with a medium population density. In the EVS data, we also include whether the individual was living in eastern or western Germany, and the quarter in which the data were collected. In the HRS data, we have added race/ethnicity with a dummy variable indicating whether an individual was white or non-white. While the non-white category is rather heterogeneous, we cannot break it down further because of our sample size.
A descriptive overview of all of the variables we use can be found in the supplementary materials.

2.6 Main results

Our main results for the U.S. and Germany are displayed in Table C.1, which shows the estimates of the net replacement rate needed to achieve a constant living standard, as well as their standard errors and the tests for shape invariance or income independence. For the nonparametric approach, the identification interval of the replacement rate is given, together with the standard errors of the lower bound and the upper bound of this interval. All of the underlying results (e.g., regression coefficients) are available upon request. For the tests of shape invariance or income independence, values below 0.05 indicate that shape invariance or income independence can be rejected. For the

Table 2.1: Main results for the U.S. and Germany.

USA	NRR	SE	95% CI	p(ESE)	n
Parametric	0.949	0.073	[0.805,1.093]	0.085	798
Semiparametric	1.052	0.118	[0.650,1.160]	0.371	878
Nonparametric	[0.867,1.156]	0.009;0.011	[0.848,1.176]	[-8305, 545]	798
Germany					
Parametric	0.976	0.035	[0.907,1.045]	0.001	2310
Semiparametric	0.996	0.066	[0.850,1.150]	0.371	2310
Nonparametric	[0.887,1.090]	0.004;0.005	[0.878,1.100]	[-1254,18]	2310

Data: HRS 2014 and EVS 2013

nonparametric approach, the 95% confidence interval of the covariance is shown. If it includes zero, then income independence cannot be rejected; if, on the other hand, zero is not in the identification set, then income independence can be rejected. All these results are based on using the share of expenditures for food at home as the welfare indicator; results for alternative welfare indicators are mostly similar and discussed in section 2.7.2 and the supplementary materials.

The parametric point estimate for Germany is 98%. This means that the net retirement income needs to be at roughly the same level as the net pre-retirement income to avoid changes in the living standard. The semiparametric point estimate is only slightly higher. The parametric estimate for the U.S. is about 95%, and is thus relatively similar to the parametric estimate for Germany; whereas the semiparametric estimate for the U.S. is 10 percentage points higher. However, the confidence intervals overlap. The standard errors are generally higher for the U.S. data than for the German data because of the smaller sample size. In the U.S. data, there are several observations with missing values for the control variables. This explains why the number of observations is lower in the parametric approach (and in the nonparametric approach) than in the semiparametric approach, which does not include these control variables. In the German data, all of the observations that fulfill the inclusion criteria are complete. Thus, the sample size is the same for all three approaches.

The identification intervals of the nonparametric approach are rather wide: the difference between the upper and the lower bound is 33 percentage points for the U.S. and is 22 percentage points for Germany. The lower bounds of the identification intervals are informative in the sense that they allow us to rule out some of the heuristic values found in the literature, like, for instance, the net replacement rate of 70% that is often quoted for Germany in the literature. The results of the parametric method and the semiparametric approach lie within the identification bounds, except for the results of the semiparametric approach applied to German data, which are slightly above these bounds. For both countries, we see a clear pattern in which the bounds of the nonparametric approach have the smallest standard errors the semiparametric approach has the largest standard errors. The latter finding might be attributable to the fact that the semiparametric approach does not include control variables.

The results regarding income independence are mixed. For both the U.S. and Germany, the semiparametric test does not reject shape invariance; and the nonparametric test does not reject income independence. The parametric approach, on the other hand, indicates a rejection of income independence in the case of Germany. For the U.S., income independence can be rejected at the 10% level. Overall, these results overall provide weak evidence for income independence, as the semiparametric test and the nonparametric test rely on assumptions that are not as strong as those of the parametric test, and should therefore be more reliable. Still, heterogeneity in adequate replacement rates might require further study.

In summary, the point estimates for both countries suggest that the adequate replacement rate is around 100%. While the wide identification intervals found when applying the nonparametric approach indicate some degree of uncertainty, they are mostly consistent with the point estimates. The tests of income independence produced mixed results, but overall provided evidence in favor of income independence.

2.7 Robustness checks

2.7.1 Overview

Our main findings presented in the previous section rely on several assumptions. Here, we provide robustness checks of these assumptions. This includes alternative measurements of welfare, and alternative definitions of income and the retirement status; estimates accounting for endogeneity of the transition into retirement; estimates which allow heterogeneity of replacement rates by age; and estimates based on extended samples also including couples. Overall, the findings of these robustness checks are largely consistent with our main findings.

2.7.2 Measurement: Welfare indicator, income concept, and retirement status

To apply our approach requires three variables, as outlined in section 2.5.2: a welfare indicator, income, and retirement status. All of these variables can be defined and measured in different ways, and there is no general consensus which variants are the most

Table 2.2: Results for the U.S. based on different welfare indicators

	NRR	SE	95% CI	p(ESE)	n
Expenditure share food (1)	0.949	0.073	[0.805,1.093]	0.085	798
Expenditure share food (2)	0.883	0.081	[0.724,1.043]	0.007	857
Expenditure share food (3)	0.933	0.169	[0.601,1.264]	0.737	192
Expenditure on nondurables	0.281	0.474	[-0.648,1.210]	0.000	184
Satisfaction with income	0.986	0.334	[0.331, 1.64]	0.000	383

Data: HRS 2014

appropriate. For instance, our main results in the previous section use the expenditure share for food at home as welfare indicator. This specific welfare indicator has been extensively used in the literature, but at the same time it has been criticized as not only being affected by the welfare level, but also time use. In particular, expenditure might be substituted with time for home production, while consumption stays constant (e.g., Aguiar and Hurst, 2007).

To deal with this and similar issues, we have conducted several robustness checks with respect to the welfare indicator, the income concept, and the definition of the retirement status. For the U.S., we use four different income concepts (e.g., net income with and without wealth), five different welfare indicators (e.g., expenditures on durable goods or satisfaction with household income), and four definitions of the retirement status. For a complete overview of the variables, see the supplementary materials. We have run all possible combinations of these variables with both the parametric and the semiparametric approach. As the semiparametric approach shows convergence issues in some cases, we have 147 estimates in total. For Germany, we also use four different income concepts, but only four welfare indicators are available, and two definitions of the retirement status. In combination with the parametric and semiparametric approach, this yields 58 models when six semiparametric estimates are dropped due to convergence issues.

Detailed results are presented in the supplementary materials. Here we provide a few examples in Table 2.2, and a brief general overview below. The findings in Table 2.2 are based on the U.S. data and show the replacement rates resulting from using different welfare indicators, but otherwise applying the parametric approach and defining income and retirement as for our main analysis. The first row of the table shows the same results as presented in the previous section; i.e., using the expenditure share for food at home as an welfare indicator. The second and the third row show alternative definitions of the food share; the fourth row uses the expenditure on nondurable goods; and the fifth row shows results for a subjective measure, the satisfaction with household income.

Note that the variables are taken from different modules of the survey, and that sample size differs considerably because of this. See the supplementary materials for a more detailed discussion of these measures.

All estimates in Table 2.2 are close to the ones presented in Table C.1, and all lie in the identification bounds of the nonparametric approach, except for the results based on expenditure on nondurable goods. However, the point estimate of the replacement rate seems rather low, and is a rather extreme outlier compared to other point estimates. This might partly be due to the small sample size on which this estimate is based. Interestingly, results on income independence are rather mixed and seem to depend on the indicator used.

More generally, for the U.S., roughly 81% of all 147 point estimates arising from the sensitivity check fall within the identification bounds of our main results shown in Table C.1. For Germany, this figure is 72%. The analysis leading to estimates that deviate from our main findings is mostly based on quite different welfare indicators: on satisfaction with household income, which on average leads to lower results than the expenditure-based estimates (the estimate in Table 2.2 being one of the few exceptions); and on nondurable expenditures, that in some cases lead to estimates that are higher than those based on food expenditures.

For equivalence scales, satisfaction-based estimates tend to be lower than expenditure-based scales. This is likely because satisfaction measures are influenced by factors other than consumption, such as a comparison of one's own situation with that of others (Ferrer-i-Carbonell, 2005). Of the nondurable expenditures, expenses for transportation tend to be especially high and strongly dependent on the need to commute. Thus, there is a big drop in these expenses at retirement that is not reflective of a decline in welfare. It therefore appears that using this welfare indicator would overstate the replacement rate needed to maintain a constant standard of living.

2.7.3 Endogeneity: Instrumenting retirement status

For our analysis, it is crucial that the effect of retirement on the welfare indicator is estimated correctly. This requires that the retirement decision must be exogenous, and the unconfoundedness assumption introduced in section 2.4 must hold. This might not be the case, as individuals might select into retirement based on their expectations of their welfare level in retirement, and their incentives to retire or to continue working (Samwick, 1998). For instance, in Germany, retiring before the nominal retirement age, such as at age 63, leads to reduced pension benefits. Individuals who are well off and have substantial savings might not be concerned about this reduction, whereas individuals who have no savings and expect to receive low pension benefits might decide against retiring at a younger age.

A potential solution to this problem is to exploit the exogenous variability in pension eligibility with a (fuzzy) regression discontinuity approach (Imbens and Lemieux, 2008). For instance, in the U.S., the lowest age at which Social Security benefits can be claimed is age 62, and the normal retirement age (NRA) is currently age 66. Both of these thresholds have been set by policy-makers, and are not influenced by individual retirement aspirations. This leads to exogenous discontinuities in the age-specific probability of retirement at the age thresholds. Regression discontinuity designs exploiting pension eligibility have been applied to U.S. and German data (Eibich, 2015; Kämpfen and Maurer, 2016). In the life-cycle literature, this approach has been used to study consumption smoothing (Battistin et al., 2009).

Here, we use linear probability models to predict the probability of retirement conditional on age, which is then used as an instrument for retirement status (Eibich, 2015). We implement this approach using the retirement status, as described in section 2.5. For the U.S., we use data on individuals aged 55 to 75, and exploit the age thresholds of 62 and the NRA, which for some of the cohorts in our data is below age 66. For Germany, we also use observations at ages 55 to 75 to estimate the probability of retirement, as well as discontinuities at ages 60 and 65. While the retirement age in Germany is increasing cohort by cohort, we are not able to exploit the variation introduced through this trend, as in order to do so we would need information on the exact day of birth. Unfortunately, our data provide information on the year of birth only. The German system allows individuals to retire before reaching age 65 without penalties provided they have a minimum number of contribution years, but the availability of this option does not induce a clear discontinuity (Eibich, 2015). More details on the implementation and diagnostic checks of the validity of the regression discontinuity design can be found in the supplementary materials.

The probability of retirement is a quantitative variable that can be easily used in combination with the parametric approach, but not with the semiparametric approach and the nonparametric approach, which require a binary retirement status. For this reason, we can present results for the parametric approach only. Moreover, the treatment effect estimated using a regression discontinuity design is a local treatment effect, which in our case means that it is essentially the effect immediately before and after retirement. Thus, these estimates might be less representative of the pension needs of the retiree population than our main findings.

The results of the second stage of estimation are shown in Table 2.3. For the U.S., the point estimate is considerably lower than our main parametric result (53% vs 95%). However, this estimate is very imprecise, and it is not statistically significantly different

Table 2.3: Results of the regression discontinuity design for the U.S. and Germany

	NRR	SE	p(ESE)	n
USA	0.532	0.449	0.082	787
Germany	0.982	0.056	0.001	1608

Data: HRS 2014 and EVS 2013.

given the relatively small sample size. For Germany, the point estimate of the regression discontinuity approach is very close to the main parametric estimate. For both the U.S. and Germany, the results of the tests of income independence are consistent with our main results. Thus, the overall endogeneity does not seem to pose a threat to our results. The estimation issues for the U.S. that lead to the large standard error of the estimate in Table 2.3 are discussed in the supplementary materials.

2.7.4 Heterogeneity: Adequate replacement rates by age

Our main analysis is restricted to individuals aged 60 to 69. This restriction was motivated by the overlap condition, which requires that individuals of all studied ages are in both the non-retiree group and the retiree group. If we had used, say, ages 60 to 89, then there would have been only a few non-retirees above a certain age.

Still, it has been argued that replacement rates might depend on age (Knoef et al., 2016; Dudel et al., 2016). On the one hand, older individuals might on average be less active and spend less on goods like leisure or transportation, leading to replacement rates declining with age. On the other hand, health expenditures might increase with age and thus the income required to keep the living standard constant. The latter might be more important for the U.S. than Germany, where health insurance coverage is comparatively generous.

Here, we present results on adequate replacement rates by age. We use non-retired individuals aged 60 to 69 as the control group, and compare them to retired individuals aged 70 to 79; otherwise we follow our main analysis. This procedure violates the overlap condition with respect to age, and the results are essentially based on extrapolation. While the findings might give some indication of whether age plays a role in adequate replacement rates, they should be viewed with care.

Estimates are shown in Table 2.4. As the semiparametric approach did not converge using the German data, no results are shown for it (see the supplementary materials for details).

For both the U.S. and Germany, the identification bounds are shifted upwards compared to the main results. However, the identification regions of the main results and the

Table 2.4: Results comparing non-retired individuals aged 60 to 69 with retirees aged 70 to 79; for the U.S. and Germany.

USA	NRR (70-79)	SE	95% CI	p(ESE)	n
Parametric	1.003	0.118	[0.773,1.234]	0.001	1059
Semiparametric	1.054	0.072	[0.990,1.200]	0.414	1172
Nonparametric	[0.939,1.169]	0.010;0.012	[0.920,1.192]	[-7357,141]	1059
Germany					
Parametric	0.920	0.037	[0.849,0.993]	0.001	3302
Semiparametric	—	—	—	—	—
Nonparametric	[0.951,1.221]	0.004;0.005	[0.942,1.160]	[-1188,-16]	3302

Data: EVS 2013 and HRS 2014

results for the older age group overlap, and it cannot be ruled out that the true value is identical. For the U.S., the point estimates are also close to the main findings. For Germany, the parametric point estimate is outside of the identification region, possibly because the income independence assumption is violated (see below).

With respect to income independence, the results are mixed for the U.S., as the parametric test rejects income independence, while the semiparametric test and the nonparametric test do not. The results for Germany are clearer, as both the parametric and the nonparametric approach reject income independence, which might explain why the parametric point estimate is outside of the identification region of the nonparametric approach. These findings imply that for Germany the replacement rates might become income-dependent with increasing age, even though they do not directly depend on age. For the U.S. there is only very weak evidence of age-dependent replacement rates. Further research might be required to gain a better understanding of how income needs change with age.

2.7.5 Sample: Adding couples

As our main findings are restricted to single-person households, couples are missing. However, large shares of the population around retirement age are married or cohabiting. For instance, in the EVS data the majority of individuals aged 60 to 69 were living with a partner. Other household constellations (e.g., parents and children) occur, but are not very common in the age range we are studying.

There are two reasons for this restriction. First, including couples within the framework described in section 2.3 would require us to assume that all functions are related to households; e.g., household utility functions. When using household utility functions, strong assumptions are needed to be meaningful and indicative of the level of welfare of the individuals within the household. Using utility functions for individuals instead requires us to look at the decision-making processes and resource-sharing within households (Chiappori, 2016). Unfortunately, the data we are using either do not allow us to assess the welfare of individuals within households (HRS), or do so to a very limited extent only (EVS). Second, even for single individuals it is not always clear whether they can be counted as retired. In case of couples this is more complicated, as one household member might have retired while the other has not; or one household member might have retired while the other one has started working again after a period of retirement. Nevertheless, one-person households are only a subset of the population in the age range we are studying, and they might be a selected group in one way or another. Couples tend to have different consumption patterns than single people because of economies of scale and economies of scope. Moreover, these two groups might also have different preferences, such as different bequest motives. It is therefore possible that our main findings are not representative of the total population around the retirement age.

Here, we present an additional analysis that also includes households of couples as well as one-person households. For households of non-retirees, we only include couples in which both partners were working, and one of the partners was in the 60-69 age range. For retirees, we include all couples in which both partners were retired and the older one was in the 60-69 age range. This group also includes couples who had been in single-earner households, and who might not perfectly match the dual-earner couples in the control group. But for many of the households in the HRS and for all of the households in the EVS, we are not able to control for the previous labor force status of both partners, as this information is not available.

Results are shown in Table 2.5. These should be viewed with care for the reasons discussed above.

The order of magnitude of the estimates is roughly comparable to our main results shown in Table C.1. For both the U.S. and Germany, the results for the parametric approach are a few percentage points below the main estimates from Table C.1, while for the semiparametric approach, the estimates are somewhat higher. The bounds of the identification intervals are rather similar, but some of the point estimates are outside of the bounds. The results of the tests of income independence are consistent across countries and are comparable to the main results; i.e., the semiparametric tests show evidence of shape invariance, and the nonparametric test does not reject income independence, while the parametric test does. As for the main results, we interpret this finding as being slightly in favor of income independence.

Table 2.5: Results including both single-person households and households of couples with a single-earner, for the U.S. and Germany.

USA	NRR	SE	95% CI	p(ESE)	n
Parametric	0.860	0.040	[0.781,0.938]	0.001	1321
Semiparametric	1.137	0.069	[0.950,1.200]	0.477	1433
Nonparametric	[0.880,1.226]	0.007;0.010	[0.866,1.245]	[-15287,240]	1321
Germany					
Parametric	0.956	0.023	[0.912,1.001]	0.004	4935
Semiparametric	1.112	0.067	[0.930,1.119]	0.390	4935
Nonparametric	[0.902,1.088]	0.002;0.002	[0.898,1.092]	[-1643,4]	4935

Data: EVS 2013 and HRS 2014.

2.8 Conclusions and policy implications

In this paper, we have described a framework for assessing the retirement income individuals need to maintain the standard of living they achieved while working, and we have proposed using this framework as a measure of pension adequacy. We estimated net replacement rates, which are defined as the net retirement income levels individuals need to maintain their living standard relative to their net pre-retirement income, with income adjusted for pensions and wealth. Applying parametric, semiparametric, and nonparametric estimation approaches to U.S. data and German data, we found adequate net replacement rates of about 100%. Thus, our results suggest that to avoid a decline in welfare in retirement, people need to have roughly the same income levels during retirement that they had while working.

We found weak evidence to support the claim that the adequate replacement rate is flat across the whole income range; i.e., that the level of the replacement rate does not depend on pre-retirement income. This finding is based on several tests of income independence, most of which did not reject it. But as some exceptions were found, additional evidence is needed to arrive at a more definite conclusion. We also conducted extensive sensitivity checks, which showed that our main findings are robust to endogeneity, and with respect to the indicators of welfare, the definition of income, and the definition of retirement status used.

The methods presented in this paper can be easily adapted for use with other welfare measures, data sets, and countries; as we have supplied the R code that implements our methods. Thus, our paper can serve as a blueprint for the empirical estimation of pension standards. As our findings show, the parametric approach yields reasonable estimates, is simple to understand, and has low data demands. However, the results for income independence were not clear. This means that the more complex approaches are needed to check the results of the simpler methods. Thus, ideally, the whole range of methods we presented here should be applied.

In the literature, replacement rate values of between roughly 60% and 100% have been deemed adequate, or at least as within a reasonable range. Our results clearly point to the upper part of this range, and are thus comparatively high. Our estimates are compatible with the results presented by Binswanger and Schunk (2012), which indicated what replacement rates individuals consider adequate and would prefer. However, our finding that the adequate replacement rate is high leads us to ask whether the retirement income levels people actually achieve are high enough to reach it.

For the U.S., the pension incomes and savings recent retirees have accumulated are, on average, roughly equal to the level we find in our analysis (Love et al., 2008). Still, as sizable shares of the older U.S. population have no or relatively little retirement savings (Lusardi and Mitchell, 2007), a large group of retirees will likely have pension incomes that are considerably lower than our standard. For Germany, our findings point to a gap in the pensions individuals receive and the pensions they need. According to Kluth and Gasche (2015), recent retirees receive a factual net replacement rate of around 70% from the public pension system, which for many people is the main source of pension income (Bönke et al., 2010). Projections by the OECD indicate that the replacement rate provided by the public system will decrease to around 55% by 2060 (OECD, 2015). Comparing these recent and projected numbers to our main findings yields a gap in pension of around 25% or more, a figure that is expected rise as high as 40% in the future.

Given the projections of the OCED, a replacement rate close to 100% seems hard to reach in the future, especially since pension benefits will need to last longer due to increasing life expectancy. For policy-makers, a replacement rate of 100% might conflict with other policy goals, such as sustainability, and is likely is only possible with certain trade-offs, such as a delay in retirement (Kitao, 2014). For individuals, this might mean that they will have to step up their saving efforts. Especially in the German context, where private retirement savings are currently quite low (Bucher-Koehnen and Lusardi, 2011), it seems unlikely that individuals will reach this level on their own. Providing individuals with better information about the pension benefits they will receive, the pension benefits they will need, and how they can reach their desired level of pension benefits might help them achieve greater financial security in retirement (Bernheim and Garret, 2003; Dolls et al., 2018).

Several other approaches have been used for estimating replacement rates. Ideally, estimates based on our approach can be combined with findings from these other methods, as each of them has different strengths and weaknesses. Our approach guarantees adequacy in the sense of a constant living standard, which the other approaches do not; and it is close to individual preferences as shown by the results of Binswanger and Schunk (2012). However, this does not necessarily mean that the income implied by the adequate replacement rate is above the poverty threshold, which might be seen as a prerequisite for adequacy; if a person is below the poverty threshold before retiring, then a constant living standard means that the person will also be below the threshold after retiring. Calculating the lowest possible replacement rate needed to avoid living in poverty as done by, e.g., Love et al. (2008) avoids this and provides a useful additional benchmark.

The life-cycle model and its extensions have also been used to derive replacement rate benchmarks (e.g., Mitchell and Moore, 1998). The life-cycle model is based on the assumption that individuals smooth the marginal utility of consumption over the life course, and it allows to derive the implications of optimal saving behavior under constraints. Our approach, by contrast, is based on the goal of keeping constant the welfare of individuals around the time of retirement, without any constraints. This could mean that our approach leads to benchmark values which are hard to reach. Thus, life-cycle-based estimates can complement our approach by providing an assessment of what replacement rate can be reached under constraints.

One important caveat of our findings is that the replacement rates we have reported are averages, and we mostly ignore heterogeneity in replacement rates, except in the tests for income independence and the results by age. Income and age might not be the only potential source of heterogeneity. For instance, replacement rates might depend on health and health-related expenditures, as medical expenses have been shown to account for large shares of household consumption in the U.S. (Banks et al., 2019; Finkelstein et al., 2013), and they might depend on further factors like leisure, over which we average. While we provided results by age, further disentangling the heterogeneity in replacement rates is possible with the methods presented in this paper, and is a potential avenue for future study.

Chapter 3

The replacement rate that maintains income satisfaction through retirement: the question of income-dependence [1]

3.1 Introduction

In many aging countries, private savings have become an essential pillar in the retirement income portfolio. The shift from public to private decision-making is accompanied by a lot of uncertainty, and requires guidance. The replacement rate – i.e., the percentage of the end-of-career employment income that is replaced by the retirement income – is a key parameter for pension planning tools based on life-cycle models, in which it represents the decline in income that the individual or the household is willing to accept after retirement (Skinner, 2007; Scholz et al., 2006). But what is a good choice for this parameter? In practice, 70% of net income (Schulz and Carrin, 1972) is often used as a benchmark replacement rate. Dudel et al. (2016) estimated for Germany that 86% of net income is needed. Both approaches assume that one benchmark fits all income levels. The survey literature, on the other hand, suggests that people with different income levels need different replacement rates (Binswanger and Schunk, 2012).

Benchmark replacement rates are also important in public policy. For example, the UK Pension Commission 2004 recommends replacement rates of 80% for annual incomes below PBS 9,500, and 50% for incomes above PBS 40,000. There is no empirical justification for these values except that those are the realized replacement rates of the recent cohorts of retirees. Arguably, these figures do not reflect whether or not the respective households are satisfied with their financial situation after retirement (Crawford and O'Dea, 2012).

In this paper, I estimate a benchmark replacement rate that has an empirical foundation, and that varies with income. I follow Dudel et al. (2016), and estimate a replacement rate that maintains income satisfaction through retirement, but I use a more flexible

[1] This is a version before revise and resubmit at the *Journal of the Economics of Ageing.*

specification that allows the rate to vary with income. The approach is based on the Generalized Absolute Equivalence Scales Exactness framework (GAESE), which was originally applied to derive income-dependent scales that equalize income across families of different compositions (Donaldson and Pendakur, 2006). The common key assumption is that welfare can be directly observed in the data.

Conceptually, the approach is applied as follows: First, I select a sample of households that are followed over a substantial number of t years after the age of 50. Second, I define a number of household situations that vary in terms of retirement status and household size. Third, I measure the relationship between household income and income satisfaction as a direct measure of welfare across the household types. I also choose a reference household; e.g., for single households, employed single individuals. Finally, I estimate a shift and a scaling parameter by Maximum Likelihood that can, in turn, be used to plot the retirement income ratio against income.

As self-reported income satisfaction is the dependent variable of the regression analysis, I apply longitudinal ordered response models, as discussed in Baetschmann et al. (2015), that take into account both the ordinal nature of the dependent variable, and the individual unobserved but time-invariant heterogeneity. The approach does not make strong assumptions regarding the interpersonal comparability of income evaluations, but instead only assumes that individuals are consistent over time. Further, as utility is measured on the individual level, no strong assumptions about intra-family resource allocation are needed (Chiappori, 2016).

I investigate Germany, which has a pay-as-you-go pension system of the Bismarckian variety in which benefits strongly depend on contributions, and retirement incomes are still mainly provided by statutory pensions. I use longitudinal data from the German Socio-Economic Panel (SOEP), which has collected yearly data on subjective income evaluations for nearly 30 years. The sample consists of 114,756 observation-years.

I find that, on average, the income-independent replacement rate that maintains income satisfaction through retirement decreases from 74% for monthly household net incomes around EUR 1000 to around 65% for incomes over EUR 3000. For couple households, the metric ranges from 90% for joint incomes of around EUR 1500 to 75% for incomes over EUR 4000. The income-independence assumption is rejected on the 1%-Level for single households whereas the results for couple households are less precise. The GAESE specification has a better Goodness-of-fit than an income-independent approach.

In the sensitivity analysis, I tested transformations of the dependent variable, additional time-varying covariates, two alternative identification strategies, stratifications along time-invarying variables, and modifications of the age threshold. The results for singles turned out to be fairly consistent across the models.

Overall, the results suggest that households at the bottom level of the income distribution have higher fixed costs to replace when they retire. They also confirm the findings of Binswanger and Schunk (2012) using qualitative data from the US and the Netherlands. From a policy perspective, the results call into question whether constant benefits-to-payments ratios, which has long been applied in Germany, is an approach in which the majority of the popualation is financially satisfied after retiring. Households with lower incomes have to save more proportional to their earnings, and because these households usually have fewer opportunities for wealth accumulation (Bernheim et al., 2001), they have to be supported in their efforts to save.

I contribute to the existing literature by applying a model that allows me, to examine not only check whether replacement rates depend on income levels, but also whether replacement rates increase or decrease with income. The previous studies that were closest to this one were conducted by Dudel and Schmied (2019) and Binswanger and Schunk (2012). In the former study, the authors investigated how much the replacement rate needs to be to maintain consumption levels. They also tested whether it is fair to use one benchmark for all income levels. They applied econometric tests which indicate whether income-independence must be rejected, but not whether replacement rates increase or decrease with income. The authors found mixed results depending on the allowed flexibility of the test. In a different study design, Binswanger and Schunk (2012) asked a sample of pre-retirees from the US and the Netherlands how much money they would need to maintain an adequate standard of living in retirement, given their current income. In both countries, low-income households expressed a desire for a larger fraction of their income, suggesting a decreasing gradient in income. While this study design has many benefits and relies on weak assumptions it is still an ex-ante approach. The question whether individuals desire a lower or a higher replacement rate *after* they retire remains unresolved, as they may change their minds. Thus, an ex-post analysis like to one conducted in this paper is an important complementary study design.

The remainder of this paper is structured as follows. In section 3.2, I review the current literature on replacement rates and corresponding benchmarks. In section 3.3, I describe the econometric framework and the identification strategy I employ. In section 3.4, I describe the dataset I use. I present the main findings in section 3.5 and offer a wide range of robustness tests. In section 3.6, I discuss the results and make some attempts to explain them. 3.7 concludes.

3.2 Related literature

Replacement rates, sometimes referred to as retirement income ratios, are used in various ways and for different purposes. For individuals, replacement rates may be used to provide a projection of their future living standards, as the rates place their projected income after retirement in relation to their income while working. The average replacement rate of a population, such as the population of a country, may be used to assess savings adequacy across cohorts (Smith, 2003; Geyer and Steiner, 2014; Knoef et al., 2016). Similarly, replacement rates are often used to study the effects of policy reforms on retirement incomes (Palmer, 1989), or as indicators of pension adequacy when comparing countries with different pension systems (OECD, 2015). In practice, actuaries and financial advisers, as well as online retirement planners, use replacement rates as benchmarks to set up individual saving plans, or to assess the adequacy of current accumulated wealth (Skinner, 2007). At both the individual level and the population level, the assessment of realized replacement rates against a benchmark replacement rate may be used as a measure of economic well-being in old-age (Dudel and Schmied, 2019). While realized replacement rates can be observed in register data and are usually higher at lower income levels, the question of whether benchmark replacement rates follow the same pattern remains open.

In the first report of the UK Pension Commission (2004), the authors suggested the use of a benchmark replacement rate that decreases with income. They recommended a threshold of 80% for annual incomes below PBS 9,500, and a threshold of 50% for incomes above PBS 40,000. The commission justified the choice of these thresholds by stating that they are the actual replacement rates of current cohorts. However, it is not clear why the realized replacement rates were selected as the benchmark replacement rates, given that the observed households might have failed to meet their financial retirement goals, or could be unsatisfied with their current resources.

Dudel et al. (2016) established a benchmark replacement rate based on an explicit objective, the maintenance of income satisfaction through retirement. Using German SOEP data from 1989-2014, they estimated this rate to be around 90%, and assumed that it applies equally to people with low and high incomes. Dudel and Schmied (2019) tested this assumption with cross-sectional expenditure data and found no clear evidence of income-dependence.

Binswanger and Schunk (2012) uncovered a different pattern using a customized questionnaire, i.e., they asked employed individuals how much of their current income (or a projection thereof) they would need to maintain an adequate standard of living in retirement. In both examined countries, the US and the Netherlands, low-income households expressed a desire for a larger fraction of their income, which suggests a decreasing gradient in income. Interestingly, the ranges in the two countries were very different. In the US, people in the lowest quintile expressed a preference for a rate of around 108%, whereas people in the top quintile expressed a preference for a rate of around 54%. In the Netherlands, the range was much narrower, from 69% to 63%.

Similar inconsistencies between empirical and qualitative evidence have been observed in the equivalence scale literature. Equivalence scales are important metrics for standardizing income between households of different compositions (Deaton and Muellbauer, 1980b). While income-independent scales had been popular for a long period of time, qualitative evidence has since indicated that the income-independence assumption must be rejected (Koulovatianos et al., 2005).[2] Therefore, Donaldson and Pendakur (2006) later proposed a more general framework that allowed equivalence scales to vary with income. At that point, equivalence scales were mostly based on revealed preferences from expenditure systems, and were estimated using detailed expenditure data. In more recent work, Biewen and Juhasz (2017) adjusted the approach so that it is applicable to satisfaction data. Many of the methodological questions discussed in the equivalence scale literature can be also applied to efforts to estimate the replacement rates that maintain income satisfaction (Dudel et al., 2016).

3.3 Methodology

The derivation of the metric I seek is essentially based on the question of how much income a comparison group needs, on average, to achieve the level of income satisfaction of the reference group. It is analogous to the identification of base-dependent equivalence scales, albeit with an additional interpretation: As the reference category is set as a household in which the adult member are in employment and are at the end of their career (defined later), and a comparison household in which the adult members are in retirement, the equivalence scale that makes the retirement income high enough that the income satisfaction while working is maintained, can be interpreted as a replacement rate (Dudel et al., 2016).

Following the notation of Biewen and Juhasz (2017), let u_{it} denote satisfaction with the household income of individual i in year t.[3] u_{it} is modeled by the equivalent income eq_{it}, i.e., the income that maintains the income satisfaction of the reference status; a vector of other time-varying covariates z_{it}' such as age, and ϕ_i that represent household-specific

[2] The assumption was referred to as base independence (IB; Lewbel, 1989b) or equivalence scale exactness (ESE; Blackorby and Donaldson, 1993).

[3] For a formal derivation of replacement rates that maintain welfare levels, including assumptions about the cost function, see Dudel and Schmied (2019).

time-invariant effects – e.g., whether a person is generally optimistic – while ϵ_{it} captures measurement error. eq_{it} is a function of the household type w_{it} – retired or not – and the household income y_{it}. The functional relationship of the equivalent income is specified as the logarithmic income and linear income satisfaction. That approach has become standard in the subjective equivalence income literature (e.g., Schwarze, 2003; De Ree et al., 2013), but has also proven to provide the best fit for the data (see Appendix C.3).

$$u_{it} = \beta_1 log(eq_{it}) + z'_{it}\gamma_k + \phi_i + \epsilon_{it} \tag{3.1}$$

$$eq_{it} = f(hhtype_{it}, y_{it}) \tag{3.2}$$

Base-dependent equivalence scales in a GAESE form allow for a translation and a scale component. For replacement rates, this implies that the specification allows for a fixed component $\alpha(d_{it})$ and a variable component $\rho(d_{it})$ of subjective retirement costs. In GAESE the change of the equivalent income functions is assumed to be constant with respect to the household income given the household type w_{it} (Donaldson and Pendakur, 2006).

$$\frac{\Delta eq_{it}(w_{it}, y_{it})}{\Delta y_{it}} = \rho(w_{it}) \tag{3.3}$$

This implies

$$eq_{it}(w_{it}, y_{it}) = \rho(w_{it})y_{it} + \alpha(w_{it}) \tag{3.4}$$

The replacement rate (or equivalence scale) that is defined by the household income divided by the equivalent income can then be obtained by

$$r(w_{it}, y_{it}) = \frac{y_{it}}{eq_{it}} = \left(\rho(w_{it}) + \frac{\alpha(w_{it})}{y_{it}}\right)^{-1} \tag{3.5}$$

Whether or not the replacement rate varies with income is indicated by the fixed component $\alpha(w_{it})$. If it is negative, the replacement rates are higher for poorer households, and vice versa. If it is zero, the metric is independent of the income level. Therefore, examining the significance of α is an econometric test to check for income independence (Biewen and Juhasz, 2017).

What is ultimately estimated is the following:

$$u_{it} = \beta_1 log\left(\rho(w_{it})y_{it} + \alpha(w_{it})\right) + z'_{it}\gamma + \alpha_i + \epsilon_{it} \tag{3.6}$$

where

$$\alpha(w_{it}) = 0 \times I(w_0 = 0) + a_1 \times I(w_1 = 1) + ... + a_k \times I(w_k = 1) \tag{3.7}$$

$$\rho(w_{it}) = 1 \times I(w_0 = 0) + b_1 \times I(w_1 = 1) + ... + b_k \times I(w_3 = 1) \tag{3.8}$$

where $a_1, a_2, ..., a_k$ and $b_1, b_2, ..., b_k$ are the estimated coefficients with k as the number of household types.

The model specified in (3.1) cannot be estimated linearly. Assuming that the rate is dependent on the reference income, the problem emerges that the reference income depends, in turn, on the replacement rate parameter. Biewen and Juhasz (2013) suggested an iterative estimation to solve that problem. Biewen and Juhasz (2017) later used an estimation strategy, which is commonly known as the blow-up-and-cluster method (BUC, Baetschmann et al., 2015). It is based on the conditional likelihood estimator (CLM, Chamberlain, 1979), which consistently estimates binary logit models while taking into account fixed effects. With the BUC method, the CLM can be applied to ordinal structured dependent variables. To that end, the dependent variable is dichotomized in $k-1$ ways. Income satisfaction ranges from zero to 10 in the SOEP data. The reported levels of income satisfaction are therefore generated by an ordered logit model, such as:

$$P(u_{it} = k | z_{it}, w_{it}) = \Lambda(\tau_k - \lambda log Q(b,a) - z_{it}\beta - \alpha_i) - \Lambda(\tau_{k-1} - \lambda log Q(b,a) - z_{it}\beta - \alpha_i) \tag{3.9}$$

where $Q(b,a)$ represents the replacement rate that sets u before and after d equal; τ_k represents a threshold for individual i who makes a certain subjective evaluation of his/her income. For any dichotomization, the parameters can be consistently estimated by the CML. The BUC method repeats this procedure for all possible combinations, while restricting the estimates to be equal across all dichotomizations.

The estimation strategy makes the following assumptions: utility can be directly observed in the data, here by income satisfaction; income satisfaction can be compared within individuals (discussed in the next section); the true utility replacement rate is non-linear (tested in section 3.5.2); and r is sufficiently controlled for by observables Z_{it} or time-invariant unobserved heterogeneity α_i (discussed in section 3.5.4).

One additional key assumption for equivalence scales that are based on families or multiple-person households is that there is a single utility function for the household/family as a whole even though it is well known that resources are not equally shared among household members (Chiappori, 2016). In this application that assumption is not needed, as in the SOEP data, the utility of all (adult) household members is individually assessed (households with children are not part of the sample).

3.4 Data

3.4.1 Sample

I use data from the German Socio-Economic Panel (SOEP), which includes over 15,000 households and 30,000 individuals in the 34th wave. I use observations until the survey year of 2017. Individuals are followed over time and not across households. Importantly, this implies that for every household with more than one adult, each adult is surveyed independently.

I reduce the sample to household situations in which the respondent is aged 50 or older, based on the fact that replacement rates usually refer to the individuals' earnings in the final years of their career. This threshold is somewhat arbitrary, and I discuss how other thresholds could affect the outcomes in section 3.5.4. I drop households with children living in the household, and households with more than two adults. Finally, I drop households with missing information on key variables. That leaves an unbalanced panel of $14,743$ households with at least two survey years and a median of seven years ranging from 1991 to 2017.

3.4.2 Household situations

In the next step, I construct binary variables to define three single and six couple household situations. Essentially, these households differ according to their labor force status, while the reference household is still employed and the comparison household is retired. The household situations are distinct from each other; i.e., one household cannot be in two or more household situations at the same time. For every couple household, there are two individual observations, but in the analysis, the standard errors are clustered on the household level.

In particular, the reference category a_{emp} denotes an employed single household situation in which the respondent is employed and did not receive any income from pensions in the previous year ($7,959$ observation-years).[4] a_{ret} denotes a single retiree household situation in which the respondent is not employed and received pension benefits in the previous survey year ($22,784$ observation-years). a_{un} denotes an unemployed single household situation in which the respondent is not employed and received no pension benefits in the previous year (709 observation-years). aa_{emp} denotes a double-earner household situation

[4] Dudel et al. (2016) showed that using alternative concepts of retirement definition leads practically to the same results.

in which both the respondent and his/her partner are employed, and neither received no pension benefits in the previous year ($18,691$ observation-years). $aa_{ret;emp}$ denotes a one-earner-one-retired household situation in which the respondent/partner is employed and received no pension benefits in the previous year, while the partner/respondent is retired and non-employed ($9,778$ observations years). $aa_{emp;un}$ denotes a one-earner-one-unemployed household situation in which the respondent/partner is employed while the partner/respondent is unemployed and received no pension benefits in the previous year. aa_{ret} denotes a two-retiree household situation in which neither the respondent nor his/her partner is employed and both received pension benefits in the previous year ($40,691$ observation-years). $aa_{ret;un}$ denotes a one-retired-one-unemployed household situation in which both the respondent and his/her partner are non-employed, and only one of them has received pension benefits in the previous year ($7,338$ observation-years). Finally, aa_{un} denotes a double-unemployed household situation in which both the respondent and his/her partner are non-employed, and neither received pension benefits in the previous year (395 observation-years).

3.4.3 Dependent variable

This approach is based on the assumption that utility is directly observed in the data. To assess the utility of the members of the household, self-reported evaluations of the household's financial situation are used. The SOEP has included such a question since 1989, and thus, almost from the beginning the of the survey. Respondents are asked how satisfied they are with their household income on a scale from zero to 10, with 10 being the most satisfied. Note again that for a couple household there are two different evaluations of the same household income.

The use of subjective evaluations of income has been a popular approach for conducting welfare analysis in general, and for examining equivalence scales and pension adequacy in particular (e.g., Ferrer-i Carbonell and Frijters, 2004; Pradhan and Ravallion, 2000; van Praag and Kapteyn, 1973; Schwarze, 2003; Biewen and Juhasz, 2017; Koulovatianos et al., 2005; Binswanger and Schunk, 2012). While the literature has been less critical of self-reported evaluation of income (or other domains) in the recent years, there are a number of assumptions and potential problems that should be taken into account when using this approach (e.g., Krueger and Schkade, 2008; Bertrand and Mullainathan, 2001; Layard et al., 2008). As I will discuss in the following, most of these reservations can be addressed within the applied study design.

First, misreporting can be systematically correlated with unobserved or observed individual characteristics (Bertrand and Mullainathan, 2001). For example, it may be the case that individuals with higher financial literacy have systematically different assessments of their financial situation than individuals with lower financial literacy. This problem is addressed in this paper by the fixed-effects estimation, which allows for correlations of measurement error with time-invariant characteristics. Problems of measurement errors are accelerated if both the dependent and independent variable are subjective (Ferrer-i Carbonell and Frijters, 2004). Here I use household income which is sufficiently objective and thus this is not a concern in this application.

Second, a number of studies have shown that the design of the questionnaire can affect the outcome variable (see Kahneman and Krueger, 2006, for a literature review). Most importantly, there is evidence that the ordering, the context, and the vagueness of the question can change the results (Bertrand and Mullainathan, 2001). It appears that using a specific domain (household income), rather than general life satisfaction, is less problematic (Krueger and Schkade, 2008). Moreover, it can be expected that respondents beyond retirement age understand the question as it is posed in a straightforward manner (see above). The question is also asked at the very beginning of the questionnaire, and is, unlike in other aging datasets, included in the core study. Finally, the question relates to each household's current financial situation, and not to their past or future situation (Pudney, 2011).

Next, there is the issue of how the variable should be treated in the econometric model, and what assumptions should accompany it. This issue is discussed extensively in Dudel et al. (2016) and in section 3.3. Essentially, the assumption of cardinality implies that an increase in satisfaction of one point is the same, regardless of whether the increase is from two to three or from four to five. In this paper, I apply different types of methods that assume either cardinality or weak ordinality.

3.4.4 Income and other explanatory variables

Household income is a key variable in the analysis. My income variable is based on what the SOEP calls household post-government income, which is reported by either the respondent or the household head. As I examine a long time period, I deflate income with the consumer price index provided by the Federal Statistical Office. Using after-tax income is key to estimating a convincing replacement rate, because the way income is taxed after retirement has changed in Germany. To make the replacement rate as close as possible to the actual income that is generated by the household, all sources of income should be taken into account. An important exception is that of public transfers, such as unemployment benefits, which represent an income component that is not generated by the household itself and does not generate earning points in the German pension system. Thus, public transfers are excluded. In addition to labor earnings, asset flows, private retirement income, private transfers, and social security pensions are used. Labor earnings include wages and salary from all employment, including training income, self-employment income, bonuses, overtime, and profit-sharing income. Asset flows include income from interest, dividends, and rent. Private transfers include payments from individuals outside of the household including alimony and child support payments. Social security pensions include payments from old age, disability, and widowhood pension schemes (Grabka, 2020). The variable does not take imputed rent into account. Homeownership is discussed in section 3.5.4.

Extreme values of household income are excluded by trimming the top and the bottom 1% of the income distribution. I do that for every household situation separately.

In the baseline model, I control for age and period dummies. Note that the applied models do not allow for the inclusion of time-invarying covariates, such as gender, cohort, or education. I examine the sensitivity of the results by including time-varying variables that affect income satisfaction around the time of retirement, such as health in section 3.5.4.

3.5 Results

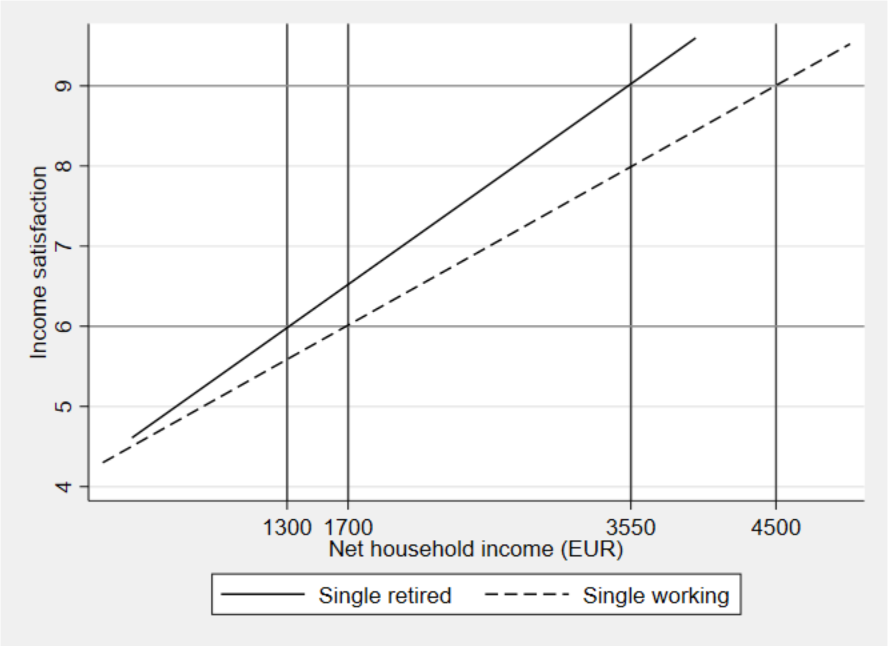
3.5.1 Visual pre-inspection

I begin the analysis with a non-parametric, visual test for income-(in)dependence that is known from the literature on Engel curves (Deaton and Muellbauer, 1986). It also aims to give the reader a better understanding of how the identification works.

Consider Figure 3.1 in which the relationship between income and income satisfaction is compared among single retirees and single workers with two linear regression curves. Income satisfaction is assumed to be cardinal in this analysis. If the curve of the reference household is located on the left-hand side of the comparison curve, the replacement rate is below 100%. At an income satisfaction level of six, the retired individuals need about EUR $1,700 - 1,300 = 400$ to achieve the same satisfaction level as working individuals. The local replacement rate that maintains that satisfaction level is $1,300/1,700 = 76.4\%$. Base-independence implies that the same rate applies for all utility levels. Here, for a satisfaction level of nine, the retiree needs about $3,550/4,500 = 78,9\%$. Hence, within this simplified model, the replacement rate would be not the same for all utility levels, and income dependence may be rejected. However, the question of whether the difference is significant cannot be assessed by this analysis.

In addition, the data are better fitted by a lin-log specification. Thus, in the upcoming graphs, I plot the relationship between the logarithm of household income and levels of income satisfaction, for singles and couples, respectively. The linear regression line,

Figure 3.1: Linear Engel curves and base dependence



Note: own illustration
Data: SOEP 1991-2017

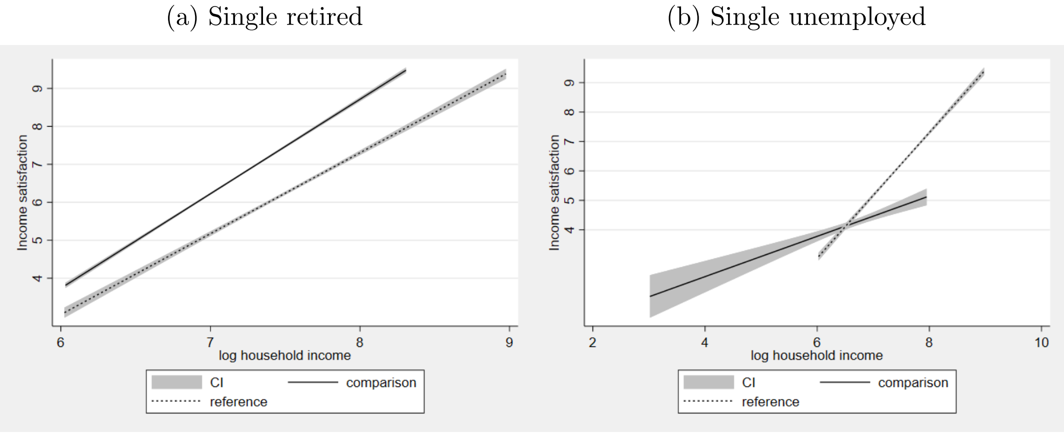
here the dotted line represents the reference household; i.e., a household in employment. While the linear fit is more representative of the applied estimation strategy, I also show a non-parametric fit in Appendix C.1.

The main idea of the analysis is as follows: if the horizontal distance between the curves increases from left to right, the replacement rate that equalizes the utility of the reference situation, moves further away from 100%. In other words, the benchmark replacement rate decreases with income because less income is needed to arrive at the same welfare level as that before retirement.

In Figure 3.2, single households are shown. The reference household is a single employed person. In comparison to the Engel curve of the single retired individual (Panel a), the lines cross after an income of EUR 500 and diverge thereafter, which suggests that at higher incomes the replacement rate is lower. For the unemployed individuals, there is only a small overlap of the income distribution, which is known to cause problems for equivalence scales (Pendakur, 1999). At earnings lower than EUR 1100, the lines converge, indicating that the benchmark replacement rates become higher at higher incomes up to that point.

In Figure 3.3, couple households are shown. The reference household is a double-earner household. In contrast to the curve for the retiree couple (Panel a), the lines appear to be parallel, but it is hard to tell whether that is the case simply by looking at them. A similar pattern is found for the two-retiree couple and for the unemployed individual (Panel c). For the household with one retiree and one employed person, as well as the household with one employed and one unemployed person (Panel b and d), the lines are very close to each other and converge slightly. This indicates that the replacement rates are close to 100% at all incomes, while the benchmark replacement rate increases with income. Finally, for the unemployed couple, the lines converge up to a joint income of EUR 2,400, which suggests increasing benchmarks. Again, the overlap of the income distribution with that of the reference household is quite small.

Figure 3.2: Linear regression lines of household income versus income satisfaction by household type: single households; reference: single working


(a) Single retired

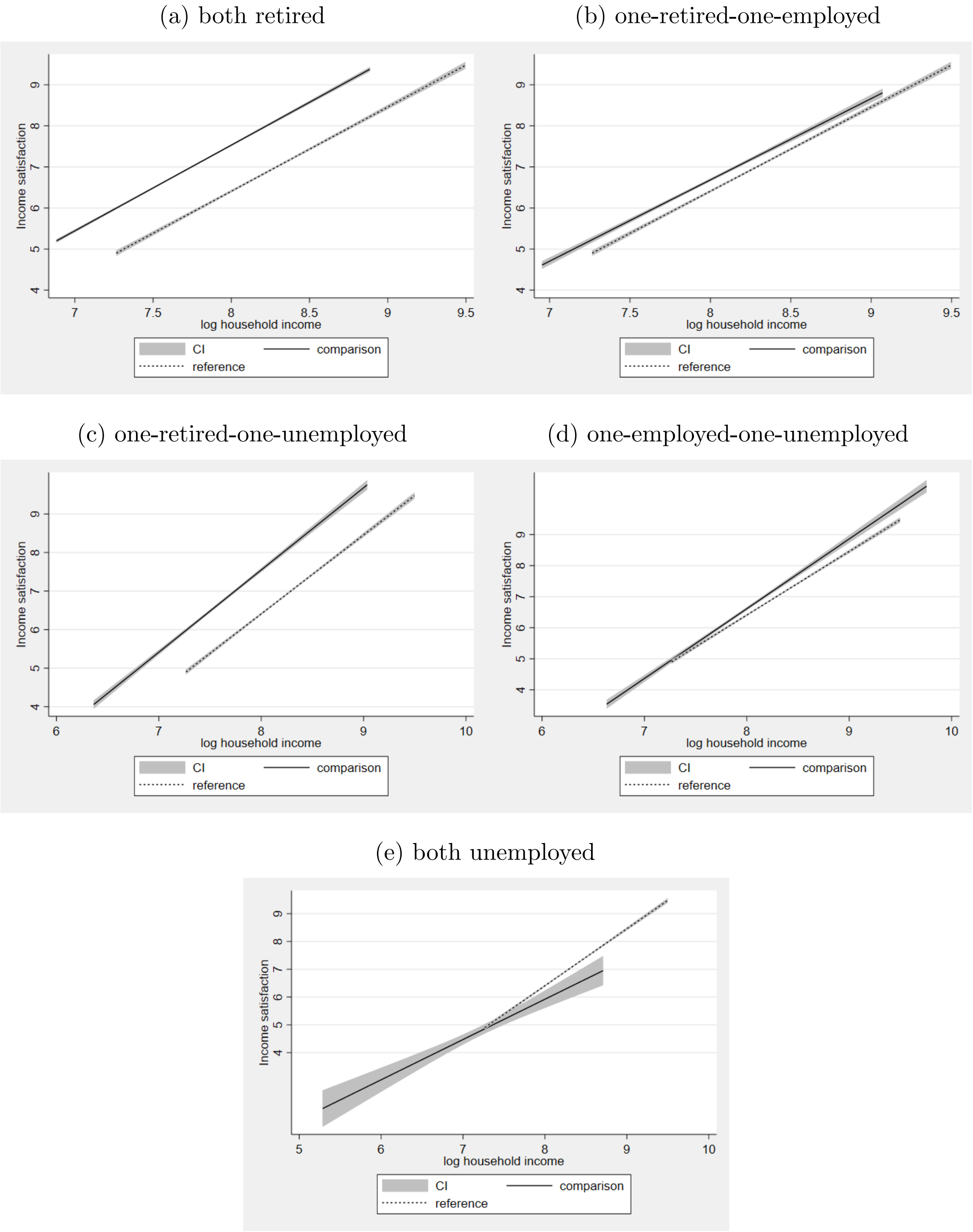
(b) Single unemployed

Data: SOEP 1991-2017

3.5.2 Main results

Table 3.1 shows regression results for equation (3.6) with a single worker as the reference category (coefficients for the main model with a double-earner household as the reference are available upon request). The model estimates the effect of transitioning from a reference household situation ($w = 0$) to several comparison household situations (w_k), on income satisfaction, holding income and age constant, while controlling for time-invariant household heterogeneity. Note again that the model is based on conditional maximum likelihood estimators following the blow-up-and-cluster strategy to identify individual thresholds. α^w indicates the scaling component of the replacement rate representing the fixed subjective costs, i.e. the loss of income satisfaction, of transitioning from the reference household situation to the comparison household situation. Again, if these are negative, the replacement rate decreases with income. If they are non-significant, the replacement rate can be assumed to be income-independent.

Figure 3.3: Linear regression line of household income and income satisfaction by household type: couple households; reference: both employed



(a) both retired

(b) one-retired-one-employed

(c) one-retired-one-unemployed

(d) one-employed-one-unemployed

(e) both unemployed

Data: SOEP 1991-2017

In Table 3.1, the α's for couple households that do not include unemployed partners are negative and significant on the 1% level. The coefficients are similar to the values found by Biewen and Juhasz (2017), e.g., for couples without children as reference.

They again, support the finding from the equivalence scales literature that economies of scale are greater at higher income levels. For couple households with an unemployed partner, α's are positive and partly significant, which indicates a metric that increases with income. When interpreting this metric, it should be taken into account that the unemployed receive a large fraction of their earnings from public transfers, which are excluded from the joint income. Further, as was noted above, the overlaps among the income distributions are small, especially with respect to the unemployed couple.

Most importantly, however, the scaling component of the retired single is negative, and is significant on the 1% level. The metric can be interpreted as the benchmark replacement rate, which, according to this finding, decreases with income for single households.

ρ^w indicates the translation component of the replacement rate, which represents the variable costs of the household transition. In Table 3.6, they are all above 1 except for household situations in which unemployed individuals are involved.

Figure 3.4 and 3.5 show the results of equation 3.5 plotted against household income. [5] A single worker is used as the reference and a dual-earner household is used as the reference, respectively. On the vertical axis, $r(w_{it}, y_{it})$ shows the equivalence scale that maintains income satisfaction from the single working situation to the other eight household situations. As was argued in Section 3.3, the scale that refers to a household situation in retirement can be interpreted as a replacement rate. The solid line in Figure 3.4 shows this metric. Around earnings of EUR 1,000 the benchmark replacement denotes 75%; whereas above earnings of EUR 3,000 the benchmark averages to around 64%.

Dudel et al. (2016) estimated an income-independent benchmark of 87%, which in this framework only applies for incomes below EUR 900. [6]

When a dual-earner household is used as the reference, r declines with income for the retired couple only. With respect to the retired single, the corresponding α is insignificant, resulting in a flat curve and income independence. For the retired couple, there is a curve that decreases with income. It averages to 86% for joint incomes below EUR 1300, and to 78% for incomes beyond EUR 4,000. However, the corresponding α has high standard errors and is not significant along common levels. The other household types increase with income, partly significantly (coefficients are available upon request).

In sum, I find that the estimated benchmark replacement rates decreases with income for both singles and couples, and that the decrease is significant for singles.

[5] I do not plot households with incomes less than the minimum pension income in Germany, which is around EUR 900 for singles and around EUR 1200 for couples. In Germany, when individuals have not contributed enough, their benefits are raised to a minimum pension value. This results naturally in higher replacement rates.

[6] There are some methodological differences in the estimation process.

Table 3.1: Coefficients Model 3.6 ; Single worker as reference category

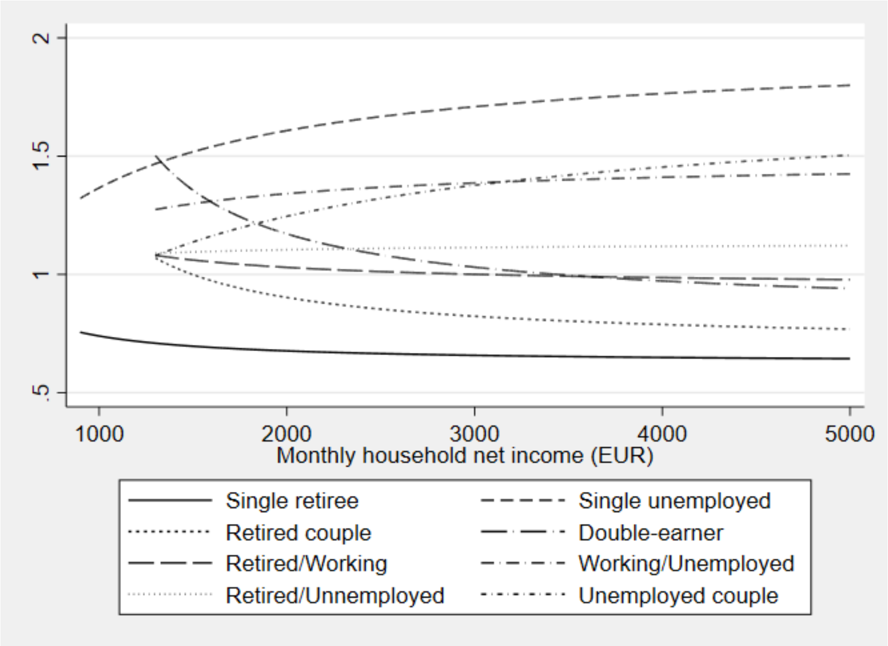
Parameter	Variable	Coefficient	SE (clustered)	P>z
β_1	log(Income)	1.068	0.054	<0.001
$\alpha^{a_{ret}}$	scaling comp retired singles	-252.3	62.7	<0.001
$\alpha^{a_{un}}$	scaling comp unemployed singles	220.6	36.2	<0.001
$\alpha^{aa_{ret}}$	scaling comp both retired	-642.4	198.0	0.001
$\alpha^{aa_{emp}}$	scaling comp both working	-698.8	218.8	0.001
$\alpha^{aa_{ret;emp}}$	scaling comp retired/working	-170.6	174.7	0.329
$\alpha^{aa_{emp;un}}$	scaling comp working/unemployed	146.0	66.1	0.027
$\alpha^{aa_{ret;un}}$	scaling comp retired/unemployed	46.8	96.4	0.627
$\alpha^{aa_{un}}$	scaling comp both unemployed	459.9	100.9	<0.002
$\rho^{a_{ret}}$	translating comp retired singles	1.604	0.163	<0.001
$\rho^{a_{un}}$	translating comp unemployed singles	0.510	0.171	<0.001
$\rho^{aa_{ret}}$	translating comp both retired	1.430	0.128	<0.001
$\rho^{aa_{emp}}$	translating comp both working	1.203	0.126	<0.001
$\rho^{aa_{ret;emp}}$	translating comp retired/working	1.056	0.122	<0.001
$\rho^{aa_{emp;un}}$	translating comp working/unemployed	0.673	0.107	<0.001
$\rho^{aa_{ret;un}}$	translating comp retired/unemployed	0.882	0.130	<0.001
$\rho^{aa_{un}}$	translating comp both unemployed	0.573	0.482	<0.001
γ_1	Age	-0.017	0.005	<0.001

Note: The dependent variable is income satisfaction scaled $0 - 10$. Period effects are included but not shown. Maximum likelihood estimation. Number of cluster: $8,436$. Standard errors are clustered on the household levels.
Data: Socio-Economic Panel 1991-2017.

3.5.3 Heterogeneity

In Germany, despite the current transition to a multi-pillar system, the retirement incomes of current cohorts are largely drawn from statutory pensions (Werding, 2016). Statutory pension benefits are, with a few exceptions, determined by contribution years. In West Germany in particular, the labor market has long been dominated the male breadwinner model. As a results, many German women have numerous gaps in their earning histories, which have resulted not only in personal earning losses, but also in low pensions. Furthermore, the household income, and thus the household income

Figure 3.4: Replacement rates that maintain income satisfaction; working single as reference category



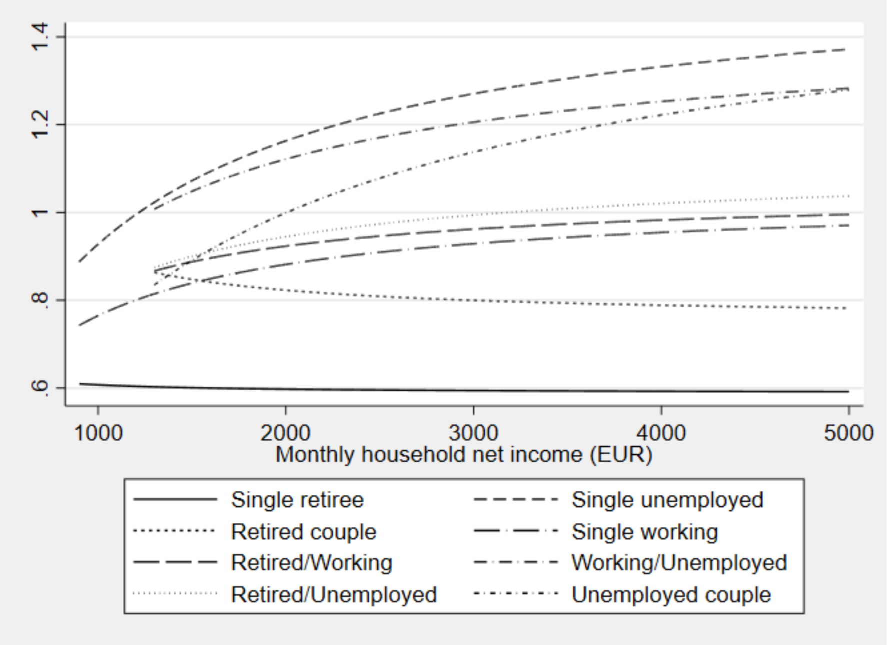
Note: The estimates are based on the coefficients outlined in Table 3.1
Data: German Socio-Economic Panel 1991-2017

satisfaction often depends strongly on the financial situation of the spouse.[7] Thus, in some cases, a woman (including a widow) who had zero labor earnings before retirement may have a high retirement income. This would result in replacement rates well above 100% and might bias the average replacement rate. Therefore, in Figure 3.6 and Table 3.2, I show results based on men only, with single working men as the reference category. The scaling component α increases to -341 leading to a steeper decline in the benchmark replacement rate than in the main result. Here, rates above 3000 EUR average to 57%, which is a little lower than in the main results.

In the former GDR, earnings were generally lower, and while retirement incomes were also lower, current cohorts from the former GDR still realize higher replacement rates. This is partly because contribution years earned in the former GDR have a different financial value than those earned in West Germany (see Kluth and Gasche, 2015, for a stratified analysis of replacement rates from statutory pensions). In Table 3.2 and Figure 3.7, I show results for individuals that were living in West Germany. For the benchmark replacement rate for singles, the results do not change much, but the decline is again a bit steeper (see row 2 in Table 3.2). For couples, the results are also similar

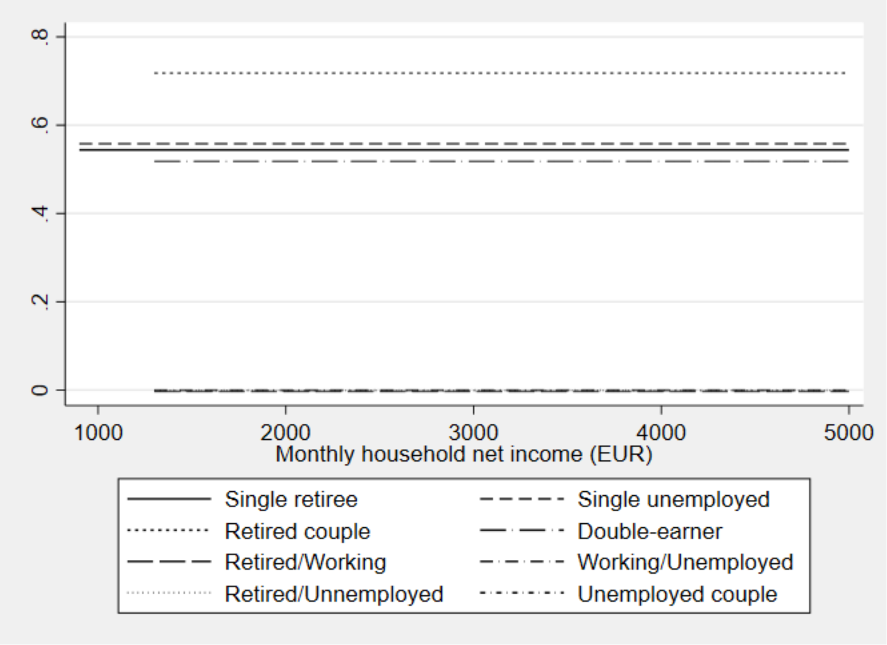
[7] In case of a divorce, earning points acquired by both partners while their marriage lasted, are allocated equally.

Figure 3.5: Replacement rates that maintain income satisfaction; working couple as reference category



Note: The estimates are based on the coefficients outlined in Table 3.1
Data: German Socio-Economic Panel 1991-2017

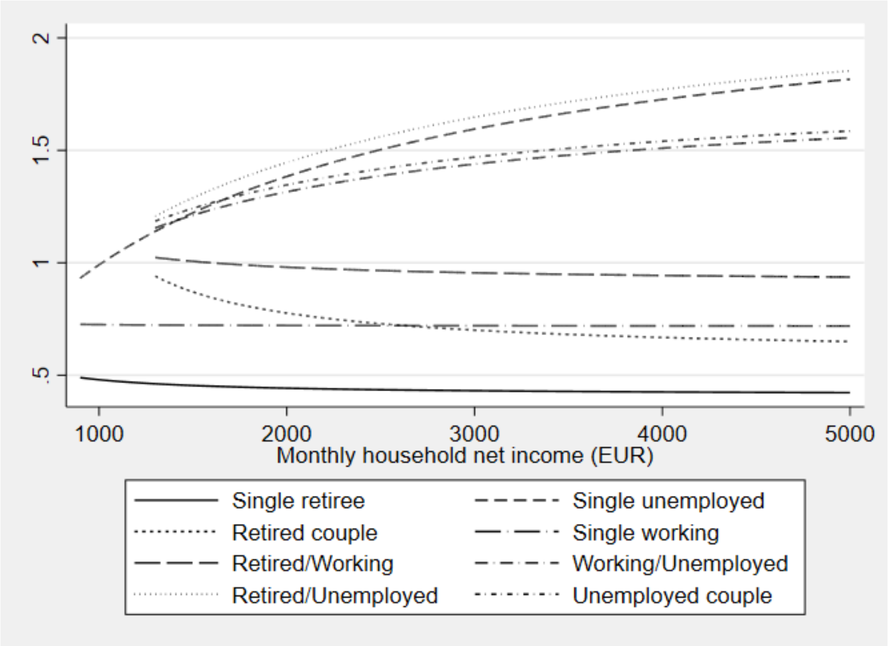
Figure 3.6: Replacement rates that maintain income satisfaction; only men; single worker as reference category



Data: German Socio-Economic Panel 1991-2017

while the standard errors are a bit higher due to the smaller size. Still, the benchmark replacement rate decreases slightly from 85% at a joint income of EUR 1,300 to 79% at an income of EUR 4,000.

Figure 3.7: Replacement rates that maintain income satisfaction; only West Germany; double-earner household as reference category



Data: German Socio-Economic Panel 1991-2017

3.5.4 Robustness

How the model uses information from self-reported income evaluations is discussed in sections 3.3 and 3.4.3. Still, I make the inevitable assumption that a cardinal notion of individual utility is behind all of the subjective indications represented by the data (Dudel et al., 2016). It is therefore reassuring if models with weaker assumptions and/or different settings come to the same conclusions. In the following section, I test how sensitive the benchmark replacement rates are to transformations of the dependent variable, different identification strategies, different sample selections, and the inclusion of other time-varying factors.

First, the dependent variable is transformed into a binary variable in which income satisfaction from 0-6 indicates *unsatisfied* with household income and 7-10 indicates *satisfied* with household income. As shown in Table 3.2, the scaling component becomes much smaller while the standard errors increase. It is, however, still negative, and the loss in the precision may well be explained by the loss of information the dichotomization

Table 3.2: Scaling component α across applied models

	Retired single	Retired couple
main result	-252.3(62.3)***	-211.1(208.0)
only men	-341.3(139.3)**	(j)
West Germany	-302-2(108.1)***	-205.1(262.1)
binary outcome	-73.9(82.5)	(k)
transformed outcome	-361.5(107.0)***	-835.2(447.5)*
control for health changes	-242.2(60.22)***	-136(204.0)
control for changes in homeownership	-251.6(62.9)***	-144.3(210.3)
Age threshold at 60	-172.0(68.8)**	-326.5(249.9)

Note: clustered standard errors in parenthesis $*$ represent significance levels; (j) not meaningful (k) did not converge

requires.

Next, I transform the dependent variable into three categories, with 0-4 indicating *unsatisfied*, 5-7 indicating *neither unsatisfied nor satisfied* and 8-10 indicating *satisfied*. The results are shown in Table 3.2 and in Appendix C.5. For singles, α increases substantially in this sensitivity test while standard remain relatively small. The decline in the benchmark replacement rate from poor to rich is steeper. For couples, α becomes substantially larger and significant on the 10% level. Thus, the benchmark for couple is also steeper in this setting, averaging to 99% for incomes below EUR 1,300 and to 50% for income over EUR 4,000.[8]

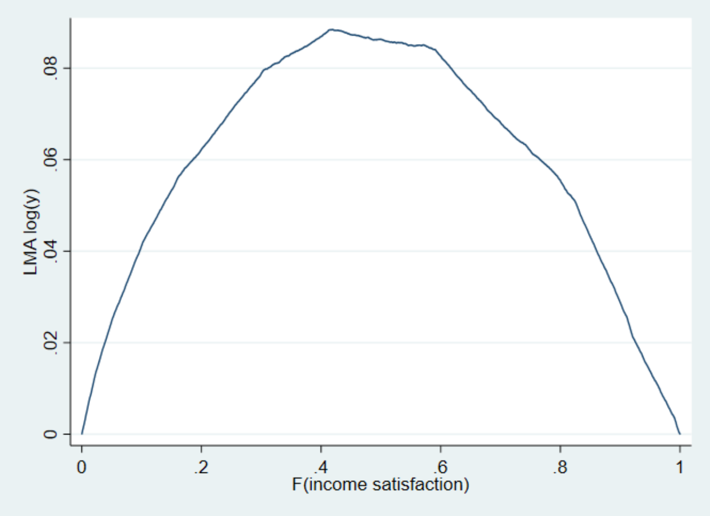
Using the procedure suggested by Schröder and Yitzhaki (2017), I check whether the key coefficients of the model are sensitive to further transformations. In Figure 3.8, the curve shows the difference between the independence of household income and the absolute concentration - which is plotted against the density of income satisfaction. It shows that there are no breaks into negative levels at any point in the distribution. Thus, there is no possibility that a transformation would change the sign of the coefficients. Age is also tested with the same conclusion.

It is known from the literature that control variables have little impact on the subjective equivalence scales (Schwarze, 2003; Biewen and Juhasz, 2017). Moreover, with the fixed-effects design, unobserved time-invarying factors are accounted for. Still, there are observable factors that might change around retirement, and it is important to check whether including them in the model changes the results.

Health is known to change in retirement. While the literature on whether the change is positive or negative is inconclusive, health is an important factor. Thus, I add health

[8] For the income-independent estimates by Dudel et al. (2016), transformation of the dependent variable did not have a strong effect.

Figure 3.8: Line of independence minus absolute concentration curve of household income with respect to income satisfaction



Data: German Socio-Economic Panel 1991-2017

to equation (3.1). I use health satisfaction, which is dichotomized with satisfaction with health being lower than five indicating *bad health*. The coefficient of bad health on income satisfaction is -0.5966, and is thus quite a large effect. However, it does not substantially change the main picture as none of the resulting ρ and α are significantly different from those of the main model (see Table 3.2).

Another important factor that might change in retirement is homeownership. Retirees who own a home have fewer fixed costs than those who rent an apartment or a house. Therefore they might be more satisfied with their income than a tenant because their expenditures are lower. The fixed-effects approach captures the effect when it is constant over the observation period. Some households may change their status when they retire, by, for example, selling their house and becoming a tenant of a smaller apartment. Thus, I add an indicator variable for tenant or owner to the main model. The effect of this variable on income satisfaction is not distinguishable from zero and it also results in estimates similar to those in the main model (see Table 3.2).

In another robustness test, I examine to what extent results are sensitive to the age threshold of the sample. For the main results, I examined adults 50 years or older in order to allow for a reasonable range of observations during their working years. In some cases, the replacement rates refer to the very last years of the respondents' careers (Kluth and Gasche, 2015). In Germany, the normal retirement age is around 66; thus I trim the sample to individuals and households aged 60 years or older. For couples, that means that both household members have to be aged 60 or older.

The bottom row of Table 3.2 shows the results of this modification. For singles, α is slightly smaller than in the main results, but it is still highly significant. The translation component ρ is also a bit smaller, resulting in higher benchmark ratios. In Germany, the average income in the final years before retirement is usually lower because many pre-retirees reduce their working hours. As a result replacement rates tend to be higher when this time span is used as the reference. For couples, α increases, but as the standard errors also increase, the decrease remains insignificant.

Next, I use an alternative identification strategy to derive income-dependent replacement rates, which can be conveniently estimated with cardinal fixed-effects modelling. Thus, the requirements for identification are relaxed. The strategy is based on Lancaster and Ray (1998), who suggested adding a variable that contains the fraction of the household type to the main specification. Hence, (3.1) becomes

$$u_{it} = \beta_1 log(y_{it}) + B_2 w_{it}^k + B_3 \frac{w_{it}^k}{y_{it}} + \beta_3 Z_{it} + \alpha_i + \epsilon_{it} \tag{3.10}$$

where B_2 and B_3 are vectors of k household dummies. By setting the utility of the household situation equal to the utility of the reference household situation, and by solving for the ratio of income the household situation and the reference situation gives

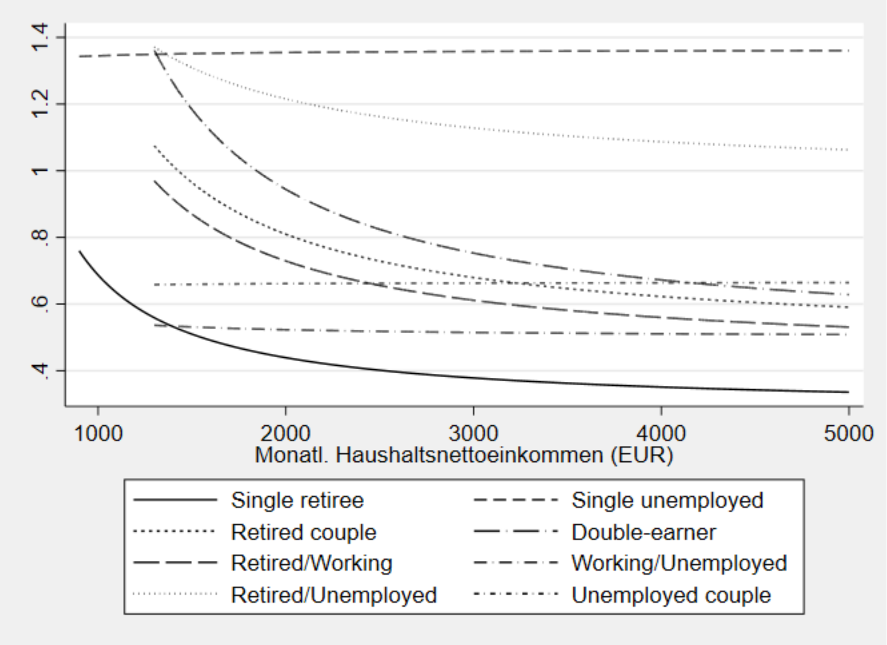
$$R(Z) = \frac{y_{it+c}}{y_{it}} = exp\left(-\frac{B_2^k}{\beta_1} + \frac{-B_3^k \frac{1}{y_{it}}}{\beta_1} + (z_{it} - z_{it+c})\frac{\hat{\gamma}}{\hat{\beta}_1} \right) \tag{3.11}$$

The replacement rate that maintains income satisfaction shown in 3.11 is now explicitly dependent on income (Dudel et al., 2017b). Calculating $R(Z)$ for our samples as described in Section 3.4 and plotted against income, results in the dotted line in Figure 3.9. Coefficients are shown in Appendix C.5. The resulting benchmark replacement rate for singles is about the same for incomes around EUR 1,000 but is lower at higher incomes, e.g., at around 40% for incomes of about EUR 3000. For couples, the benchmark replacement rate decreases from 100% for joint incomes of EUR 1,500 to around 60% for joint incomes of EUR 4,000. Thus, both benchmarks are steeper than in the main results but they are decreasing which provide reassurance that the main results are not driven by the identification strategy.

The identification strategy outlined in section 3.3 nicely supports the idea of maintaining the living standards individuals had during their working years (Dudel et al., 2016), but the estimation procedure is complicated and the implementation is tricky (code is available upon request). Moreover, the computational effort is substantial. To address concerns that the main results are not a statistical artifact from the methodology, I demonstrate an alternative way to identify benchmark replacement rates.

Coming back to the initial question what a good choice of replacement rate in the

Figure 3.9: Replacement rates that maintain income satisfaction using an alternative identification strategy applying cardinal fixed effect modeling; single worker as reference category

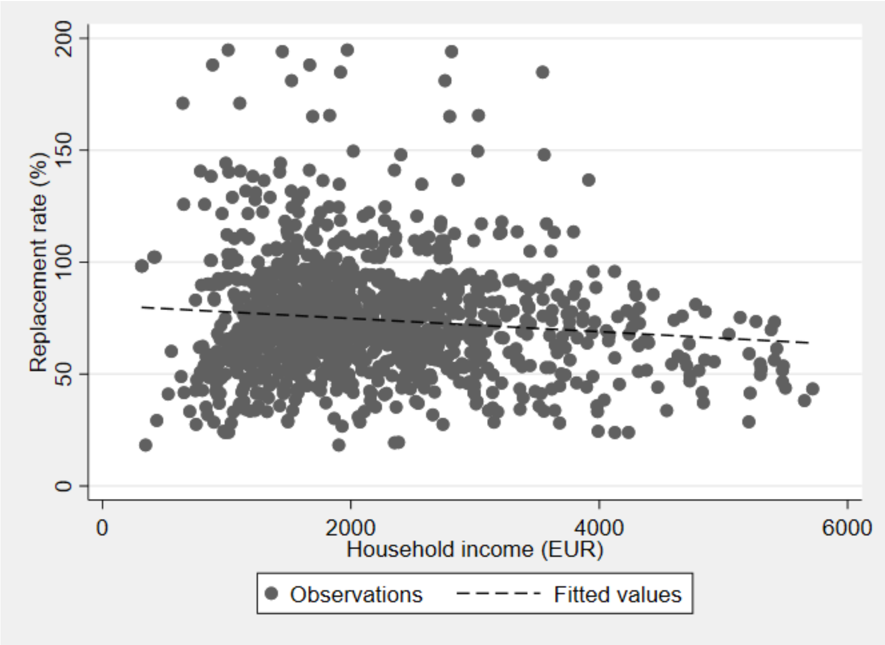


Data: German Socio-Economic Panel 1991-2017

retirement plan would be, we could simply calculate the empirical replacement rates; i.e., the rates that retired households realized. However, using administrative data, there is no obvious way of assessing whether the replacement rate they realized were high enough. Thus, I use the household income from SOEP, calculate the replacement rates for recent cohorts, and select on retirees who have more than median income satisfaction after retirement. The assumption is that retirees who evaluated their household income with value higher than seven on a scale from zero to 10 are satisfied. It turns out that, plotted against income, the replacement rate also decreases with income, with values similar to those in the main results (For simplicity, I only calculated singles), see Figure 3.10.

To sum up the findings of the sensitivity analysis, many modifications of the default strategy lead to similar results. The benchmark replacement rate for singles turns out to be fairly robust. While the exact relationship between the metric and income is not equal across models, it is consistently negative, which indicates that the benchmark replacement rates are higher at lower incomes and vice versa. For couple households, the standard errors are higher, leading to mostly insignificant results. Still, the scaling component is also always negative (See Table 3.2).

Figure 3.10: Empirical replacement rates for satisfied retirees (only singles)



Note: Each dot represents one household. Replacement rates are calculated by dividing the first household income after retirement by the final working income before retiring. Data: German Socio-Economic Panel 1991-2017

3.5.5 Goodness-of-fit

In this section, I compare the fit of the main model to more general models with stronger assumptions. In particular, I compare the AIC of the model using the BUC strategy and a binary fixed-effects estimator implementing an income-dependent identification as outlined in 3.1 against an income-independent identification estimated with a fixed-effects ordered logit model.

I calculate the respective Akaike Information Criteria from the log likelihood as follows:

$$AIC = -2 \times (LL) - 2k \tag{3.12}$$

where k indicates the number of model parameters which is equal across the two models. The log likelihood of the main model amounts to 114,036 whereas the base independent model results in 193,969. Thus the AIC of the main model is smaller than the AIC of the more general form, indicating that the fit of the main model is superior.

3.6 Discussion

3.6.1 Comparison to existing results

Dudel et al. (2016) found that the replacement rates that maintain income satisfaction are around 86%. For singles, my benchmarks were similar, but only for incomes up to EUR 2,000; whereas for incomes above EUR 2,500, my benchmark replacement rates were substantially lower, at around 60%. Thus, given that Binswanger and Schunk (2012) pursue a very different approach, my results are surprisingly similar to their finding for the Netherlands, (see section 3.2). Finally, Dudel et al. (2020) found benchmarks of around 95% for Germany based on expenditure data. These benchmarks match my benchmarks only for incomes up to EUR 900. While the differences are likely attributable to the welfare indicator applied (subjective equivalence scales have been lower than expenditure-based equivalence scales), they could also mean that the existing income-independent approaches suggest benchmarks that are too high for more affluent households.

An obvious policy implication would be that the benchmark replacement rates should be compared with the realized replacement rates from recent cohorts in Germany. For the empirical replacement rates, when they are available based on register data, the replacement rates levels strongly depend on the reference time span applied. Using a similar time span to that in this in this paper, Kluth and Gasche (2015) found replacement rates of around 50 percent. Across income classes, they only provided, what they called, life-cycle replacement rates across income classes, which implies, that the reference period was the average income from the life-cycle. For West Germany, these replacement rates increased with income from 30 to 45 percent (they only considered individuals). Note however, that the authors did not take into account income from wealth. Still, it is important to note that the realized replacement rates were lower for low earners while the results of this paper suggest that they should be higher for low-earners. The opposite is the case for Eastern Germany, where replacement rates are higher for low earners (Kluth and Gasche, 2015).

3.6.2 Explanation attempts

My results have shown that the benchmark replacement rates are higher at low incomes and vice versa. But why might this be the case?

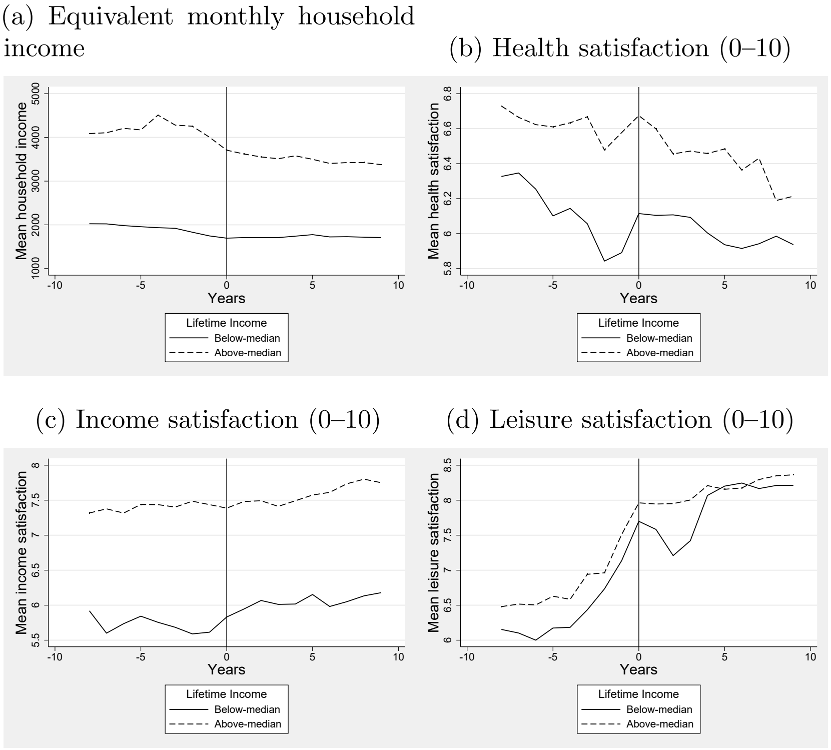
When households transition into retirement their expenses usually change (e.g. Bernheim et al., 2001; Schwerdt, 2005). In particular, their work-related costs, such as their expenditures on commuting, business clothes, accommodation or travel tend to decrease or drop after retirement. Also savings, when regarded as a form of costs, drop or at least decrease upon retirement. Conversely, because households have fewer time constraints, other types of expenses, particularly for leisure, typically increase. Finally, households' age-related costs, such as their health expenditures tend to increase, because they are, on average, older in retirement than when working. A priori, it is unclear whether those changes in expenses are proportionally larger or smaller across income classes. However, if the decrease in retirement costs is, in sum, larger for more affluent households, they would need less of their income to achieve their desired utility level than poorer households would. There is some evidence, from the US, that people with low wealth accumulations experience higher expenditures cuts after retirement (due to poor health), whereas for the richer part expenditures increase (Moran et al., 2021). For Germany, Schwerdt (2005) finds that people with high replacement ratios have consumption increases after retirement whereas for low replacement ratios consumption drops up to 30%. While the SOEP core study does not have detailed expenses on expenditures, the SOEP included a self-reported variable of saving until 2014. The respondents estimated how much of their monthly income is left to save. Because the variable is not available for the whole sample, and prone to measurement error, I did not include it in the income variable but it may be used to assess whether savings are proportionally larger or smaller across income classes. A simple scatter plot indicates that the saving rate tends to be lower for higher incomes (results available upon request) which contradicts the hypothesis stated above, and thus, this approach is not pursued further.

Recent literature has suggested that as people spend more time in retirement, they perform activities such as shopping and meal preparation more efficiently (Aguiar and Hurst, 2005; Luengo-Prado and Sevilla, 2013). Therefore, they need a smaller amount net income to achieve the same utility level that they did while working. It may be the case that those efficiency gains are more pronounced among the more affluent households. Again, there is currently no consent that this is the case, and the data do not allow for such an analysis.

Another potential explanation for my findings, albeit only for couple households, is that retired couples benefit from economies of scale. From the equivalence scale literature we know that economies of scales are greater among richer households. It is possible that more affluent couples are able to increase their economies of scale after retirement, and are thus able to achieve higher utility with less income.

Testing these these potential explanation is beyond the scope of the paper, but to get a clearer picture of what – apart from income – distinguishes poorer households from more affluent households during the years before and after retirement, I restructure the

Figure 3.11: Income and satisfaction before and after retirement across income classes: magnitudes



(a) Equivalent monthly household income

(b) Health satisfaction (0–10)

(c) Income satisfaction (0–10)

(d) Leisure satisfaction (0–10)

Notes: The dotted line represents the sample with average lifetime income above the median, and the solid line represents the sample with average lifetime income below the median. The reference period 0 represents the period of retirement. For the discrete variables (a)-(c), I show the unweighted population average of the absolute change with respect to the reference period. The graph for the continuous variable income (d) is illustrated without standardisation. The modified OECD scale is used for calculating equivalent household income.
Data: German Socio-Economic Panel 1991-2017

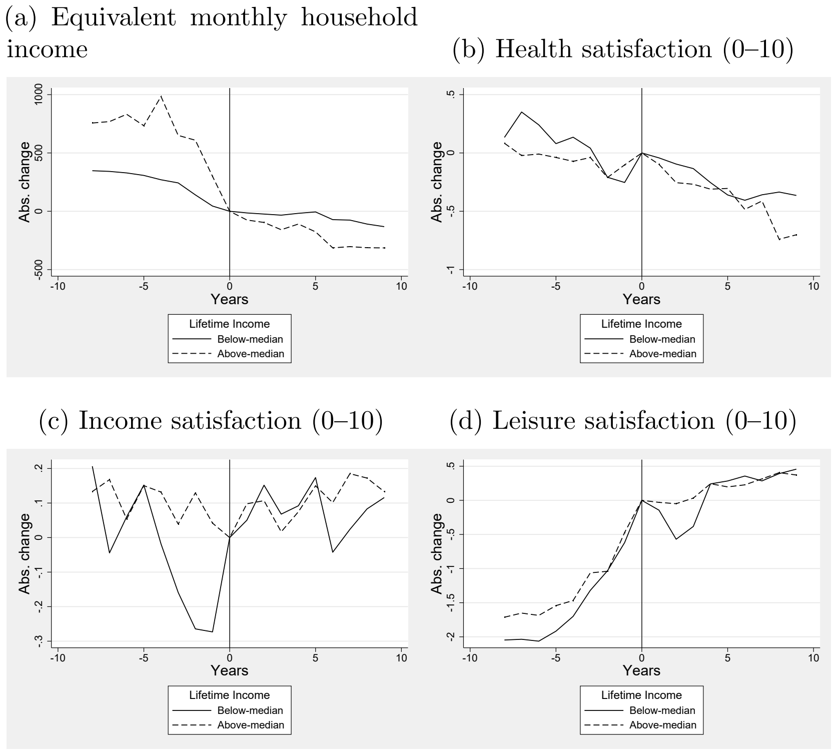
data in an event study design in which the point of retirement is the reference year. I calculate the equivalent income for all households in the sample (see section 3.4) using the square root scale[9] income of all households; and average over all reported incomes within the households, deflated by the consumer price index. Based on the average lifetime income, I split the sample into above and below the median. This enables me to study the variables that have been found to play a key role during the retirement process (e.g., Bönke et al., 2018).

Figure 3.11 a-c displays levels of income, health and leisure satisfaction with Panel (a) showing equivalent monthly income in EUR. Figure 3.12 does the same, except in terms of the absolute change in value with respect to the point of retirement.

All panels show that throughout the considered time span, the more affluent households are, on average, much more satisfied not only with their income, but also with their

[9] Dudel et al. (2020) find that empirical equivalence scales for Germany are very close to this heuristic

Figure 3.12: Income and satisfaction before and after retirement across income classes: Absolute changes with respect to the year of retirement


(a) Equivalent monthly household income

(b) Health satisfaction (0–10)

(c) Income satisfaction (0–10)

(d) Leisure satisfaction (0–10)

Notes: The dotted line represents the sample with average lifetime income above median and the solid line below median. The reference period 0 represents the period of retirement. For the discrete variables (a)-(c), I show the unweighted population average of the absolute change of respect to the reference period. The graph for the continuous variable income (d) is illustrated without standardisation. The modified OECD scale is used for calculating equivalent household income.
Data: German Socio-Economic Panel 1991-2017

health and leisure. While these results indicate that money does have an effect on health and leisure, this association might simply stem from the effect of income on general life satisfaction, which in turn correlates with satisfaction in those domains.

As the identification strategy, outlined in section 3.3, is based on within household variation, it is the change in these variable that is more important. As Figure 3.12 a) shows, in both groups, income drops shortly before retirement and steadily stabilizes after retirement. Panel (b)-(d) show the absolute changes in health, income, and leisure satisfaction with respect to the point of retirement. Despite the mostly positive effect of retirement on health (Eibich, 2015), health satisfaction deteriorates after retirement, presumably due to age. Leisure satisfaction increases steadily around retirement, but the marginal increase declines as people move further away from the point of retirement. For the absolute changes, there are no notable differences between the two income groups for either health or for leisure satisfaction.

With respect to changes in income satisfaction, the graphs follow a pattern that is often found in the literature on life and/or income satisfaction. Namely, that satisfaction is low in mid-life and increases around retirement (e.g., Blanchflower and Piper, 2022). However, for the poorer households there is a large drop in the five years preceding retirement.

Income satisfaction is very low for the households with low incomes before they retire. This could be the result of end-of-career work fatigue (the utility from work is usually higher at higher incomes), or concerns about having a low retirement income. At retirement, the satisfaction levels of these households go up, and develop similarly to those of the more affluent group thereafter. Because the relative changes in income are similar, these households need a larger fraction of their income to compensate for the relatively sharp increase in their income satisfaction.

3.6.3 Policy implications

The results of this study show that, after retirement, lower-income households have to make an extra effort to maintain themselves financially at the welfare level they had before retirement (which might have been low in the first place; see limitations section below). Extra effort implies, that while they are working, these households have to consume less of their net income to accumulate wealth, as they have replace proportionally more of their earnings than higher earners do. However, given that they usually have a lower capacity to save than high earners, because they have less wealth and higher fixed costs, they might not be able to do so on their own (Crossley and O'Dea, 2010).

Many pension systems have a tax-financed redistributive component in which low earners are supported by the high earners (e.g., Sweden, Denmark). Others have a large progressive component where larger contributions result in relatively smaller benefits. In Germany, the constant benefit-to-contribution rates are historically deeply rooted in the Bismarckian system. Following the "Äquivalenzprinzip", pension benefits are to a large degree determined by earning contributions, with high contributions being rewarded with high pension benefits (until a certain income level, beyond which the system is no longer mandatory). However, parts of the system have already been changed under the current government.

3.7 Conclusion

In this paper, I addressed the question of what a good choice of a replacement rate would be when making retirement saving plans. I found that the replacement rates that maintain income satisfaction through retirement are higher for households with lower incomes and lower for household with higher incomes. That was shown to be the case for both single and couple households, whereby the results for singles are more precise and consistent across a wide range of changes in the applied approach. While I offered some potential explanations, I leave further investigations of these findings to future research.

It is important to stress the relative dimension of this analysis. Maintaining income satisfaction does not rule out the possibility that a household has difficulties to making ends meet. Therefore, poverty thresholds should always be taken into account additionally (e.g. Love et al., 2008). Furthermore, while income from wealth was considered, households may hold wealth that has not been reported. Similarly, potential inheritances could increase the households' retirement income were not taken into account.

Research on benchmark replacement rates is still scarce, which is surprising given their practical relevance. This approach could be applied to other countries, where datasets include income satisfaction. Meanwhile, measuring realized, desired, and welfare maintaining replacement rates for the very same households is a potential avenue for future research.

Chapter 4

In and out of unemployment – labour market transitions and the role of testosterone[1]

4.1 Introduction

'Joblessness leaves permanent scars on individuals' (Arulampalam, 2001), partly because unemployed individuals might be perceived (and might perceive themselves) as violating a social norm. On the other hand, it can also be a rational decision to remain unemployed for a period to hold out for a better job offer and improve the job match. The economic literature has shown that there are various factors which explain why individuals become unemployed or stay in unemployment. However, the focus has been on observable factors, such as individual and household characteristics or the past unemployment experience and duration (see, e.g., Gregg, 2001). More recent evidence points to personality traits and non-cognitive skills as influential factors of job search behaviour and unemployment duration. Studies have investigated, e.g., the locus of control (Caliendo et al., 2015; Heckman et al., 2006; Schurer, 2017), impatience (DellaVigna and Paserman, 2005), the Big 5 personality traits (Viinikainen and Kokko, 2012), or self-efficacy and interpersonal skills (Uysal and Pohlmeier, 2011).

Hormones have been linked to a number of non-cognitive skills and personality aspects. In particular, testosterone is prominently linked to risk-attitude and aggression (Dabbs, 1992; Dabbs et al., 2001; Hughes and Kumari, 2019), but also to skills such as motivation, pro-social behaviour, persistence, or numerical ability (Apicella et al., 2008; Carré and McCormick, 2008; Dabbs et al., 2001; Welker and Carré, 2015). Likely related to these attributes, testosterone has also repeatedly been found to predict men's labour market performance (Dreher et al., 2016; Gielen et al., 2016; Nye et al., 2017). Moreover, testosterone also seems to affect occupational choices (Dabbs, 1992; Greene et al., 2014). Yet, surprisingly testosterone has not been investigated as an explanatory factor of

[1] This manuscript is a preprint of a version that is now published at *Economics and Human Biology* DOI: 10.1016/j.ehb.2022.101123

unemployment, something we seek to address in this paper.

We investigate whether differences in serum testosterone levels of men can explain transitions in and out of unemployment. We use data from *Understanding Society (UKHLS),* a longitudinal household survey covering about 40,000 households from the United Kingdom, and link it with the *Health and biomarkers Survey*, which holds a range of biomarker data, including the circulating level of testosterone. We examine two samples of initially employed or initially unemployed men aged 20 to 60, and we standardise their testosterone levels for age and time of the sample.

Taking advantage of the longitudinal nature of the data, we apply dynamic probit random-effects models to estimate labour market transitions.

We contribute to the literature by providing novel evidence on latent biological mechanisms which affect labour market trajectories. Previous studies have only considered inflammation markers in relation to unemployment but not hormones such as testosterone (Sumner et al., 2020). Moreover, unlike previous studies we examine actual testosterone levels measured in a recent blood sample rather than 2D:4D ratio, which is a prominent marker for prenatal exposure to testosterone (see, e.g., Gielen et al., 2016). The closest study to ours is Hughes and Kumari (2019), who examined the impact of testosterone on risk tolerance, gross earnings, household net income, and socio-economic status. In contrast to our study, they only considered the likelihood of being in work at a single point in time, whereas we consider labour market transitions.

Findings from our preferred regression specification indicate that the risk of remaining unemployed significantly declines in testosterone level for unemployed men. In contrast, testosterone has no significant effect on the unemployment risk of employed men.

Cognitive and non-cognitive skills, such as numerical skills or logical reasoning, might partly explain these findings as these are associated with high testosterone levels. In line with previous studies, our descriptive evidence shows that men with high testosterone levels indeed performed better in these areas. In addition, we find suggestive evidence that individuals with higher testosterone search differently for a job.

Our findings highlight how latent biological processes beyond illness and disease affect labour market outcomes. For example, when designing job search assistance programs, policymakers must be aware that biological mechanisms can drive differences in job search behaviour. Thus, due to their inherent skills, some individuals might require specific forms of assistance to thrive - for example, individual training rather than group sessions. This study contributes to our understanding of such mechanisms by providing comprehensive evidence on the role of testosterone.

The rest of this paper is set out as follows. Section 2 reviews the literature on biomarkers, and based on this literature we discuss how testosterone could affect labour market transitions. Section 3 presents our data and Section 4 outlines our empirical estimation strategy. Section 5 contains descriptive statistics. Section 6 documents the results from our regression specifications. Section 7 discusses further possible mechanisms and descriptive evidence for these potential pathways. Section 8 concludes.

4.2 Testosterone and the labour market

4.2.1 Existing literature

Testosterone has been related to different forms of health issues, e.g., cardiovascular disease (Elagizi et al., 2018), but the evidence has not been very conclusive and causal pathway not fully understood (Bann et al., 2015; Hughes and Kumari, 2019). More convincingly, among men, testosterone seems to affect risky health behaviours and thus, different forms of health hazards (Booth et al., 1999).

Testosterone also plays a role for demographic outcomes, such as fertility, divorce and mating (e.g., Bütikofer et al., 2019), fitness and sport (e.g., Hsu et al., 2015), but also for labour market outcomes (e.g., Coates et al., 2009; Dabbs, 1992; Dabbs Jr. et al., 1990; Parslow et al., 2019).[2] For example, in a twin study on Dutch men, more prolonged prenatal testosterone exposure led to higher earnings during the working life (Gielen et al., 2016).[3] Other studies found education to be lower among people with low testosterone levels (Bann et al., 2015; Nye et al., 2017). Coates and Herbert (2008) followed the daily business of 300 traders in London and found that high levels of testosterone lead to higher profits on that day. Testosterone also affects the choice of occupation. Low testosterone individuals seem to choose more people-oriented jobs, whereas high testosterone individuals choose more things-oriented jobs (Dabbs Jr. et al., 1990; Hell and Päßler, 2011; Nye and Orel, 2015).[4] Typical jobs that have been related to high testosterone are sportsmen, sales men, actors, or politicians (Dabbs Jr. et al., 1990). The evidence is not conclusive, though. A more robust finding is that individuals with high testosterone levels have a higher probability to be self-employed (Greene et al., 2014; Nicolaou et al., 2017; Sapienza et al., 2009).

[2] While testosterone is present in both sexes, most of the experimental studies in the literature have focused on men. Important exceptions looked at both sexes (Dabbs et al., 2001; Gielen et al., 2016; Nye et al., 2017; Sapienza et al., 2009) or exclusively at women (Bütikofer et al., 2019; Parslow et al., 2019).

[3] Among women, high testosterone levels are expected to be associated with higher earnings as well, as women with higher testosterone levels tend to work in male-dominated occupations, which tend to be better paid. However, recent empirical evidence found the opposite or no effect (Bütikofer et al., 2019; Gielen et al., 2016; Nye et al., 2017).

[4] Women that have higher testosterone levels tend to choose jobs that are male-dominated, whereas women with low levels choose more female-dominated jobs (Nye and Orel, 2015). This observation has been used to explain parts of the gender pay gap (e.g., Gielen et al., 2016).

The findings discussed above are usually attributed to non-cognitive skills and individual characteristics associated with high testosterone levels. Typical characteristics that have been stressed in the literature are, among others, being independent, self-centred, adventurous, achievement-oriented, and focused on personal goals (Greene et al., 2014). Further, high testosterone is associated with risk-taking (Apicella et al., 2008; Coates and Herbert, 2008; Hughes and Kumari, 2019; Stenstrom et al., 2011), dominant behaviour and aggression (Archer, 2006; Chance et al., 2000; Dabbs, 1992; Dabbs et al., 2001; Schaal et al., 1996), but also status-enhancing pro-social behaviour.[5] For example, Dreher et al. (2016) injected testosterone or a placebo to 40 young men and found that in an economic bargaining game, treated individuals were indeed more aggressive towards others. Still, at the same time, they were also more generous when it promoted social status. Similarly, individuals with high testosterone levels show more initiative forming friendships and are, therefore, able to build up larger social networks (Booth et al., 2006; Cheng et al., 2013). In other game studies, men with high testosterone levels were more willing to engage in competitive tasks (Carré and McCormick, 2008) and they showed more persistence solving an undoable task (Welker and Carré, 2015). Cognitive abilities have also been related to testosterone. While early work reported that young boys with high testosterone levels lack intelligence (Chance et al., 2000; Dabbs, 1992), more recent work showed that individuals with high testosterone levels have higher numeric capabilities and thus perform better in computer science or related occupations (Brookes et al., 2007; Brosnan et al., 2011). Similarly, individuals with more prolonged prenatal exposure to testosterone performed better in the cognitive reflection test (Bosch-Domènech et al., 2014), a test which measures the tendency to override an intuitive incorrect answer, and which has therefore been used as a measure of reflection in decision making (Frederick, 2005). Finally, a series of studies showed that people with high testosterone levels perform better in face-to-face situations (e.g., Dabbs et al., 1997; Mazur, 1985). For example, Dabbs et al. (2001) interviewed and filmed male college students and found that individuals with high levels of testosterone appeared more forward and independent and focused directly on the target. They were also more restless and oriented toward action.

4.2.2 Testosterone and employment transitions

There are multiple pathways of how testosterone might relate to unemployment. We focus on differences in job search behaviour and self-selection by occupational choice while distinguishing between entry into unemployment and exit from unemployment.

[5] The effect of testosterone on prosocial status-promoting behavior and risk has been found to be moderated by cortisol (e.g., Mehta and Prasad, 2015).

As noted above, high testosterone levels are associated with aggression (in the broadest sense), which includes competition-seeking and dominant behaviour (Archer, 2006; Chance et al., 2000), or even pro-social behaviour (Dreher et al., 2016). If pro-social behaviour associated with higher testosterone levels leads to larger social networks, then these networks might constitute an important resource for the job search (Ponzi et al., 2016). Moreover, the job search in general, and assessment centres or job interviews in particular, might favour competitive, dominant and pro-social individuals. Thus, individuals with high testosterone might invest more effort into their job search, since adopting the required behaviour comes more natural to them (Dabbs et al., 2001, 1997) and exerts less mental strain than it might for individuals with low testosterone. For similar reasons, individuals with high testosterone might perform better in such situations, and might thus be more likely to receive a job offer. Yet, testosterone might also affect individuals' likelihood to accept a job offer. Individuals with low testosterone, who are less willing to take risks, might accept a job offer earlier. In contrast, high testosterone individuals might be more inclined to take a risk and look for a better position. This is in line with the evidence that individuals' with a higher level of testosterone are more reflective in the decision-making process (Bosch-Domènech et al., 2014). Re-employment, therefore, would take longer for individuals with high testosterone but might result in a better job match. Conversely, due to the perceived social stigma of unemployment, high testosterone individuals, worried about their social status, might be more inclined to take first job offers to move out of an economically disadvantaged position.

For individuals in employment, once employers learn about their employees' productivity competition-seeking and dominant behaviour may become less critical. To some extent, such behaviour might even be considered detrimental, e.g., for the performance in teams. Hence, individuals with high testosterone levels may be at an increased risk of entering unemployment compared to individuals with normal testosterone levels.

In terms of occupational choice, workers with high testosterone levels might select into jobs that are perceived as offering greater rewards at higher risks. For example, positions with performance-based remuneration and where redundancies are more common, like in sales or self-employment. Besides, higher numeric capabilities associated with high testosterone levels (Brookes et al., 2007; Brosnan et al., 2011) would also imply a selection into certain occupations or sectors. Individuals with low testosterone tend to be more risk-averse and might prefer jobs that offer more stability (e.g., in the public sector). Such occupational sorting would imply that high testosterone individuals are more likely to face unemployment, but are able to find re-employment relatively quickly. In contrast, individuals with low testosterone are less likely to lose their job but stay longer in unemployment if they become unemployed.

In summary, the existing evidence suggests that testosterone might affect transitions both in and out of unemployment, but the direction of the effect is ambiguous, and it may differ for exits and entries into unemployment.

4.3 Data

The UK Household Longitudinal Study *Understanding Society* is one of the few surveys available that collects both data on testosterone levels (among other biomarkers) as well as annual longitudinal data on individuals and households characteristics. *Understanding Society* is the successor of the British Household Panel Survey (BHPS), started in 2009, and at the time of writing 9 waves of data are available. With approximately 40,000 households (at Wave 1) in the United Kingdom, it collects a range of individual- and household-related information that also enables the researcher to trace labour market trajectories. Approximately five months after their Wave 2 or Wave 3 (2010-2013) mainstage interview adult participants received a health assessment visit from a registered nurse ('Health and biomarkers survey').[6] A range of bio-medical measures was collected from over 20,000 adults, including testosterone levels.

4.3.1 Health and biomarkers Survey

To be eligible for a nurse interview survey respondents must have completed a full face-to-face interview in the most recent mainstage wave, lived in Great Britain, completed their interview in English and, for women, were not pregnant. Among those eligible, approximately 20,700 (57%) took part, of which 13,107 (68.5%) had at least one biomarker which was successfully obtained and processed (Benzeval et al., 2014). During the nurse visit, blood samples were taken to extract a range of biomarker data, including measures of growth hormones (testosterone, DHEA's, IGF-1). Serum testosterone, the specific biomarker of interest for this study was measured using an electrochemiluminescent immunoassay on the Roche Modular E170 analyser.

Testosterone levels show wide variation and are considered within a normal range between 9-25 nmol/L. Testosterone varies by time of day such that values in the morning are higher than those found in the afternoon or evening (See Table 4.1). The level of testosterone also declines in age (See Figure 4.1).

[6] The nurse health visit was conducted among adult survey participants from the General Population Sample (GPS) which comprises of households in the UK and BHPS sample only. The nurse visit took place after wave 2 (May 2010-July 2012) for those individuals in the GPS and after wave 3 (June 2011-July 2012) for BHPS sample respondents.

Table 4.1: Level of testosterone (nmol/l) and interview time

The start time of the interview (hour)	Testosterone (nmol/l)	
	Mean	Std Dev
9	17.50	6.31
10	17.68	5.87
11	17.38	6.25
12	17.31	6.03
13	16.05	6.48
14	16.05	5.76
15	14.49	5.39
16	14.83	5.62
17	15.14	5.21
18	14.64	5.52
19	14.06	4.94
20	13.24	4.42

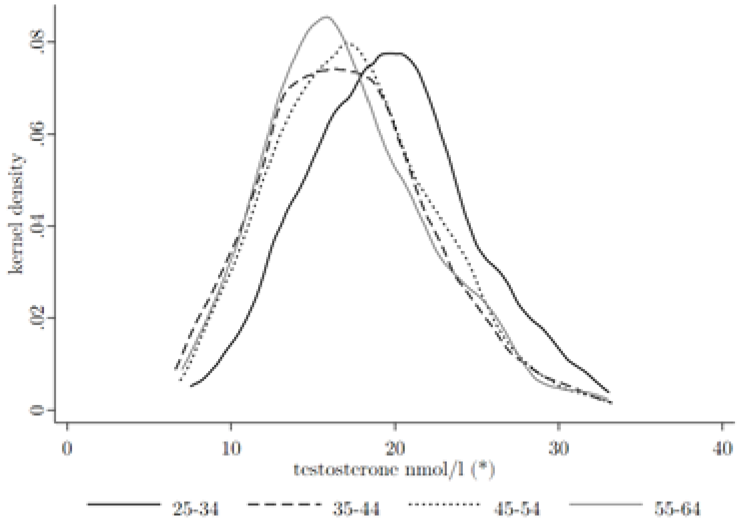
Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. $N = 4,605$ men with a positive level of testosterone in the age range 16 to 70 who had their interview started between 9 am and 8 pm.

Apart from time and age, differences in testosterone levels are expected to originate from prenatal development, particularly in-utero exposure to testosterone. The sex difference in testosterone is almost non-existent before puberty but up to 20 times higher for men thereafter (e.g., Handelsmann et al. 2018). However, where the variation for testosterone levels among men comes from is not entirely clear. There is evidence from mice that maternal stress alters plasma testosterone levels in fetal males (Ward and Weisz, 1980). Similarly, testosterone levels have been found to interrelate with other hormones like cortisol and hence to stress, but evidence for humans is scarce (Braude et al. 1999). In the Health and biomarkers Survey, there are 3,597 men with a plausible level of testosterone in the age range 25 to 64 who had their interview started between 8 am and 8 pm.

4.3.2 Longitudinal data

In each wave of Understanding Society, survey respondents are asked about their current labour force status. This information is used to estimate the transition into unemployment as well as the degree of persistence. We start by trimming the Health and biomarker Survey to men who are between 20 and 60 during the nurse visit. This survey also holds information about participants' social and economic circumstances, including

Figure 4.1: Level of testosterone (nmol/l) and age



Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. N=3,597 men with a positive level of testosterone in the age range 25 to 64 who had their interview started between 8 am and 8 pm. (*) showing the level of testosterone (nmol/l) corrected by the time of the nurse visit.

the current labour force status.[7] We restrict the sample to those men who state being either unemployed or, if employed, an employee. We drop self-employed individuals since this group of individuals is likely to differ from employees on unobservable characteristics (such as personality type) as well as their labour supply behaviour. Also, the sample size is insufficient to include them as a separate group.

In the next step, we merge this sample to the mainstage wave of Understanding Society. As noted above, the nurse visit took place shortly after either Wave 2 or Wave 3. The interviews in the primary survey were conducted, on average, virtually one year apart. However, the time difference between the nurse visit and the follow-up interview at the primary survey is less than one year. We restrict our sample to individuals who are either employed or unemployed at the follow-up interview. Our final sample consists of 2,115 individuals, out of which 111 (5.25%) were unemployed during the nurse visit, and 2,004 were employed (94.75%).

[7] Possible answers are: (1) self employed, (2) paid employment(fulltime/parttime), (3) unemployed, (4) retired, (5) on maternity leave, (6) family care or home, (7) full-time student, (8) long-term sick or disabled, (9) government training scheme, (10) unpaid, family business, (11) on apprenticeship, (12) doing something else.

Table 4.2: Database and number of days between consecutive interviews

		Days until the next interview	
Database of *Understanding Society*	Period t	Mean	Std Dev
Primary survey	-2	366	23
Primary survey	-1	151	25
Health and biomarkers survey (Nurse visit)	0	217	30
Primary survey	1	369	36
Primary survey	2	377	82
Primary survey	3	-	-

Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. $N = 1,880$

4.4 Methodology

We aim to understand how the individual's testosterone level impacts labour market changes between the nurse visit and follow-up interview one year later in the primary Understanding Society survey. We distinguish between employed and unemployed men aged between 25 and 64. The reduced form model for unemployment can be written as follows:

One way to assess labour market dynamics is to consider transition matrices. In Table 4.3, the probability of being (un)employed at t, conditional on the labour market position at $t-1$ is presented. The first number in each cell shows the conditional probability for the full sample, and unsurprisingly there is a high level of state dependence. This means that the chances to stay employed are higher for someone who was already employed in the previous period, and similarly for the unemployed. However, these probabilities differ substantially with respect to the initial labour market position. For example, the risk to stay unemployed is 75 per cent for the initially unemployed, but only 32 per cent for those initially employed.

As the focus of our study is to identify the effect of testosterone on the unemployment risk, we differentiate the transition matrix further according to the testosterone groups (see Table 4.4). On the one hand, we find that for initially unemployed men the conditional probability of staying unemployed is highest in the first decile (86 per cent) – especially with respect to the 2^{nd} – 9^{th} decile (70 per cent). For initially employed men, we find that those in the top decile have the highest conditional probability of entering unemployment (3.6 per cent). A general pattern in Table 6 is that low testosterone is

Table 4.3: Transition matrix of labour market status

	employed$_t$	unemployed$_t$	Total$_{t-1}$
employed$_{t-1}$	98.14 (90.48) [98.27]	1.86 (9.52) [1.73]	94.49 (27.18) [98.52]
unemployed$_{t-1}$	36.21 (25.33) [68.42]	63.79 (74.67) [31.58]	5.51 (72.82) [1.48]
Total$_t$	94.73 (43.04) [97.83]	5.27 (56.96) [2.17]	

Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. $N = 5,460$. Numbers in ( ) / [ ] refer to the sample of initially unemployed / initially employed.

Table 4.4: Unemployment risk differentiated according to testosterone level

Testosterone (categorical)	Full Sample	Initially unemployed	Initially employed
	unemployed$_t$|unemployed$_{t-1}$		
1st decile	74.42	86.49	-*
2nd – 9th decile	59.90	70.00	33.33
10th decile	70.59	81.58	38.46
	unemployed$_t$|employed$_{t-1}$		
1st decile	1.30	12.50	1.13
2nd – 9th decile	1.73	8.70	1.61
10th decile	3.78	14.29	3.61

Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. $N = 5,460$. * Due to the short panel, we do not observe individuals who stay unemployed with a low testosterone level for those initially employed.

associated with higher persistence of unemployment, while high testosterone seems to be associated with a higher risk of entering unemployment.

4.5 Results

4.5.1 Base regression

Our base model controls for the labour market position in the initial periods, i.e., the nurse visit and the two waves before it, the labour market position in the previous period, and additional covariates. Furthermore, we include the level of testosterone in three different alternative specifications. Table 4.5 shows only the effect of testosterone and unemployment risk (complete output tables are available on request).

The first regression uses the full sample (see the first three columns of Table 4.5) and includes the level of testosterone in a linear trend (4.1), a second-degree polynomial (4.2), and as a categorical variable (4.3). In all three specifications, we find that the unemployment risk increases with the level of testosterone[8]. However, the magnitude is always small, and estimates are not significantly different from zero.

In section 4.2, we outlined potential reasons why we might expect the effect of testosterone to differ between the initially unemployed and initially employed. Once we condition on initial labour force status (columns four to nine of Table 4.5), we see that the direction of the effect and the magnitude change substantially. For those who are initially unemployed, Table 4.5 (column 4-6) indicates that the risk of staying unemployed declines in the level of testosterone. This finding is independent of the specification and significantly different from zero.

We also find a significant impact of testosterone on the risk of becoming unemployed for the sample of initially employed. In contrast to the sample of initially unemployed, there is a positive sign, indicating that higher levels of testosterone increase the risk of becoming unemployed. Model (4.3) suggests that this effect is largest among individuals with a very high level of testosterone – no significant difference is observed between the low and the medium levels of testosterone. Independent of the sample we use, we do not find any support for a second-degree polynomial relationship of testosterone and the unemployment risk.

For the two subsamples, we also calculate the average partial effects (APE) for Model (4.3) at the individual level. The partial effect for the initially unemployed is the difference (in percentage points) of becoming unemployed if the person was unemployed at $t-1$ and had a testosterone level in the 2^{nd} to 9^{th} decile, resp. 10th decile, compared to the first decile.

Compared to the lowest category, the unemployment risk is reduced by 25.4 percentage points for the medium category and by 30 percentage points by the higher category. In the case of the initially employed, the previous labour market status is set at being employed. The magnitude of the APE is positive but much smaller and not detectable for the 2^{nd} to 9^{th} decile. For the highest decile, we find an elevated unemployment risk of, on average, 2.2 percentage points.

4.5.2 Robustness checks

In order to assess the sensitivity of our main results for each subsample, we carry out a number of robustness checks.[9] Our primary focus is to understand the link

[8] In the case of specification (2), when looking at a range of 5-30 nmol/l testosterone, there is in the beginning a reducing effect but the slope turns positive from around 10 nmol/l.

[9] Additional robustness estimations which are not described in detail here include dropping covariates. However, none of the tests lead to qualitatively different findings.

Table 4.5: Effect of testosterone on unemployment risk

	Full Sample			**Initially unemployed**			**Initially employed**		
Model	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
testosterone nmol/l	0.0167	-0.0273		-0.0843*	-0.274*		0.0270*	-0.00565	
	(0.0127)	(0.0502)		(0.0443)	(0.165)		(0.0138)	(0.0521)	
(testosterone nmol/l)^2		0.00123			0.00524			0.000908	
		(0.00136)			(0.00417)			(0.00141)	
testosterone									
1^{st} decile					*reference category*				
2^{nd} – 9^{th} decile			-0.132			-2.027**			0.162
			(0.242)			(0.812)			(0.274)
10^{th} decile			0.355			-2.331**			0.622*
			(0.299)			(1.001)			(0.339)
Observations	5,460	5,460	5,460	309	309	309	5,151	5,151	5,151
LogLikelihood	-556	-555.6	-559.7	-93.76	-92.90	-98.25	-435.9	-435.6	-438.8

Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. $N = 5,460$. All models include controls for the labour market status in the initial periods (prior to the nurse visit) as well as the previous period (i.e., lagged labour market status). Model 1 includes a linear trend in circulating testosterone, Model 2 models testosterone with a quadratic polynomial, and Model 3 includes testosterone as a categorical variable.

Table 4.6: Average partial effects

	Initially unemployed	**Initially employed**
1^{st} decile	*reference category*	
2^{nd} – 9^{th} decile	-0.254*	0.004
	(0.137)	(0.007)
10^{th} decile	-0.300*	0.022
	0.168	0.016
Individuals	109	1,771

Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. ***,**,* refers to statistically significant at 1%, 5% and 10% level respectively.

between testosterone and labour market dynamics among working-age men. One concern, therefore, is that our results are driven by the age restriction (20-60) we imposed when choosing our initial sample. However, for those at the bottom and the tail of the sample, labour market decisions might be influenced by external factors. For example, individuals might consider delaying entry to the labour market due to attending university or reducing labour supply prior to entering retirement. We first narrow the age window by iteratively dropping the youngest and oldest age, until the age range includes those between 25 and 55. The respective coefficients for the categorical testosterone variable are shown in Figure 4.1. For the group of initially unemployed, we find a stable level for both coefficients, and in all iterations, the coefficients are significantly different from zero at the 10 per cent level (in most cases also at the 5 per cent level). These findings also hold if we instead estimate a specification where testosterone is included as a linear term. Turning to the results for those who are initially employed, the findings are also relatively stable with respect to the highest decile, though we note that in two out of five cases, the effect is insignificant. If we include testosterone as a continuous variable, we find in all specifications a significant effect. Moreover, we find no significant difference between the medium and low-level testosterone groups.

We next consider our categorisation of the variable defining the low, medium and high-level groups. In our main specification, we use the bottom and top decile as cut-off points to determine the three categories. We re-run the model and use as cut-off points between the bottom (top) five to 20 per cent, moving in one percentage points (see Figure D.2). For the initially unemployed, we find in line with our expectations that the coefficients for both groups are declining marginally. However, with very few exceptions, the coefficients are statistically significantly different from zero at the five per cent level.

For the group of the initially employed, the coefficient for those with a high level of testosterone stays rather stable, but significance increases.

In order to address the correlation between observable characteristics and individual-specific effects, we apply the Mundlak-Chamberlain approach as outlined in section 4.4.3. A subset of these results can be found in D.1 in the appendix. Once again, we estimate the model conditional on initial labour force status and classify individuals into three groups based on their testosterone level. The results support the findings in Table 4.5, namely that relative to the low testosterone group, individuals with medium and high levels of testosterone are significantly less likely to remain unemployed (APEs of 35 and 28 percentage points, respectively).[10] Moreover, the magnitude of these effects is larger than our main results. Turning to the initially employed group, consistent with our main findings we find those individuals with high levels of testosterone are significantly more likely to become unemployed (APE: 2.1 percentage points) in line with the results reported in Table 7.[11]

4.5.3 Continuous (un)employment

To integrate out the random-effects error terms, we have to impose an assumption on the distribution of the individual-specific unobserved effect ('auxiliary distributional assumption'). However, if the auxiliary distributional assumption is not valid, the random-effects estimator is not consistent. One option to circumvent this issue is to consider dynamic linear probability models, which impose no assumption on the individual-specific effects. Due to time-differencing, it does not impose any assumption on the distribution for the individual-specific effects. One example is the Arellano and Bond (1991) GMM estimator, which uses (among others) lags of the dependent variable as instruments. A crucial restriction is that there is no second-order serial correlation in the idiosyncratic shocks (there will be first-order correlation due to including the lagged dependent variable). However, to test this assumption, the dynamic sequence of the panel data set must hold at least four periods, and in our basic specification, our panel does not exceed three periods. Furthermore, due to differencing individual-specific time-invariant characteristics (e.g., testosterone level) are excluded.

To run a specification where we do not have to make any assumption about the distribution of the individual-specific effects, we use a simple probit model to estimate the probability of being unemployed in the first three periods after the nurse visit. For those initially unemployed, we run three different specifications, depending on whether the individual was unemployed in (4.1) one or more waves, (4.2) two or more waves, or

[10] Both are significant at the 1% level.

[11] Significant at the 5% level.

(4.3) in all three waves. For those initially employed, we are only able to control for one or more wave of unemployment in the first three periods after the nurse visit. We use the same covariates as in the basic specification, which were collected at the nurse visit, and account for the level of testosterone in a categorical way. Table 4.7 presents the marginal effects of the level of testosterone. In line with our previous findings, we see for the initially unemployed that those individuals with a medium level of testosterone are significantly less likely to experience unemployment. For those individuals with a high level of testosterone, the findings indicate that they are significantly less likely to be unemployed in all three waves. However, we do not find any significant effect on experiencing some unemployment spells. These findings are robust to changes in the cut-off point.

When turning to the initially employed, we find that individuals with a high level of testosterone are more likely to experience a spell of unemployment, though the effect is not significant. When changing the cut-off point to the top two deciles, individuals with a high level of testosterone are 4.1 percentage points more likely to experience at least one wave of unemployment compared to individuals with a low level of testosterone, and the effect is significant at the 5% level.

Table 4.7: Marginal effects

	Initially unemployed			**Initially employed**
Waves unemployed	≥1	≥2	3	≥1
1^{st} decile		*reference category*		
2^{nd} – 9^{th} decile	-0.267***	-0.286**	-0.285**	0.001
	(0.080)	(0.114)	(0.121)	(0.016)
10^{th} decile	-0.169	-0.160	-0.313**	0.035
	(0.140)	(0.158)	(0.145)	(0.026)
Individuals	91	91	91	1,609

Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. ***,**,* refers to statistically significant at 1%, 5% and 10% level respectively

To analyse the labour market trajectories of our sample members, we observe individuals for up to three waves following the nurse visit. The focus of our study is the transition between unemployment and employment (and vice versa), and therefore we only include these two states. We allow an individual to exit the panel but not to re-enter. As shown in Table 4.8, our final sample consists of 1,880 individuals who contribute 5,460 observations, out of which 309 (5.7%) were unemployed during the nurse visit, and 5,151 were employed (94.3%).

To estimate labour market transitions, we follow the economic literature by utilising

Table 4.8: Sample size

	Number of individuals	Number of observations
Employed during nurse visit	1,771	5,151
Unemployed during nurse visit	109	309
Total	1,880	5,460

Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey.

dynamic non-linear models (Arulampalam, 2001; Bhuller et al., 2017; Biewen and Steffes, 2010; Stewart, 2007). The idea is that the labour market dynamics follow a first-order Markov process, which means that the status in the previous period has a *genuine* effect on the position in the subsequent period. Moreover, if individual effects are persistent over time, not accounting for unobserved heterogeneity will lead to an over-estimation of state dependence (Stewart, 2007). The dynamic reduced-form model to estimate state dependence in unemployment can be written as follows:

$$y_{it}=\mathbf{1}(\alpha_1 y_{it-1} + x'_{it=0}\beta + v_i+\varepsilon_{it} > 0) \tag{4.1}$$

where the subscripts i= $1,...,N$ are individuals and $t = 1, \dots, T$ refer to the waves of the dynamic sequence (thus the three waves following the nurse visit, which are labelled as the post-period). The dependent variable (y_{it}) equals 1 if i was unemployed at wave t and 0 otherwise. Following the assumption of a first-order Markov process, y_{it} is explained by its lagged outcome y_{it-1} and the coefficient α_1 reveals the magnitude of state dependence in unemployment. Furthermore, $x_{it=0}$ is a vector of explanatory variables which were collected during the nurse visit at $t = 0$. As covariates, we include: age (linear and second order polynomial), highest qualification, self-rated health, region, urban identifier, household size, long-term disability and legal marital status. To ensure that the explanatory variables refer to the same time point as the measurement of testosterone, in our main specification, the explanatory variables are not time-varying. Lastly, v_i is an individual-specific time-invariant shock and ε_{it} is an idiosyncratic shock. Note, that we make the assumption that $v_i \sim$ iid $N\left(0, \sigma_v^2\right)$ and $v_i \perp x_{it=0}, \varepsilon_{it}\ \forall\ i, t.$ i.e., the individual-specific effects are randomly distributed across individuals and independent of observable characteristics. As part of our robustness checks and discussed further below, we follow Mundlak (1978); Chamberlain (1992) and relax this assumption by including time-varying covariates and their means for the period $t > 0$ (see also Wooldridge, 2005).

As explained in section 4.3.3, we trim the sample to individuals who were either employed or unemployed in the two waves before the nurse visit ($t = -1$ and $t = -2$). However, the time-invariant error-term might be correlated with the outcome in these periods, discussed as the 'initial conditions problem' (Wooldridge, 2005). As $y_{it=0} = y_{it=-1}$, we have four possible combinations of employment sequences in the pre-periods:

- Continuously employed ($y^1_{it=0}$): $y_{it=-2} = y_{it=-1} = 0$
- Short-term employed ($y^2_{it=0}$): $y_{it=-2} = 0$ and $y_{it=-1} = 1$
- Continuously unemployed ($y^3_{it=0}$): $y_{it=-2} = \ y_{it=-1} = 1$
- Short-term unemployed ($y^4_{it=0}$): $y_{it=-2} = 1$ and $y_{it=-1} = 0$

The individual-specific error term takes the following form when we condition on the initial period values (see also Wooldridge, 2005):

$$v_i = \sum_{r=2}^{4} \lambda_r y^r_{it=0} + a_0 + \alpha_i \tag{4.2}$$

Plugging (4.2) into the original specification (4.1) results in:

$$y_{it} = \mathbf{1}(\alpha_1 y_{it-1} + \sum_{r=2}^{4} \lambda_r y^r_{it=0} + x'_{it=0}\beta + a_0 + \alpha_i + \varepsilon_{it} > 0) \tag{4.3}$$

Kroft et al. (2013) provide evidence that the probability of exiting unemployment depends on the unemployment duration. To account for this aspect, we interact the lagged dependent variable with the variables referring to the labour market status in the pre-period.[12] Our model takes the following form for the full sample:

$$y_{it} = \mathbf{1}\left(\sum_{r=2}^{4} \gamma_r y^r_{it=0}\left(y_{it-1} = 0\right) + \sum_{r=1}^{4} \delta_r y^r_{it=0}\left(y_{it-1} = 1\right) + x'_{it=0}\beta + a_0 + \alpha_i + \varepsilon_{it} > 0\right) \tag{4.4}$$

with being employed in the previous period ($y_{it-1} = 0$) and continuously employed in the pre-period ($y^1_{it=0}$) as the reference category. We assume that both error terms follow a normal distribution, e.g., $\alpha_i \sim N\left(0, \mathrm{s}^2_\alpha\ \right)$ and $\varepsilon_{it} \sim N\left(0, \mathrm{s}^2_\varepsilon\ \right)$ and that ε_{it} is iid. As

[12] Note that our findings on the effect of testosterone are robust to various specifications of including the initial labour market status (e.g., interacting with the lagged labour market position, no interaction, not accounting for the initial labour market status).

α_i is time-invariant, the composite error term $u_{it} = \alpha_i + \varepsilon_{it}$ is correlated over time and the correlation between two (different) time points is constant and takes the following equi-correlation structure:

$$\rho = corr\,(u_{it}, u_{is}) = \frac{\mathrm{s}_{\alpha}^{2}}{\mathrm{s}_{\alpha}^{2}\ + \mathrm{s}_{\varepsilon}^{2}} \tag{4.5}$$

with $t \neq s$ and $t, s = 1, 2, 3$. As the outcome variable is dichotomous, a normalisation of ε_{it} is required. We take $\varepsilon_{it} \sim N\,(0, 1)$ and the outcome probability is:

$$P_{it}\,(\alpha *) = \mathbf{\Phi}\left[(\sum_{r=2}^{4} \gamma_r y_{it=0}^{r}\,(y_{it-1} = 0) + \sum_{r=1}^{4} \delta_r y_{it=0}^{r}\,(y_{it-1} = 1) + x'_{it=0}\beta + a_0 + \mathrm{s}_{\alpha}^{2}\ \alpha^{*})\,(2y_{it} - 1) \right] \tag{4.6}$$

Note that $\mathbf{\Phi}\,[\bullet]$ refers to the cumulative standard normal distribution. The likelihood function is the product of all time-point specific probabilities across all individuals. Namely,

$$L = \prod_{i=1}^{N} \int_{\varphi^{*}} \left\{ \prod_{t=1}^{T} P_{it}\,(\alpha^{*}) \right\} \mathrm{d}F\,(\alpha^{*}) \tag{4.7}$$

where F is the distribution function of $\alpha^{*} = \alpha / \mathrm{s}_{\alpha}$. Equation (4.7) does not have a closed-form, and therefore α has to be integrated out. As we assume that α is normally distributed, the integral can be evaluated using Gaussian-Hermite quadrature. All the equations are estimated using Gauss-Hermite quadrature (Butler et al., 1989).

4.5.4 Subsample estimations

As we outlined in section 4.2.2, we expect different (potentially conflicting) effects for those who are initially employed compared to those who are initially unemployed. Therefore, we also estimate separate regressions based on the labour force status at $y_{it=-1}$. For those initially employed, Equation (4.4) changes to:

$$y_{it} = \mathbf{1}\left(\gamma_2 y_{it=0}^{2}\,(y_{it-1} = 0) + \sum_{r=1}^{2} \delta_r y_{it=0}^{r}\,(y_{it-1} = 1) + x'_{it=0}\beta + a_0 + \alpha_i + \varepsilon_{it} > 0 \right) \tag{4.8}$$

Where the reference category is the continuously employed ($y_{it=0}^{1}$) who were employed at $t-1$ ($y_{it-1} = 0$) is. For those initially unemployed it changes to:

$$y_{it} = \mathbf{1}\left(\gamma_3 y_{it=0}^{3}\,(y_{it-1} = 0) + \sum_{r=3}^{4} \delta_r y_{it=0}^{r}\,(y_{it-1} = 1) + x'_{it=0}\beta + a_0 + \alpha_i + \varepsilon_{it} > 0 \right) \tag{4.9}$$

Where the reference category is the short-term unemployed ($y^4_{it=0}$) who were employed at $t-1$ ($y_{it-1}= 0$).

4.5.5 Including testosterone as a covariate

To ensure comparability across groups, we adjust the circulating testosterone levels for age and time of the day when the blood sample was taken. We use two different approaches to that end:

In the regression model, we include the absolute level of testosterone (nmol/l) as a covariate and control for the hour of the nurse visit (Model 1). In a further specification, we include the absolute level of testosterone as a second-degree polynomial (Model 2). First, we use the Health and biomarkers Survey to construct a sample of men with a positive level of testosterone in the age range 16 to 70 who had their interview started between 9 am and 8 pm ($N = 4,605$). We utilise an OLS model to estimate the deviation from the time- and age-corrected mean.[13] Second, we order the distribution of the deviation and form three groups: (i) low level of testosterone if the deviation belongs to the lowest decile, (ii) medium level of testosterone if the deviation is in 2^{nd} to 9^{th} decile, and (iii) high level of testosterone if the deviation belongs to the highest decile (Model 3).

In a robustness check, we re-estimate our regression specifications adjusting the cut-point defining low, medium and high levels to ensure our results remain stable and are not driven by these definitions.

4.5.6 Observable characteristics and individual-specific effects

Based on our modelling setup, two potential sources of bias may affect our results: (i) unobserved heterogeneity caused by individual-specific differences and (ii) the correlation of the unobserved heterogeneity with the initial conditions (Heckman, 1981). If unobserved heterogeneity is present and persists over time, this will lead to an overstatement of true state dependence (Stewart, 2007). The modelling framework outlined in section 4.4.1 controls for unobserved heterogeneity, which we assume follows a specific distribution. In order to address the initial conditions problem, we follow the approach of Wooldridge (2005). By construction, the model in section 4.4.1 assumes that the covariates used in the regression analysis and the random effect error term are uncorrelated. For example, we rule out a correlation between testosterone and (unobserved) ability (which would be captured in the error term). Wooldridge (2005) addresses the initial conditions by

[13] We include time as a categorical variable on the full hour. Age is included in a linear version (in a robustness check, we included age as a second degree polynomial, but results remain similar).

extending the so-called Mundlak-Chamberlain approach. In this case, one specifies an approximation for the individual unobserved time-invariant heterogeneity *given* the initial conditions, where we also condition on exogenous variables likely to be correlated with the unobservable component. This then allows for correlation between the initial observation (labour force status in our case) and the unobservable individual effects. This approach hinges on the auxiliary conditional distribution for the unobserved heterogeneity to be correctly specified (Wooldridge, 2005). In our case, the individual level (time constant) unobserved effect is a function of the initial labour force status, the mean of time-varying covariates and an individual specific error term. Thus, the specification changes to:

$$y_{it} = \mathbf{1}(\alpha_1 y_{it-1} + x'_{it}\beta + \theta_i + \varepsilon_{it} > 0) \tag{4.10}$$

Note that in order to provide sufficient variation within an individual, we use up to 5 observations per individual (therefore, the number of time-points we consider is larger than in the base specification). To account for the Mundlak specification, we specify:

$$\theta_i = \sum_{r=2}^{4} \lambda_r y^r_{it=0} + \overline{x}'_i a + a_0 + \eta_i \tag{4.11}$$

Where $\overline{x}_i$ refers to the time-mean of the observable characteristics of the dynamic sequence ($t \geq 1$) (see also Akay, 2012; Rabe-Hesketh and Skrondal, 2013). Inserting Equation (4.11) into Equation (4.12) leads to:

$$y_{it} = \mathbf{1}(\alpha_1 y_{it-1} + x'_{it}\beta + \sum_{r=2}^{4} \lambda_r y^r_{it=0} + \overline{x}'_i a + a_0 + \eta_i + \varepsilon_{it} > 0) \tag{4.12}$$

Following the concept in the basic specification, we extend Equation (4.12) by interacting the lagged labour market position with the initial period position. This leads to our final specification:

$$y_{it} = \mathbf{1}\left(\sum_{r=2}^{4} \gamma_r y^r_{it=0}\,(y_{it-1} = 0) + \sum_{r=1}^{4} \delta_r y^r_{it=0}\,(y_{it-1} = 1) + x'_{it}\beta + \overline{x}'_i a + a_0 + \eta_i + \varepsilon_{it} > 0\right) \tag{4.13}$$

It is important to note that if the estimation results following this approach are similar to those based on our basic specification, then this implies that by not accounting for time-varying means of the covariates the random effects assumption is likely to hold. Put another way, the basic framework controls for much of the individual level heterogeneity and therefore the approach of Wooldridge (2005), which primarily deals with the residual heterogeneity between covariates and the error term by incorporating the mean of particular covariates (relative to our basic framework), should not change the main conclusions drawn from our baseline estimates.

4.6 Descriptive statistics

The economic literature has shown that unemployment risk is influenced by factors like the qualification, age, health etc. In our study, we test the explanatory power of testosterone. When we split the sample into initially unemployed and employed (column three and four of Table 4.9, we can see that these groups differ with respect to observable characteristics. For example, among the initially unemployed is a significantly higher share of individuals with no qualification or in poor health. However, we also find that there are significant differences in the distribution of testosterone. The sample of initially unemployed has, on average, a higher level of testosterone, and the difference is statistically significant. One explanation is that due to time restriction of work, employed individuals provide their blood sample later in the day (not shown), and as levels naturally decrease over the days, they have a lower level of testosterone, on average. However, we still find a higher share of individuals with a testosterone level in the 10^{th} decile (as well as a higher share of individuals in the 1^{st} decile).

4.7 Mechanisms

We showed that testosterone affects men's transitions in and out of employment. Now we examine whether these transitions can be explained by observed behaviour among individuals and personality traits. In section 4.2.2, we discussed potential channels through which testosterone may affect an individual's employment status. Although we cannot provide conclusive evidence for these mechanisms in this study due to data limitations, we consider several potential channels in a descriptive analysis to examine whether the associations in our dataset are consistent with the literature.

For example, the numerical ability was related to testosterone in experimental studies (Brosnan et al., 2011), and it was also collected during wave 3 in the UKHLS mainstage interview.[14] Survey respondent's practical numerical knowledge is assessed by testing whether they can understand percentages and fractions in typical real-life settings. Such ability measures have been shown to be highly related to wealth (McArdle et al., 2009; McFall, 2013). Individuals are presented with three initial problems, and if none

[14] This is approximately 7 months after the wave 2 nurse visit and 5 months before the wave 3 nurse visit, and hence relatively close to the date circulating testosterone is measured.

Table 4.9: Descriptives at nurse visit

	Full Sample	Initially unemployed	Initially employed	t-test (p-value)
Testosterone (nmol/l)	15.24 (5.60)	16.67 (6.89)	15.15 (5.50)	0.0057
Testosterone (categorical)				
1^{st} decile	10.59	14.68	10.33	0.1525
2^{nd} – 9^{th} decile	80.16	70.64	80.75	0.0102
10^{th} decile	9.26	14.68	8.92	0.0441
Age	43.7(8.0)	43.1(11.8)	43.8(9.8))	0.5279
Highest qualification				
Degree	29.95	16.51	30.77	0.0016
Other higher degree	11.76	10.09	11.86	0.5787
A-level etc	23.62	17.43	24.00	0.1173
GCSE etc	22.07	28.44	21.68	0.0988
Other qualification	8.4	11.93	8.19	0.1722
No qualification	4.2	15.60	3.50	0.0000
General Health				
excellent	17.82	9.17	18.35	0.0151
very good	40.59	32.11	41.11	0.0634
good	28.94	33.94	28.63	0.2350
fair	11.28	20.18	10.73	0.0024
poor	1.38	4.59	1.19	0.0031
Region of residence				
England	84.52	95.41	83.85	0.0012
Wales	6.7	2.75	6.95	0.0894
Scotland	8.78	1.83	9.20	0.0083
Rural area	22.18	13.76	22.70	0.0293
Number of people in household				
1	13.35	30.28	12.31	0.0000
2	28.14	23.85	28.4	0.3056
3	20.96	16.51	21.23	0.2405
4+	37.55	29.36	38.06	0.0688
Long-standing illness or disability	24.84	30.28	24.51	0.1762
Legal marital status				
single	26.81	54.13	25.13	0.0000
married	61.60	28.44	63.64	0.0000
separated/ divorced/ widowed	11.60	17.43	11.24	0.0500
N	1,880	109	1,771	-

Notes: Author's own calculations, using data from the Understanding Society subsample Health and biomarkers Survey. $N = 1,880$.

are answered correctly a further (simple) question is asked. On the other hand, if all questions are answered correctly, then an additional (more difficult) question is asked; if this was also answered correctly, a further final question is asked.[15] Thus, an individual's final score is between zero (no correct answers) and five (all correct answers) and a clear ordering exists. Regression results (see Table D.1 in the appendix) show that relative to individuals with a low testosterone level, the log odds of reporting a higher test score are 1.29 times higher (significant at the 5% level) among those with a medium level of testosterone.[16]

Alongside numerical ability, Understanding Society assesses an individual's fluid reasoning using logic puzzles (number series). Such measures have been found to be related to individuals financial knowledge (Delavande et al., 2008) and are negatively associated with age (Salthouse, 2010). [17] We, therefore, control for age in the regression analysis. Individuals in households were randomly allocated to a set of questions.[18] Within each 'set', individuals were asked six questions. An individual's final score ranged between zero (no correct answers) and six (all correct answers). Regression results **Table A.5** in the appendix) show that relative to individuals with a low testosterone level, the log odds of reporting a higher test score are 1.28 times higher (significant at the 5% level) among those with a medium level of testosterone. It is important to note that, given our main sample follows individuals aged between 20 and 60 years old when initially observed, one could argue that individuals underlying ability is relatively stable across time (as opposed to ability at very young ages). Indeed, this is one of the underlying assumptions we make when controlling for initial conditions and time-invariant unobserved heterogeneity following Wooldridge (2005). Thus, even though these differences in numerical ability and fluid reasoning are in line with the literature on testosterone, they are unlikely to be the mechanism through which testosterone affects employment status in our base model as we control for such time-invariant unobserved heterogeneity.[19]

A similar line of reasoning applies to occupational class. Given that occupational class is likely to be time-invariant (at least in the short panel considered in this paper), it is unlikely to drive our results. Moreover, in our data, we do not observe an association

[15] This test was adopted from the English Longitudinal Study of Ageing and some parts have also been included in the US Health and Retirement Study and Survey of Health Ageing and Retirement in Europe.

[16] Low level is defined as bottom quintile of deviation from mean, medium level is between second and fourth quintile and high level is top quintile.

[17] This test was developed for use in the US Health and Retirement Study.

[18] We only analyse the relationship between testosterone and responses to 'Set 1' as there was a CAPI coding error in 'Set 2'. See McFall (2013) for further details.

[19] Our robustness checks show that (i) our main results hold even when restricting the sample to those aged at least 25 and (ii) if we assume individual's ability is appropriately captured by the specification described in section 4.4.3, and hence is time constant then our main findings hold after controlling for such unobserved factors.

between occupational class and testosterone levels, which earlier studies have found (Dabbs, 1992).

Research suggests that males with higher levels of testosterone are more likely to express certain personality traits and behaviours in social and professional situations (Greene et al., 2014; Dabbs et al., 2001). These same traits may help such individuals overcome adverse situations, such as unemployment. Understanding Society fielded a General Health Questionnaire at wave 3 which included attitudinal questions relating to whether individuals felt they have recently been losing confidence in themselves and, separately, whether individuals feel they have recently been able to face up to problems. In this case, individuals with medium levels of testosterone were significantly more likely to report a response which suggested they were had not lost confidence or the ability to face problems, compared to individuals with low levels of testosterone. We also consider risk-taking, which has been associated with high testosterone levels (Apicella et al., 2008; Coates and Herbert, 2008). Respondents in Understanding Society were asked to rate their willingness to take general risks on a scale between 0 and 10, where higher values indicate a greater willingness to take risks. In a regression model controlling for age and log earnings, we found a positive and statistically significant association between being in the high testosterone group and reporting a higher score (OR=1.23*).

Individual's behaviour is also strongly correlated to their personality. For example, one might expect that individuals with higher levels of testosterone are willing to search more intensely for a job ceteris paribus. In response to a question about job search in the last 4 weeks and asked to individuals who did not report being in paid work in the last week or having a job, those with medium level of testosterone were more likely to report using the internet to search for a job compared to unemployed individuals belonging to the low testosterone group. In addition, individuals were also asked about whether they used their network to explore employment opportunities. Based on purely descriptive evidence, the data suggest a higher proportion of the high testosterone group mentioned such a strategy, however, this result was not statistically significant. We also examine individuals' self-reported likelihood to lose their job in the next 12 months (very unlikely, unlikely, likely or very likely), and find a strong positive association between those individuals who belong to the high testosterone group and the likelihood of job loss (OR=1.37**, controlling for age and log earnings).

In summary, the associations found in Understanding Society are in line with earlier studies, showing that testosterone is positively associated with numerical ability and cognition as well as personality traits such as risk-taking and self-confidence. Moreover, we also find some descriptive evidence for differences in job search behaviour. While we cannot conclusively prove that these potential mechanisms explain the observed relationship between testosterone levels and unemployment, we interpret our descriptive findings as suggestive evidence that such mechanisms are likely to play a role.

4.8 Conclusion

This paper examines the relationship between testosterone levels and unemployment dynamics among men in the UK. Based on existing studies on testosterone and individual behaviour, we expect that individuals with higher testosterone levels are more likely to exit unemployment, but employed individuals with high testosterone levels are at a higher risk of entering unemployment. The results from our dynamic random effects labour market model confirm these expectations. Among initially unemployed individuals, those with medium and high testosterone levels are significantly more likely to leave unemployment compared to those with testosterone levels in the lowest decile. In contrast, for our sample of initially employed men, those with testosterone levels above the 9^{th} decile were more likely to enter unemployment than those with medium and low levels of testosterone.

Descriptive evidence suggests that these mechanisms might be driven by differences in personality traits and job search behaviour as well as occupational sorting. While "aggressive" behaviour such as competition-seeking or dominance might put individuals with high testosterone levels at an advantage during their job search, these same traits might prove detrimental to remain in employment. Moreover, individuals with high testosterone levels tend to choose occupations in which spells of unemployment and re-employment are more common. In contrast, men with lower testosterone levels tend to prefer stable and secure occupations. Our findings have important implications for labour market policy. They demonstrate that latent biological processes can affect job search behaviour and labour market outcomes, without necessarily relating to illness and disability. While it would surely be impractical to determine testosterone levels of unemployed men to improve their labour market outcomes, such differences can still be taken into account. For example, when considering how much assistance should be provided to job seekers (or whether sanctions should be applied), it is important to recognise that some differences in job search behaviour are driven by biological processes outside the control of the job seeker. Hence, some individuals might require more assistance than others. One specific example could be training programmes, where individuals with lower testosterone levels might benefit more from individual coaching rather than group sessions. Our results also suggest that individuals with high testosterone levels are at an advantage during the job search, although such hormonal differences do not necessarily translate into better productivity. Awareness of the impact of personality and behavioural traits on performance during job interviews can potentially improve the quality of the job match.

While our results are robust to a wide variety of specification changes, including approaches to account for unobserved individual heterogeneity and initial conditions, there are nevertheless some limitations. Most importantly, the causality of our findings is not clear. We control for time-invariant unobserved heterogeneity, and the dynamic structure of our models should ensure that testosterone levels were measured before changes in employment status. Nevertheless, our models assume that the levels of testosterone (adjusted for age and time of day) remain broadly stable. Unfortunately, our data does not allow us to test this assumption since only one measurement of testosterone is available for each individual. Repeated measures of testosterone could be used to examine whether this assumption holds, as well as if and how testosterone levels change during labour market transitions. In a future extension of the paper, we plan to draw on genetic predictors of testosterone levels to identify random variation in testosterone levels that remains stable over the life course. These genetic predictors will be used as instrumental variables in aian Randomization design to assess the causality of the findings from our dynamic labour market model. Moreover, we recommend that future research should examine the long-term cumulative effects of testosterone levels on labour market outcomes. Finally, it would be worthwhile to study the mechanisms for which suggestive evidence was presented in this paper in more detail and determine whether they extend to women as well.

Chapter A

Appendix to Chapter 1

A.1 Calculation of inequality and poverty indices based on different equivalence scales

Table A.1: Measurement of inequality and poverty.

Approach	Gini coefficient	At-risk-poverty rate (%)	Interquartile range (Euro)
Quadratic expenditure system	0.23	14.6	3623
Quadratic almost ideal demand system	0.24	14.6	3900
Semiparametric (modified)*	0.23	14.9	3886
Matching	0.23	14.4	3760
Modified OECD scale	0.25	14.8	3964
Square root scale	0.24	14.8	4221

Note: This observation is based on income data from the EVS 2013, while using the more plausible equivalence scales from Table 6. The scales are partly based on 2003, 2008, and 2013 data.

In Table A.1, we present our findings on the degree to which different equivalence scale estimates influence the measurement of inequality and poverty. To calculate equivalence income, each of the more plausible equivalence scales was applied to EVS household income data in 2013 (matching; QAI), with the exception of the nonparametric approach, as its interval estimates were not well-suited for this exercise. We also added two equivalence scales that were less plausible (QES, Stengos et al., 2006), but that displayed equivalence scale elasticities close to those of the plausible estimates. The modified OECD scale and the square root scale were also applied. In a second step, equivalence income was used to calculate three commonly used indicators: the Gini coefficient, the at-risk-of-poverty rate (ARP), and the interquartile range (IQR). As the estimation was done without additional weighting, these values indicated the household level.

When using the modified OECD scale for this calculation, we obtained values of 0.25 for the Gini coefficient, 14.8% for the ARP rate, and around EUR $4,000$ for the IQR.[1] The equivalence scales of all of the methods shown in Table A.1 generated very similar findings. For instance, for the equivalence scale obtained from the Quadratic Almost Ideal Demand System, we calculated values of 0.24 for the Gini coefficient; 14.9% for the ARP rate, and around EUR $4,200$ for the IQR. In contrast, the methods with implausible equivalence scales led to deviating results (not shown in Table A.1). For instance, the equivalence scale of the AI demand system generated a Gini coefficient of 0.27, an at-risk-poverty rate of 19%, and an IQR of EUR 6,300.

Overall, we conclude that applying our plausible equivalence scales leads to consistent assessments of inequality and poverty. We also observe that applying less plausible scales seems to lead to similar results if their equivalence scale elasticities are similar.

[1] When applied, the weights for children differ. This is considered in the application. The household-specific OECD scale can vary by 0.2 per child depending on the ages of the household members.

A.2 The QAI demand system and income independence

Figure A.1: Income independence test based on the QAI (Banks et al., 1997), Household type single (A)

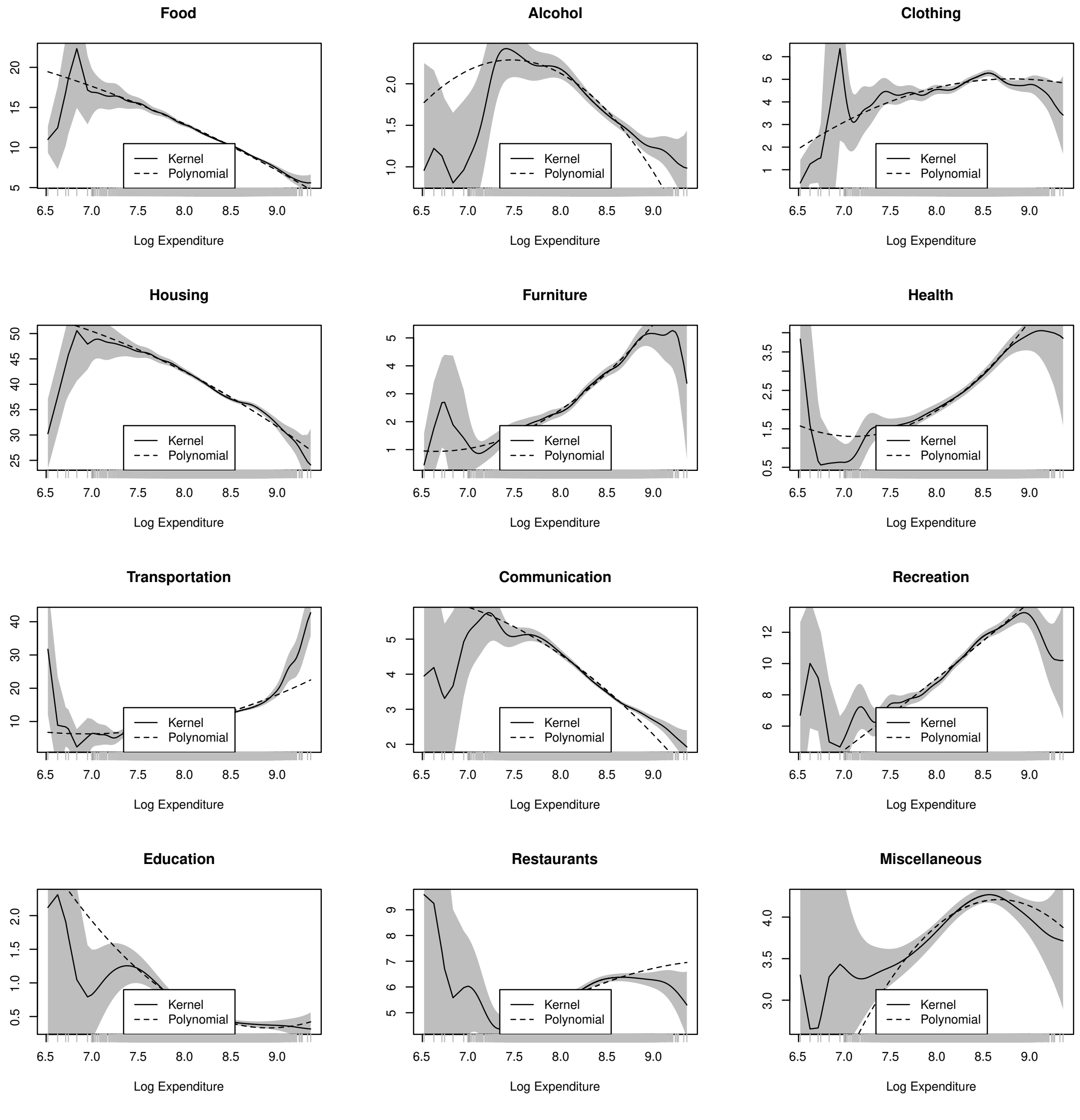


Figure A.2: Income independence test based on the QAI (Banks et al., 1997), Household type couple (AA)

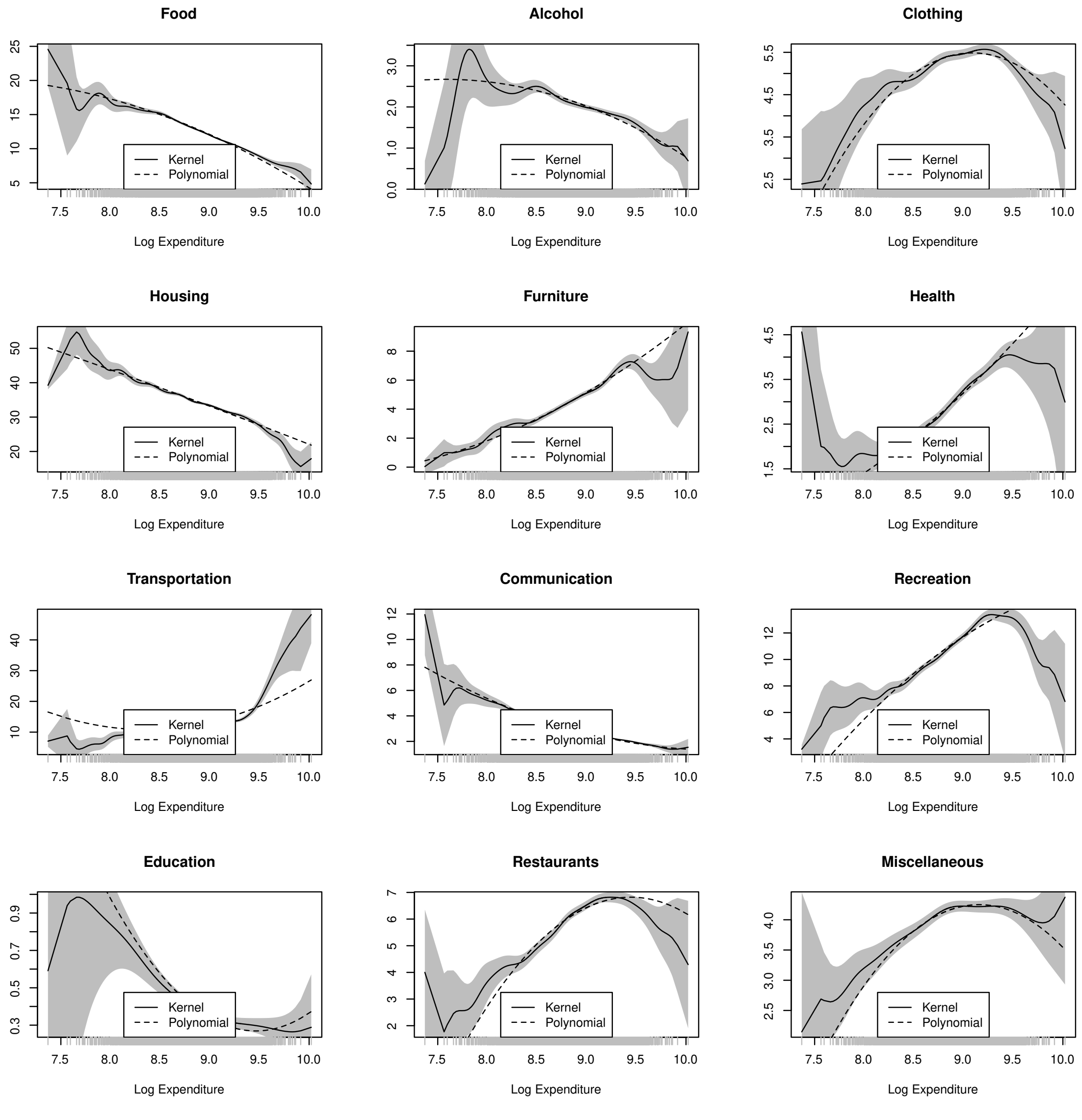


Figure A.3: Income independence test based on the QAI (Banks et al., 1997), Household type couple with one child (AAC)

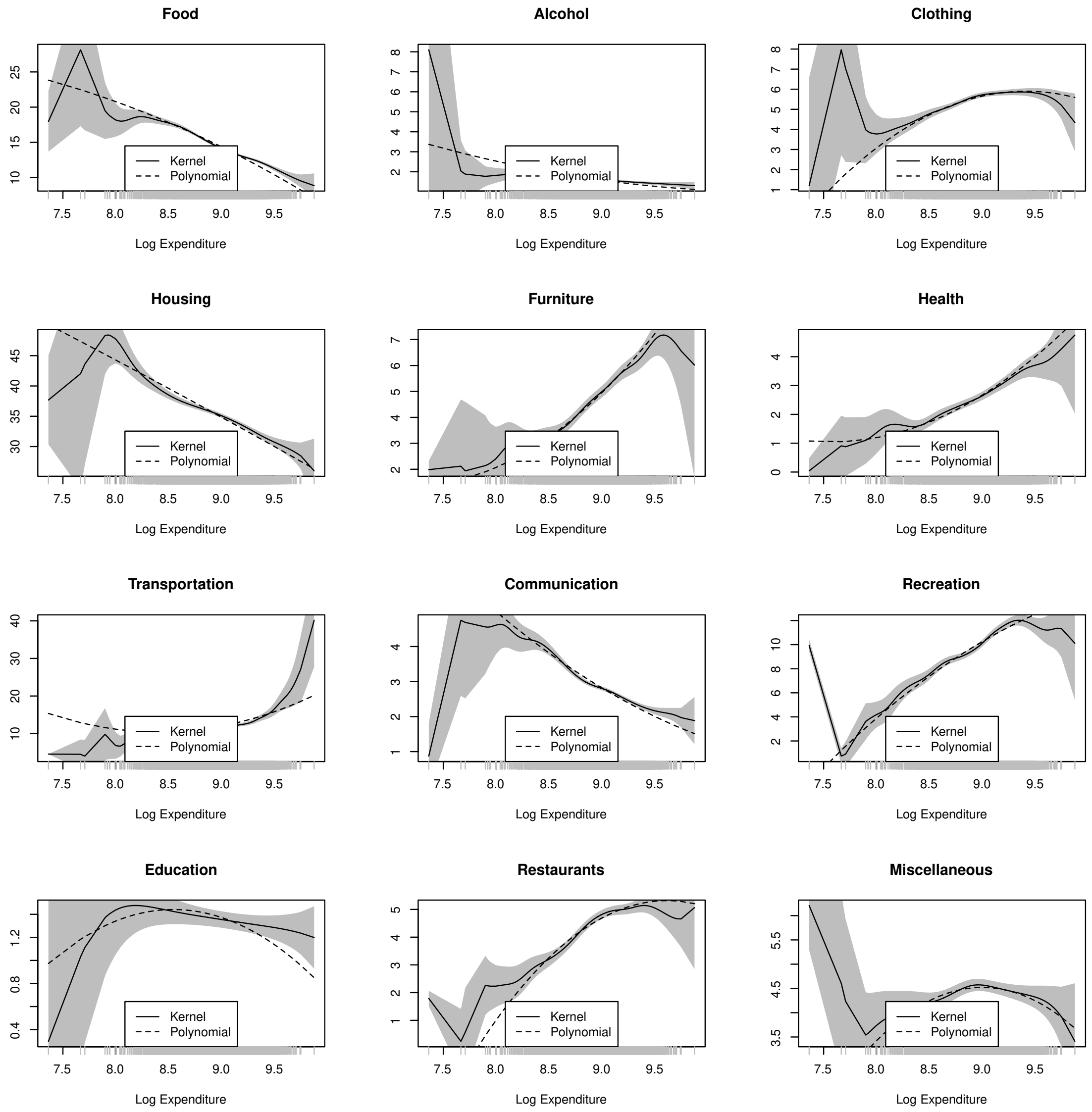


Figure A.4: Income independence test based on the QAI (Banks et al., 1997), Household type couple with two children (AACC)

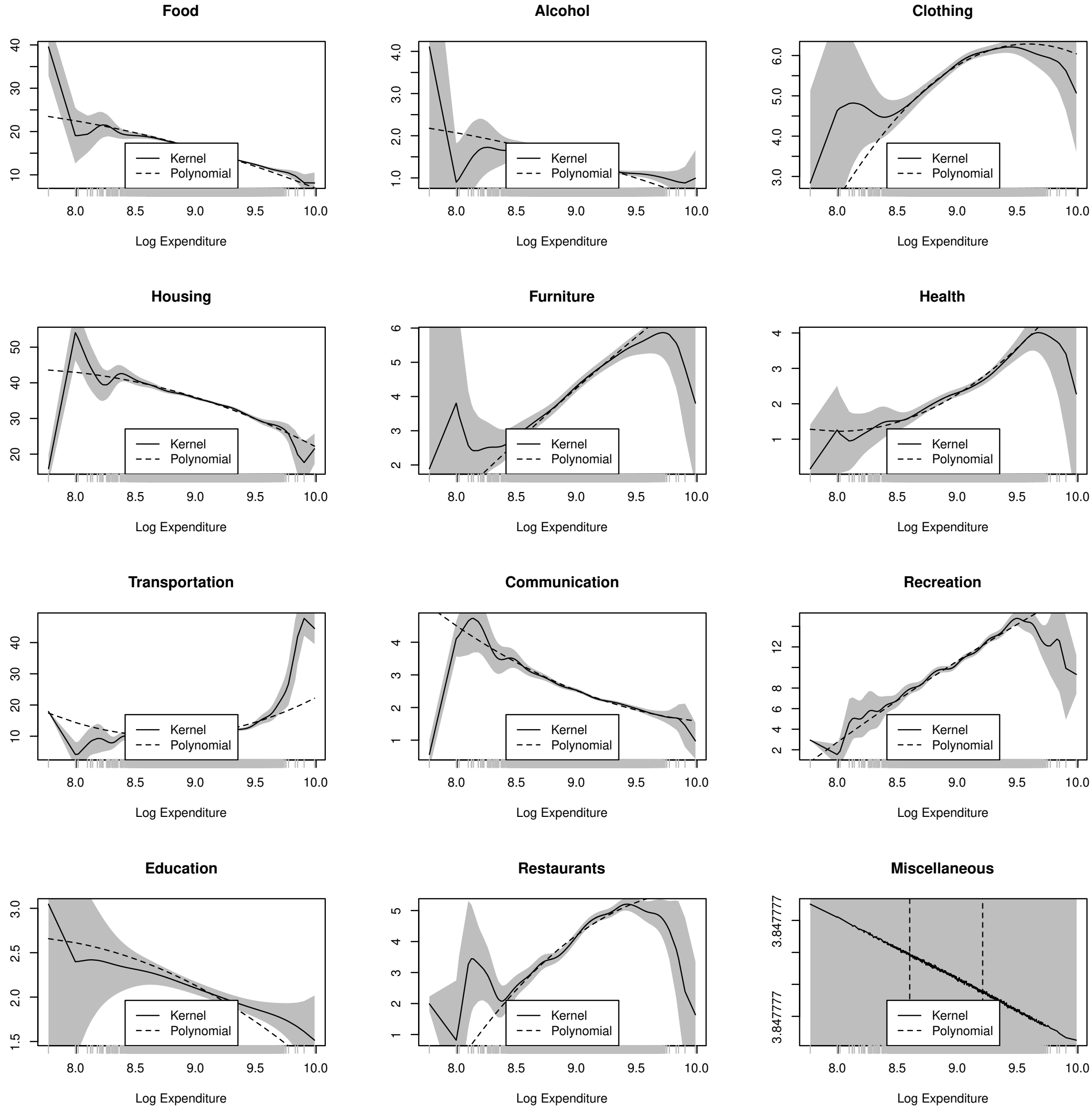
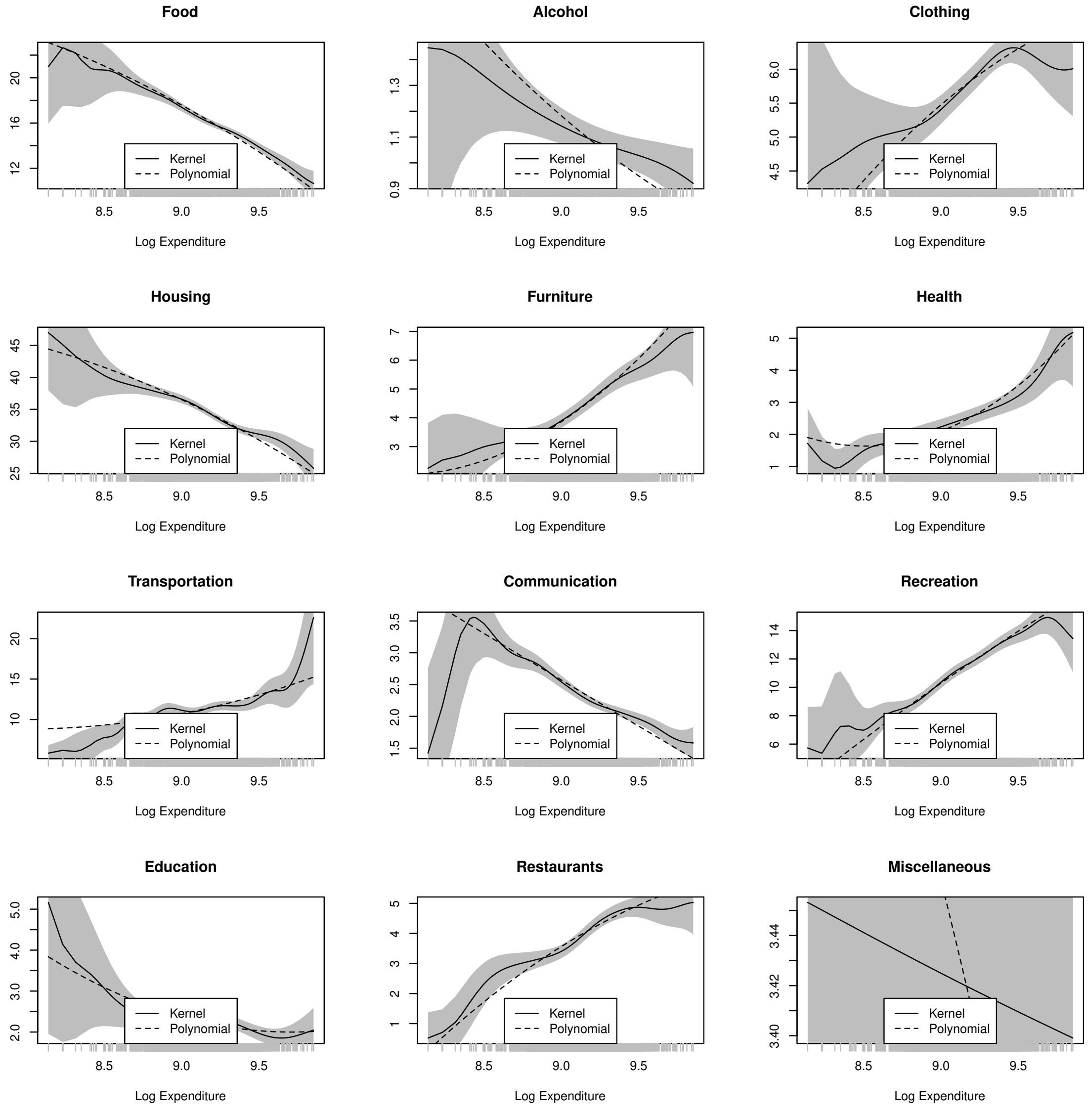


Figure A.5: Income independence test based on the QAI (Banks et al., 1997), Household type couple with three children (AACCC)



A.3 Calculation of confidence intervals for nonparametric bounds

If (L_s, U_s) denote the bounds resulting for the sth bootstrap replication, then we choose the bounds l and u of the confidence interval, such that $l < L_s$ and $U_s < u$ for 95% of the bootstrap replications. As l and u usually will not be unique, we choose the values of l and u for which the interval width, $u - l$, is smallest. To calculate the smallest interval, an iterative procedure is used and re-run several times. Each run takes the 5% percentile of the lower bound L and the 95% percentile of the upper bound U as starting values, perturbed with noise $e_L \sim \mathcal{N}(0, \mathrm{sd}(L))$ or $e_U \sim \mathcal{N}(0, \mathrm{sd}(U))$, respectively. Let the resulting bounds be denoted by $L^{(0)}$ and $U^{(0)}$. The coverage achieved with these values is equal to $\rho^{(0)}$. If $\rho^{(0)}$ is smaller than $1 - \alpha$, $L^{(0)}$ and $U^{(0)}$ are decreased and increased, respectively, by a stepsize $\lambda_L = 0.1\mathrm{sd}(L)$ or $\lambda_U = 0.1\mathrm{sd}(U)$ to get new values: $L^{(1)} = L^{(0)} - \lambda\epsilon_L^{(0)}$ and $U^{(1)} = U^{(0)} + \lambda\epsilon_U^{(0)}$, where $\epsilon_U^{(0)}$ and $\epsilon_L^{(0)}$ follow a uniform distribution. If $\rho^{(0)}$ is larger than $1 - \alpha$, the signs for λ are instead changed to decrease the interval width. $\rho^{(1)}$ is the coverage achieved after these adjustments. Depending on whether it is above or below $1 - \alpha$, the adjustments are applied to obtain updated values $L^{(2)}$ and $U^{(2)}$; $\rho^{(2)}$ is checked against $1 - \alpha$ again, etc.; until $\rho^{(k)} = 1 - \alpha$. This procedure is re-run for 100 different starting values and the interval with the smallest width is reported.

Chapter B

Appendix to Chapter 2

B.1 Descriptive statistics

Table B.1: The Health and Retirement Study 2014, summary sample statistics by retirement status

	Pre-retirees	Retirees
Retirement status defined by:		
1) Labor force status from RAND, n	423	710
2) as 1) but obs. working >20hrs are treated as pre-retired, n	449	684
3) Whether or not receive any pension income, n	828	291
4) Self reported retirement status, n	370	734
...applying definition 2)		
Welfare indicators		
Food share narrow (%), mean	14.3	17.4
Food share wide (%), mean	14.0	16.6
Monthly expenditures for nondurables (in 1000 Dollar), mean	17.7	11.6
Income satisfaction, mean	3.1	3.2
Income concepts		
Monthly net income (in 1000 Dollar), mean	2.5	2.0
Monthly net income (in 1000 Dollar), min	0.1	0.1
Monthly net income (in 1000 Dollar), max	6.8	7.2
w/o annuities w/o housing (in 1000 Dollar), mean	2.3	1.7
w/o housing (in 1000 Dollar), mean	2.4	1.9
w/t imputed rent (in 1000 Dollar), mean	2.8	2.2
Covariates		
Age (years), mean	63.2	64.6
Share of males (%)	34.5	34.6
Share of homeowner (%)	54.8	49.4
Share of highly educated (%)	32.1	27.9
Share of non-whites (%)	41.2	43.1

Table B.2: The German Income and Expenditure Survey 2013, summary sample statistics by retirement status

	Pre-retirees	Retirees
Retirement status defined by:		
1) Self reported labor force status, n	807	1510
2) as 1) but obs. working >20hrs are treated as pre-retired, n	816	1501
... applying definition 2)		
Welfare indicators		
Food share narrow (%), mean	12.2	14.6
Food share wide (%), mean	15.8	18.2
Monthly exp. for nondurables narrow (in 1000 Euro), mean	0.5	0.5
Monthly exp. for nondurables wide (in 1000 Euro), mean	0.7	0.6
Income concepts		
Monthly net income (in 1000 Euro), mean	2.4	1.8
Monthly net income (in 1000 Euro), min	0.2	0.3
Monthly net income (in 1000 Euro), max	11.5	10.8
w/o annuities w/o housing (in 1000 Euro), mean	2.3	1.6
w/o housing (in 1000 Euro), mean	2.3	1.7
w/t imputed rent (in 1000 Euro), mean	2.6	2.0
Covariates		
Age, mean	61.9	65.6
Share of males (%)	31.5	29.5
Share of former GDR (%)	21.8	25.9
Share of homeowner (%)	45.5	43.4
Share of highly educated (%)	33.9	24.5

B.2 Replacement rates using gross income

As discussed in the main text, replacement rates based on gross income can potentially differ from those based on net income. To assess this, results for replacement rates based on gross income (gross replacement rates; GRR) are shown in Table B.3. NRRs are shown for comparison. Gross income here is defined as monthly income from work and pensions before taxes, plus annuitized wealth and housing wealth, as described in the main text for net income. The analysis is based on the samples of single-person households in the age range from 60 to 69.

Table B.3: Results comparing estimates based on net income and gross income, for the U.S. and Germany

USA	NRR	GRR
Parametric	0.912	0.924
Semiparametric	1.052	1.014
Nonparametric	[0.888,1.194]	[0.837;1.169]
Germany		
Parametric	0.968	0.980
Semiparametric	1.076	1.001
Nonparametric	[0.863,1.061]	[0.827;1.037]

Data: HRS 2014 and EVS 2013.

The difference between GRR and NRR are overall not large, and GRR estimates are rather close to our main findings. Results on income independence do also not differ (details available upon request from the authors). A potential reason for this is that both gross income and net income account for annuitized wealth and housing wealth in a similar way. Moreover, given the same gross income level, differences in taxation of retirees and non-retirees are not very large, at least not in our sample.

B.3 Sensitivity analysis for the main variables: Income, welfare indicator, and retirement status

B.3.1 Variables: Welfare indicators

For the main results presented in the text, we use the expenditure share of food at home as the welfare indicator. Expenditure for food at home is known to be only an approximate measure of household consumption (Bernheim et al., 2001). Apart from not capturing food expenditure away from home at restaurants, at work, etc. this indicator also might mismeasure consumption as expenditure might be substituted with time for home production, while consumption stays constant (Aguiar and Hurst, 2005; Luengo-Prado and Sevilla, 2013).

Because of this, we use several other indicators to assess the robustness of our results. This includes the income share of expenditure for food at home and away from home; the absolute level of expenditure for nondurable goods defined in several ways; and satisfaction with household income. Expenditure for nondurable goods has been argued to give a more complete picture of consumption (Battistin et al., 2009). Satisfaction with household income differs from the expenditure based indicators in that it is a subjective measure, assumed to capture a subjective assessment of household welfare (van Praag, 1991). While satisfaction with household income is a qualitative variable, we follow earlier literature (Ferrer-i Carbonell and Frijters, 2004; Dudel et al., 2016) and model it similar to the other welfare indicators and thus as a metric variable.

While the expenditure based measures are readily available for the EVS, this is not the case for the HRS. Information on food away from home is collected, but excludes expenditures for food at work or school, making this variable potentially incomplete and not directly comparable to the EVS data. Nondurable expenditures are not covered in the HRS itself, but in the Consumption and Activities Mail Survey (CAMS), which was collected in 2015 after the 2014 HRS wave. This data is only available for a part of the HRS sample, leading to a small number of observations. Satisfaction with household income was covered as part of the HRS 2014 survey, but only in a leave-behind questionnaire administered to half of the respondents. This leave-behind questionnaire was not answered by all respondents, also leading to a small number of observations.

For all expenditure-based welfare indicators we removed some outliers. For both the EVS and HRS we excluded households from the analysis if the income share of food was above 70%, or if the amount spent on nondurables amounted to more than 80% of household income.

B.3.2 Variables: Income concepts

For our main analysis, we used the net income plus annuitized wealth and annuitized housing wealth. While this income concept can easily be implemented for both the German and the U.S. data, other income concepts could be used instead (for a discussion see Munnell and Soto, 2005; Crawford and O'Dea, 2012). For instance, housing wealth might not generate steady income, and could be illiquid (Angelini et al., 2014).

To assess the sensitivity of our results with respect to the income concept use, we conduct robustness checks using net income excluding wealth; net income plus annuitized non-housing wealth, i.e., excluding housing wealth; and net income plus annuitized non-housing wealth plus imputed rent instead of housing wealth. Imputed rent is readily available for the EVS data as calculated by the Federal Statistical office, while for the HRS we generate imputed rent assuming a 5% rental yield (Munnell and Soto, 2005; Crawford and O'Dea, 2012).

B.3.3 Variables: Retirement status

For our main analysis, for the US we used the labor force status as defined by RAND in combination with the number of weekly hours worked. In the sensitivity checks, we use three variants: only using the RAND indicator, without dropping individuals who work more than 20 hours per week from the group of retirees; setting retirement status based on receipt of any pension income, including pension plans and annuities; and self-reported retirement status, i.e., whether respondents consider themselves to be retired or not.

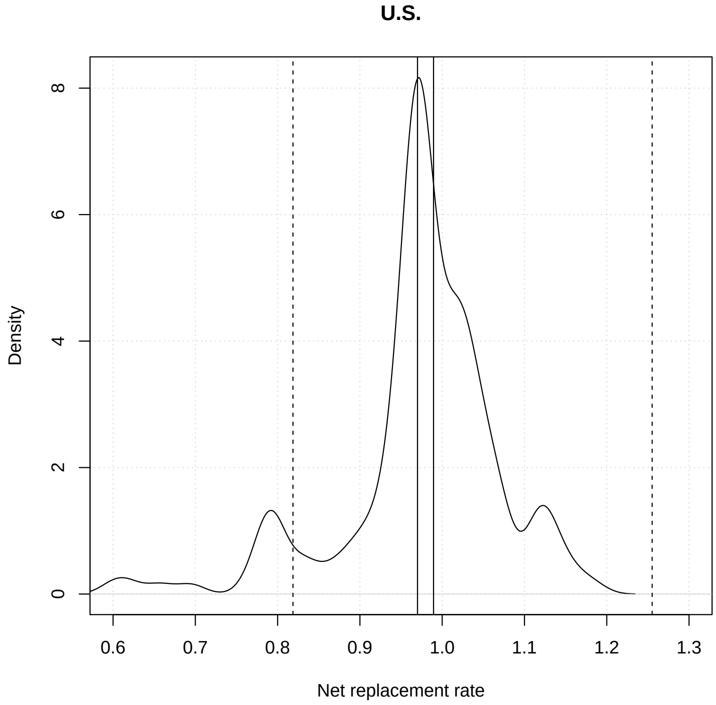
For Germany, in the main analysis we used the labor force state as defined by the Federal Statistical Office and the number of hours worked to define whether an individual is retired. In the sensitivity analysis, we drop the hours worked per week.

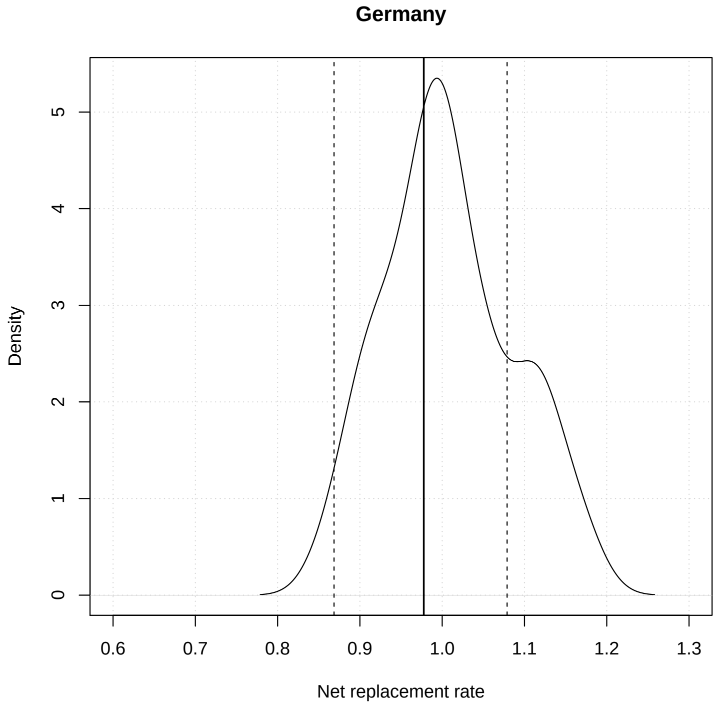
B.3.4 Results and discussion

The robustness checks are summarized in Figure B.1. It shows kernel densities based on 157 net replacement rate estimates for the US and 58 estimates for Germany, all either based on the parametric or semiparametric approach. Each of these models is based on a different combination of welfare indicator, income concept, and retirement status. For some, the semiparametric approach exhibited convergence issues (3 for the US, 2 for Germany). The point estimates of the parametric and semiparametric approach of our main results are shown as solid lines, and the identification bounds as dashed lines.

For the US, 80% of all estimates fall within the identification bounds of our main results. The exceptions to this form two groups. The first group are estimates based on nondurable expenditure, which are sometimes below and sometimes above the identification bounds. And the second groups are estimates based on satisfaction with household income, which in some cases are lower than expenditure based estimates. The latter might be due to the fact that satisfaction measures are influenced by other things than consumption, e.g., comparison of one own's situation to others (Ferrer-i-Carbonell, 2005). Somewhat surprising is the fact that in our robustness checks parametric and semiparametric estimates based on satisfaction can differ, as the contrary has been reported for equivalence scales (Bellemare et al., 2002). The results for nondurable

Figure B.1: Kernel density estimates of the net replacement rates resulting from different welfare indicators, income concepts, and definitions of the retirement status





Note: For the US results are based on 157 estimates of net replacement rates; for Germany on 58 estimates.
Data: HRS 2014 and EVS 2013.

expenditure which are lower than our main findings are more difficult to explain, while the ones higher than our main findings might be explained as follows. Nondurable expenditure includes expenses that can be rather high trongly dependent on the need to commute; with retirement there is a big drop in these expenses, which is not reflective of a drop in the welfare level. Thus, using this welfare indicator likely overstates the replacement rate needed to maintain a constant standard of living, explaining the estimates which are higher than our main results.

For Germany, the results of the robustness checks are insofar different compared with the US in that the overall range of estimates is lower, and there are more estimates above the identification bounds, while there are no estimates below the bounds. This is due to the different welfare indicators used: no satisfaction based measure is available in the EVS, but different definitions for nondurables can be applied, while still maintaining a large sample size (in contrast to the HRS and CAMS). High estimates of the replacement rate around 110% and above mostly result when using a rather wide concept of nondurables which as described above might lead to overestimation of the replacement rate needed to maintain a constant standard of living.

The sensitivity checks also included several different definitions of the retirement status and income concept. The variation of results with respect to these variables is less systematic than in case of the welfare indicators. Given the same welfare indicator and using the same econometric approach the impact of the income concept and the retirement status on results is small for Germany. For the US, where we explore more definitions of the retirement status, variability is larger. Still, except for some outliers resulting when retirement is only defined through pension receipt, results are overall largely consistent.

B.4 Additional information on the RDD

B.4.1 Implementation

To estimate the probability of retirement, we use a linear probability model. d_i denotes the binary retirement status (1=retired,0=not retired). Let x_i be individual age, and let $\mathbb{I}(\cdot)$ be the indicator function. Roughly following Eibich (2015), we specify the model for Germany as

$$\begin{aligned} d_i &= a + b_x x_i + b_{60}\mathbb{I}(65 > x_i \geq 60) + b_{65}\mathbb{I}(x_i \geq 65) \\ &+ b_{x60} x_i \mathbb{I}(65 > x_i \geq 60) + b_{y65} x_i \mathbb{I}(x_i \geq 65) + \epsilon_i, \end{aligned}$$

where ϵ_i is an error term, and b_x, b_{60}, b_{65}, b_{x60}, and b_{x65} are the coefficients to be estimated. Coefficient estimates are then used to predict $\mathrm{E}(d_i|x_i)$. These predicted values are then used to instrument retirement status in equation (2.4). For the U.S. we use 62 instead of 60, and 65 is replaced with the NRA applying to each individual.

For the German data we use, only the year of birth is known for the respondents and not their exact age. It is therefore not possible to determine exactly how close respondents are to retirement age. Still, as the German system provides strong financial incentives to retire at certain thresholds, the cutoffs at ages 60 and 65 can be expected to to be relatively clear (Eibich, 2015), and thus the instrument rather strong (Börsch-Supan, 2000). The HRS, on the other hand, provides the exact date of birth.

B.4.2 Diagnostics: Discontinuities

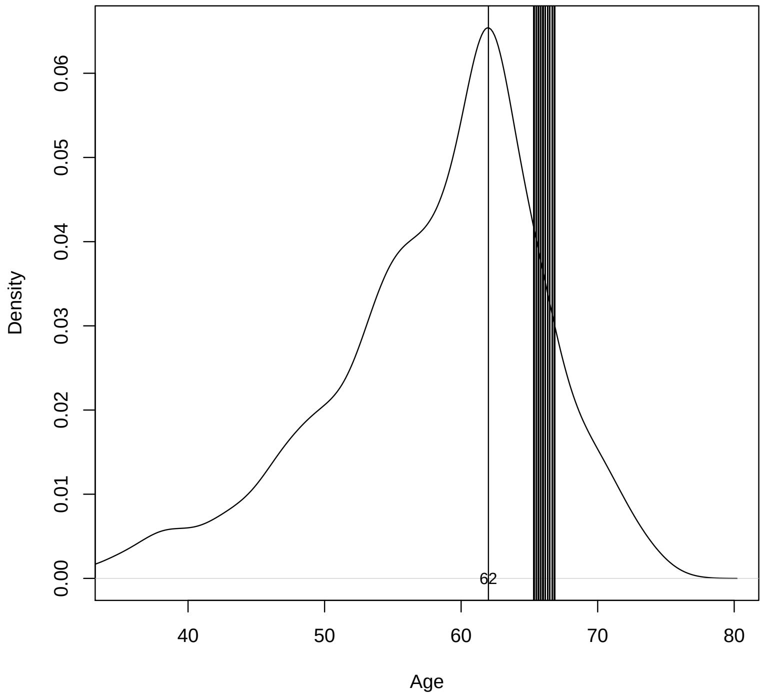
Following Imbens and Lemieux (2008) we test the validity of the RDD mostly graphically. As noted above, for Germany age can only be calculated based on the year of birth. This means that the running variable has to be treated as discrete and consequently it is not possible to compute averages very close to the the cutoff point (Lee and Card, 2008).

The first question we want to assess graphically is whether or not there is a clear cutoff in the probability of retiring in the running variable, i.e., age. The left panel of Figure B.3 demonstrates that in the German EVS sample the share of retiring individuals jumps at the age of 60 and the age of 65; i.e., these retirement eligibility ages are important thresholds. Potentially, the system offers incentive to retire early at the age of 63 (Börsch-Supan, 2000), but the graph suggests that this threshold is less relevant here.

For the US, a small discontinuity is recorded at the age of 62, which is the youngest age where US Americans become eligible for retirement benefits (Kämpfen and Maurer, 2016). This is confirmed by Figure B.2 showing that most US Americans in fact retire at this age. The NRA, which differs depending on the year of birth, is less clear cut and it has to be questioned whether that age can serve as a relevant instrument. This likely explains the high standard error we find when applying the RDD approach to the HRS, as reported in the main text.

To test the relevance of the instruments we calculate the F-statistic for excluding the retirement eligibility ages when estimating the probability of retiring (See C.1). For both the age of 62 and the NRA combined, we find the F-value to be 5.3 which is below 10, a benchmark often used in this literature (Staiger and Stock, 1997). For comparison, the F-statistic for joint instruments of 60 and 65 in the EVS sample amounts to 253.

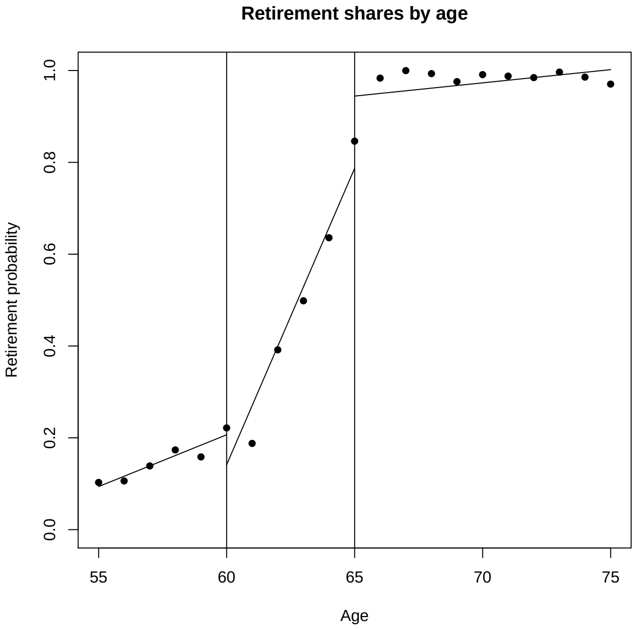
Figure B.2: Distribution of self-reported age of retirement taking into acoount cohorts born in 1939-1959



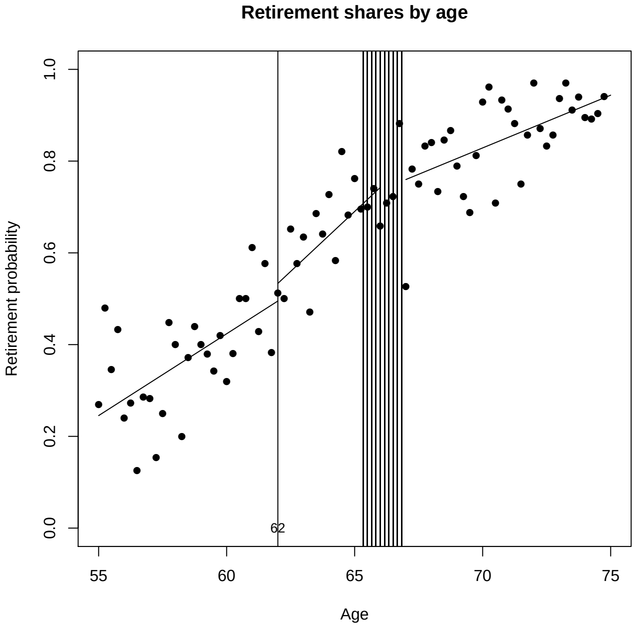
Note: the vertical lines represent eligibility ages at age 62 and the Normal Retirement Age (NRA) depending on the year of birth.
Data: HRS 2014 and EVS 2013.

Figure B.3: Age specific retirement probabilities and retirement eligibility

(a) GERMANY in year bins



(b) The U.S. in quarter bins



Data: HRS 2014 and EVS 2013.

B.4.3 Diagnostics: Further checks

We checked several other assumptions underlying the RDD approach. We briefly summarize the findings here. Overall, these findings confirm that the RDD approach should work well for the German data, but might be weak in case of the US. Detailed results are available upon request.

First, for the RDD to be valid respondents are not allowed to have influence on the running variable. In our case age is the running variable. Apart from misreporting their birthday or year of birth the respondents cannot influence this variable. While age might be misreported in the data we use, it is likely not a big issue. Second, the outcome variable (welfare indicator) must be a smooth function of the assignment variable. Assessing this visually showed that i is the case for Germany, while results are less conclusive for the US. Third, we assessed the sensitivity of the RDD results with respect to the choice of bandwidth, to the polynomial degree, and to the choice of cutoff points. Again, results showed no issues for Germany; results for the US were rather sensitive to these choices, though, again showing that the RDD approach potentially is not valid for the subset of HRS data we use.

B.5 Details on the semiparametric approach and the nonparametric approach

The semiparametric approach and the nonparametric approach make use of nonparametric kernel regression and nonparametric estimates of quantile functions. We use the implementation in the `np` package for R, developed by Hayfield and Racine (2008). This package implements generalized product kernels, which allow mixing of continuous and categorical explanatory variables (Racine and Li, 2004). These kernels are defined as

$$K(\mathbf{x}_i - \mathbf{x}) = \prod_{k=1}^{m_1} \frac{1}{h_k} K_k^{(c)}(x_{ki} - x_k) \prod_{l=m_1+1}^{m} K_l^{(d)}(x_{ki} - x_k),$$

where $\mathbf{x}$ is a vector of explanatory variables with elements x_k. m is the number of elements of $\mathbf{x}$, with the first m_1 elements being continuous and the other $m - m_1$ elements being categorical. h_k is the bandwidth for variable x_k. The type of the kernel function depends on the type of variable, where $K^{(c)}$ indicates the kernel function for continuous variables, and $K^{(d)}$ the kernel function for categorical variables. For continuous variables, we used a second-order Gaussian Kernel and for the categorical case the kernel function proposed by Aitchison and Aitken was utilized (Hayfield and Racine, 2008). For bandwidth selection, see Hall et al. (2004).

For statistical inference, for the semiparametric estimation approach we use a residual bootstrap as proposed by Pendakur (1999). For the nonparametric estimation approach, we use the resampling bootstrap. This procedure could potentially be conservative (see Härdle and Mammen, 1993), but residual-based approaches are not well defined for estimates of conditional quantile functions. For both the semiparametric estimation approach and the nonparametric estimation approach standard errors and confidence intervals are based on 1000 bootstrap replications.

Confidence intervals are based on percentiles of the bootstrap replications. In case of the nonparametric identification region, we construct an interval which covers the complete identification region with a fixed probability (95%), roughly similar to Horowitz and Manski (2000). If (L_s, U_s) denote the bounds resulting for the sth bootstrap replication, then we choose the bounds l and u of the confidence interval such that $l < L_s$ and $U_s < u$ for 95% of the bootstrap replications. As l and u will usually not be unique, we choose the values of l and u for which the interval width, $u - l$, is smallest.

To calculate the smallest interval, an iterative procedure is used and re-run several times. Each run takes the 5% percentile of the lower bound L and the 95% percentile of the upper bound U as starting values, perturbed with noise $e_L \sim \mathcal{N}(0, \mathrm{sd}(L))$ or $e_U \sim \mathcal{N}(0, \mathrm{sd}(U))$, respectively. Let the resulting bounds be denoted by $L^{(0)}$ and $U^{(0)}$. The coverage achieved with these values is equal to $\rho^{(0)}$. If $\rho^{(0)}$ is smaller than $1 - \alpha$, $L^{(0)}$ and $U^{(0)}$ are decreased and increased, respectively, by a stepsize $\lambda_L = 0.1\mathrm{sd}(L)$ or $\lambda_U = 0.1\mathrm{sd}(U)$ to get new values: $L^{(1)} = L^{(0)} - \lambda\epsilon_L^{(0)}$ and $U^{(1)} = U^{(0)} + \lambda\epsilon_U^{(0)}$, where $\epsilon_U^{(0)}$ and $\epsilon_L^{(0)}$ follow a uniform distribution. If $\rho^{(0)}$ is larger than $1 - \alpha$ the signs for λ are changed to instead decrease the interval width. $\rho^{(1)}$ is the coverage achieved after these adjustments. Depending on whether it is above or below $1 - \alpha$, the adjustments are applied to get updated values $L^{(2)}$ and $U^{(2)}$; $\rho^{(2)}$ is checked against $1 - \alpha$ again etc. until $\rho^{(k)} = 1 - \alpha$. This procedure is re-run for 100 different starting values and the interval with the smallest width is reported.

To find the values of α and μ which minimize equation (2.6), we inserted values between 0.4 and 1.2 for α, starting from 0.4 and using increments of 0.001. Conditional on α, μ is identified and can easily be calculated, allowing in turn to calculate (2.6). α was then chosen as the value in the interval $(0.4, 1.2)$ which yields the lowest value of equation (2.6). If the lowest value of α was either 0.4 or 1.2 it indicates that (2.6) is likely not a convex function over $[0.4, 1.2]$, and we say that the semiparametric approach did not converge.

The semiparametric approach makes use of simulation-based inference to assess income independence, where we follow Pendakur (1999). Income independence is assessed by comparing the empirical value of $L(\alpha, \mu)$, as defined in equation (2.6), to simulated values of L arising from assuming income independence. These are generated by using the residual bootstrap, where predicted values for retirees are generated using the kernel regression function of non-retirees shifted by the estimates of α and μ. Using this bootstrap sample, the semiparametric approach is fitted again, yielding a value of L^* based on shape invariance and thus income independence. The distribution of simulated values L^* is used to assess the probability of the empirical value of L.

Chapter C

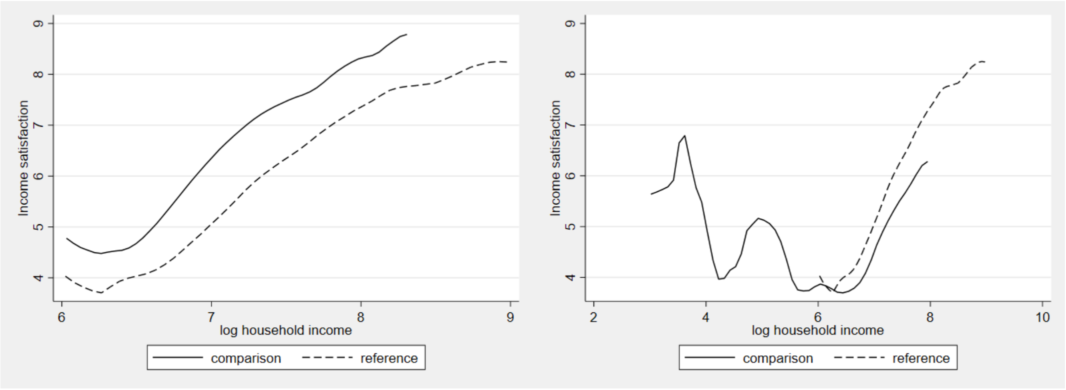
Appendix to Chapter 3

C.1 Non-parametric Engel curves

Figure C.1: Non-parametric regression line of log household income and income satisfaction by household type:; single households; reference: single working

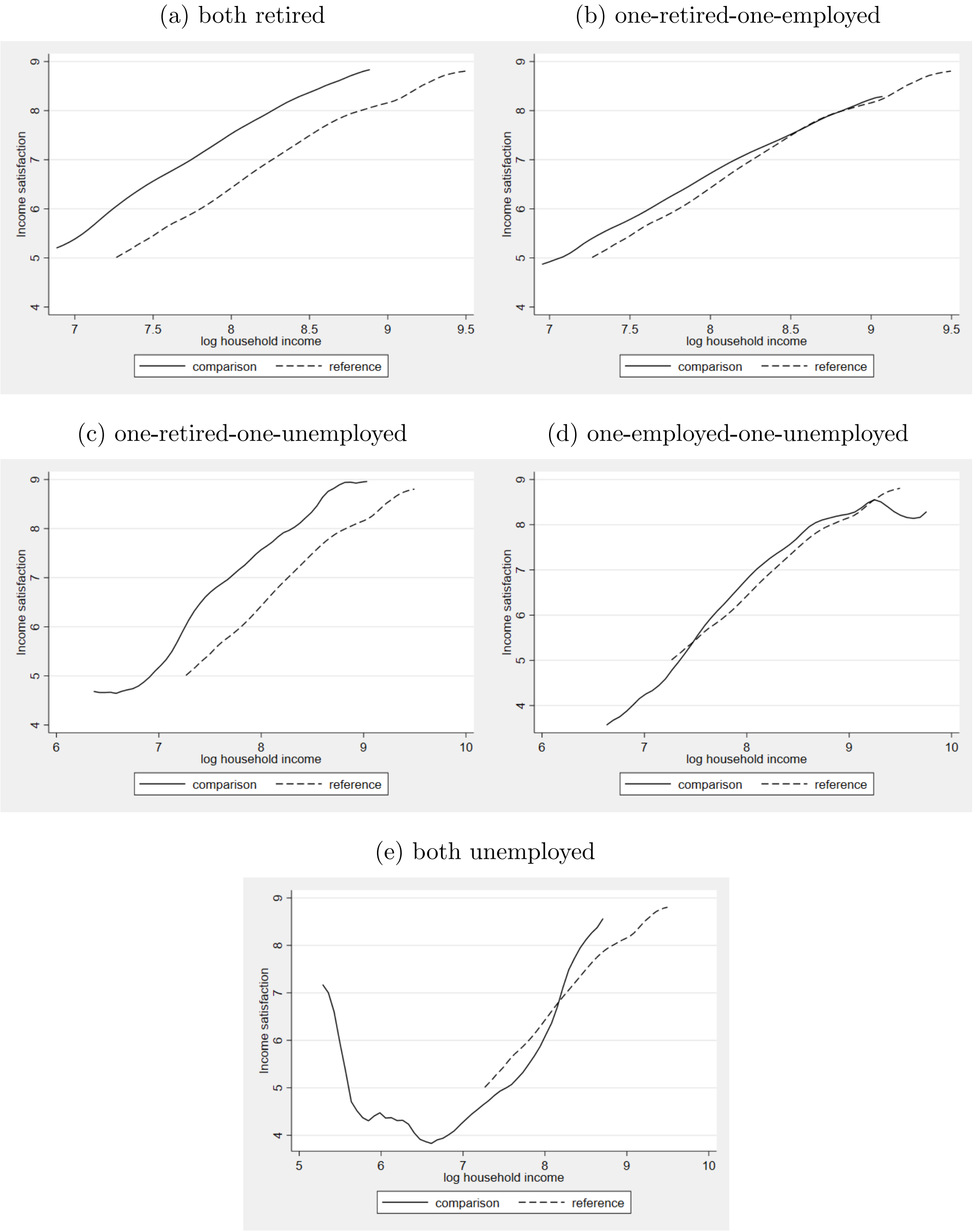
(a) Single retired

(b) Single unemployed



Data: SOEP 1991-2017

Figure C.2: Non-parametric regression line of log household income and income satisfaction by household type: couple households; reference: both employed



(a) both retired

(b) one-retired-one-employed

(c) one-retired-one-unemployed

(d) one-employed-one-unemployed

(e) both unemployed

Note: polynomial smooth ... bla bla Kernel bla bla
Data: SOEP 1991-2017

C.2 Coefficients for the sample 60 years and older

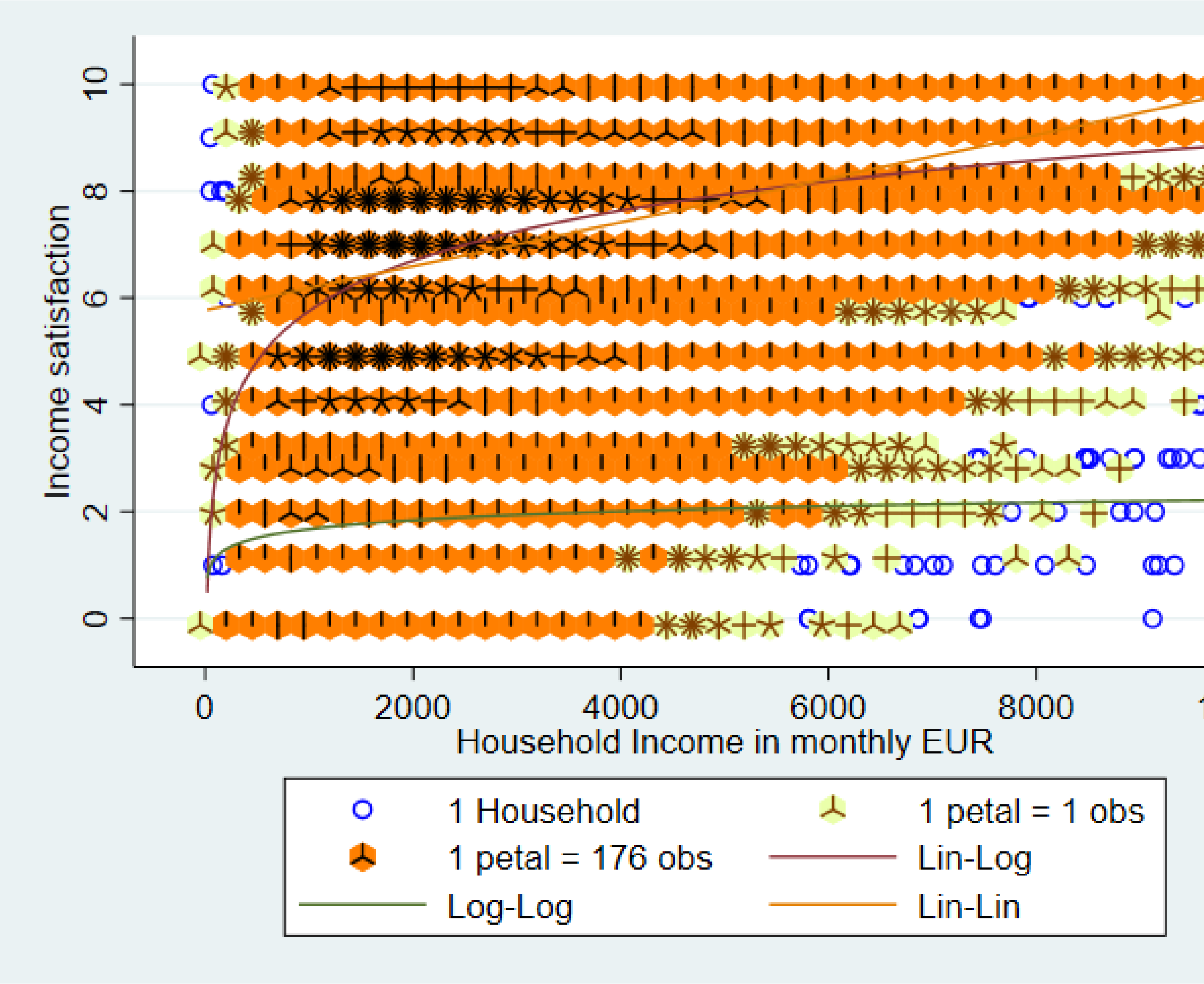
Table C.1: Coefficients Model 3.6 ; Single worker as reference category

Parameter	Variable	Coefficient	SE	P
β_1	log(Income)	1.446944	0.113	0.001
$\alpha^{a_{ret}}$	scaling comp retired singles	-879.8	185.0	0.003
$\alpha^{a_{un}}$	scaling comp unemployed singles	-267.0	192.4	0.189
$\alpha^{aa_{ret}}$	scaling comp both retired	-821.5	195.9	0.001
$\alpha^{aa_{emp}}$	scaling comp both working	-817.4	212.4	0.001
$\alpha^{aa_{ret;emp}}$	scaling comp retired/working	-850.1	197.4	0.001
$\alpha^{aa_{emp;un}}$	scaling comp working/unemployed	-1016.0	167.8	0.001
$\alpha^{aa_{ret;un}}$	scaling comp retired/unemployed	-776.6	212.1	0.002
$\alpha^{aa_{un}}$	scaling comp both unemployed	-666.0	720.0	0.676
$\rho^{a_{ret}}$	translating comp retired singles	1.728	0.163	0.001
$\rho^{a_{un}}$	translating comp unemployed singles	0.894	0.171	0.001
$\rho^{aa_{ret}}$	translating comp both retired	1.231	0.128	0.001
$\rho^{aa_{emp}}$	translating comp both working	1.072	0.126	0.001
$\rho^{aa_{ret;emp}}$	translating comp retired/working	1.141	0.122	0.001
$\rho^{aa_{emp;un}}$	translating comp working/unemployed	1.114	0.107	0.001
$\rho^{aa_{ret;un}}$	translating comp retired/unemployed	1.122	0.130	0.001
$\rho^{aa_{un}}$	translating comp both unemployed	0.899	0.482	0.001
γ_1	Age	-0.018	0.005	0.002

Note: The dependent variable is income satisfaction scales $0-10$. Period effects are included but not shown. Maximum likelihood estimation. N=28,381; Standard errors are clusted on the household levels.
Data: Socio-Economic Panel 1991-2017.

C.3 Fit of functional form

Figure C.3: Sunflower graph to show fit of functional forms



Notes:
Data: German Socio-Economic Panel 1991-2017

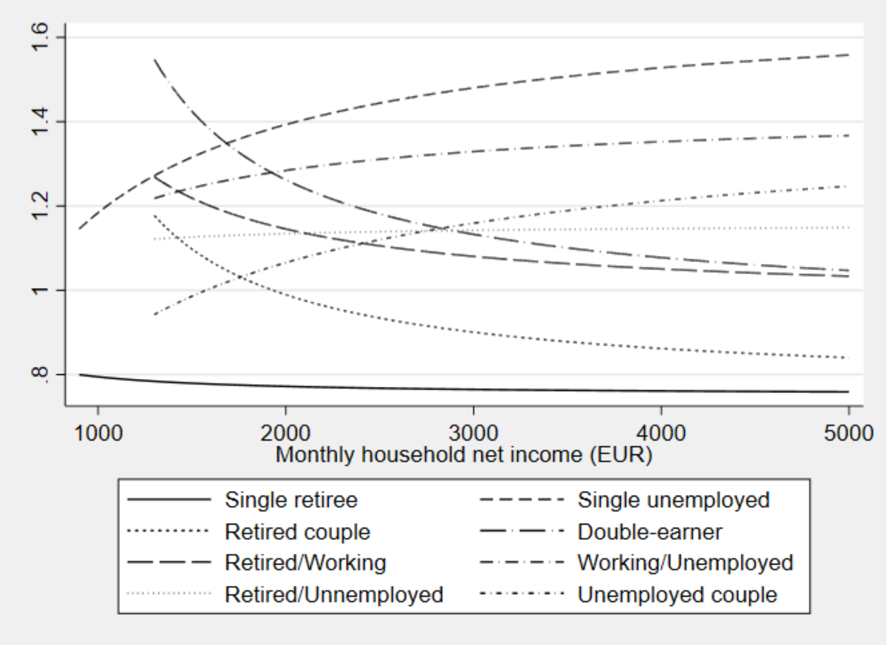
C.4 Summary statistics

Table C.2: Sample summary statistics

Percent Female (in %)	50,5
Year of birth (min)	1923
Year of birth (max)	1957
Share singles at age 65 (%)	12.5
Share singles at age 80 (%)	36.0
Age (min)	50
Age (max)	92
Retirement year (mean)	63.4
Retirement year (min)	60
Retirement year (max)	69

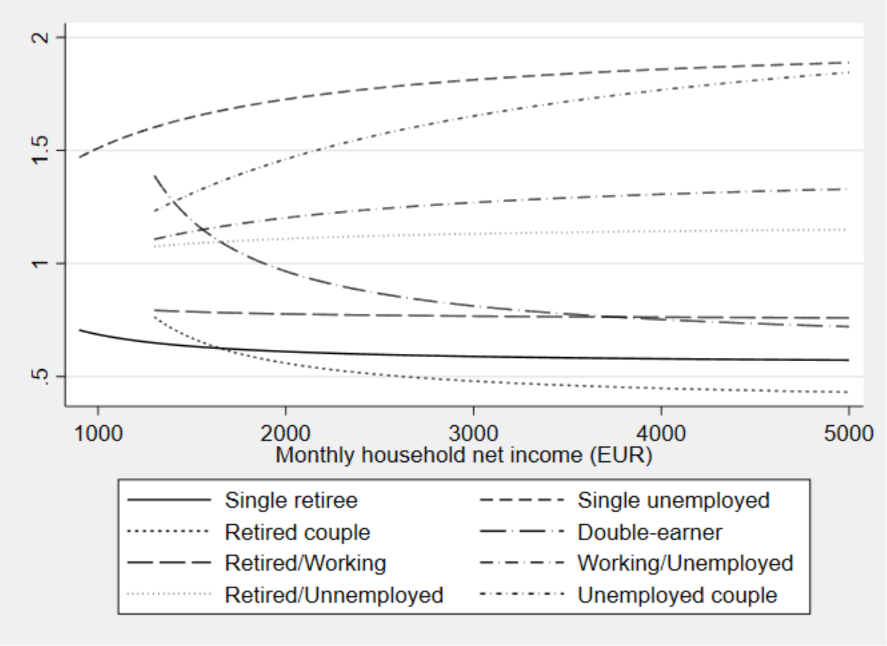
C.5 Further results from the sensitivity analysis

Figure C.4: Results with a transformed binary outcome variable ; single worker as reference category



Data: German Socio-Economic Panel 1991-2017

Figure C.5: Results with a transformed categorized outcome variable ; single worker as reference category



Data: German Socio-Economic Panel 1991-2017

Table C.3: Coefficients Model 3.10 ; Single worker as reference category

Parameter	Coefficient	SE	P
β_1	0.942	0.026	0.001
$dy^{a_{ret}}$	-833.0	127.9	0.003
$dy^{a_{un}}$	-1216.3	320.6	0.023
$dy^{aa_{ret}}$	-966.5	136.5	0.001
$dy^{aa_{emp}}$	-1500.5	166.7	0.001
$dy^{aa_{ret;emp}}$	-1276.9	166.7	0.001
$dy^{aa_{emp;un}}$	-1779.9	183.1	0.001
$dy^{aa_{ret;un}}$	-1117.6	193.2	0.002
$dy^{aa_{un}}$	-1310.3	604.9	0.485
$\beta_2^{a_{ret}}$	0.968	0.090	0.001
$\beta_2^{a_{un}}$	0.607	0.267	0.001
$\beta_2^{aa_{ret}}$	0.631	0.062	0.001
$\beta_2^{aa_{emp}}$	0.718	0.058	0.001
$\beta_2^{aa_{ret;emp}}$	0.683	0.068	0.001
$\beta_2^{aa_{emp;un}}$	0.725	0.078	0.001
$\beta_2^{aa_{ret;un}}$	0.270	0.096	0.001
$\beta_2^{aa_{un}}$	0.743	0.384	0.030
$\gamma_1(Age)$	0.057	0.012	0.001
$\gamma_2(Age^2)$	<-0.001	<-0.001	0.001

Note: The dependent variable is income satisfaction scales $0 - 10$. Period effects are included but not shown. Two way fixed effect estimation. Standard errors are clustered on the individual level. N=138,631, n = 17,207
Data: Socio-Economic Panel 1991-2017.

Chapter D

Appendix to Chapter 4

D.1 Different age ranges and cut-off points

D.2 Unobserved heterogeneity

Table D.1: Effect of testosterone on unemployment risk controlling for initial conditions and unobserved heterogeneity

	Initially unemployed	**Initially employed**
Testosterone		
1^{st} decile	*reference category*	
2^{nd} – 9^{th} decile	-4.13***	0.4
	(1.47)	(0.43)
10^{th} decile	-3.21**	1.18**
	(1.42)	(0.58)
Observations	371	6,408
LogLikelihood	-50.59	-297.4

Notes: Author's own calculations using data from Understanding Society subsample Health and biomarkers Survey. N= 6,771. ***,**,* refers to statistically significant at 1%, 5% and 10% level respectively.

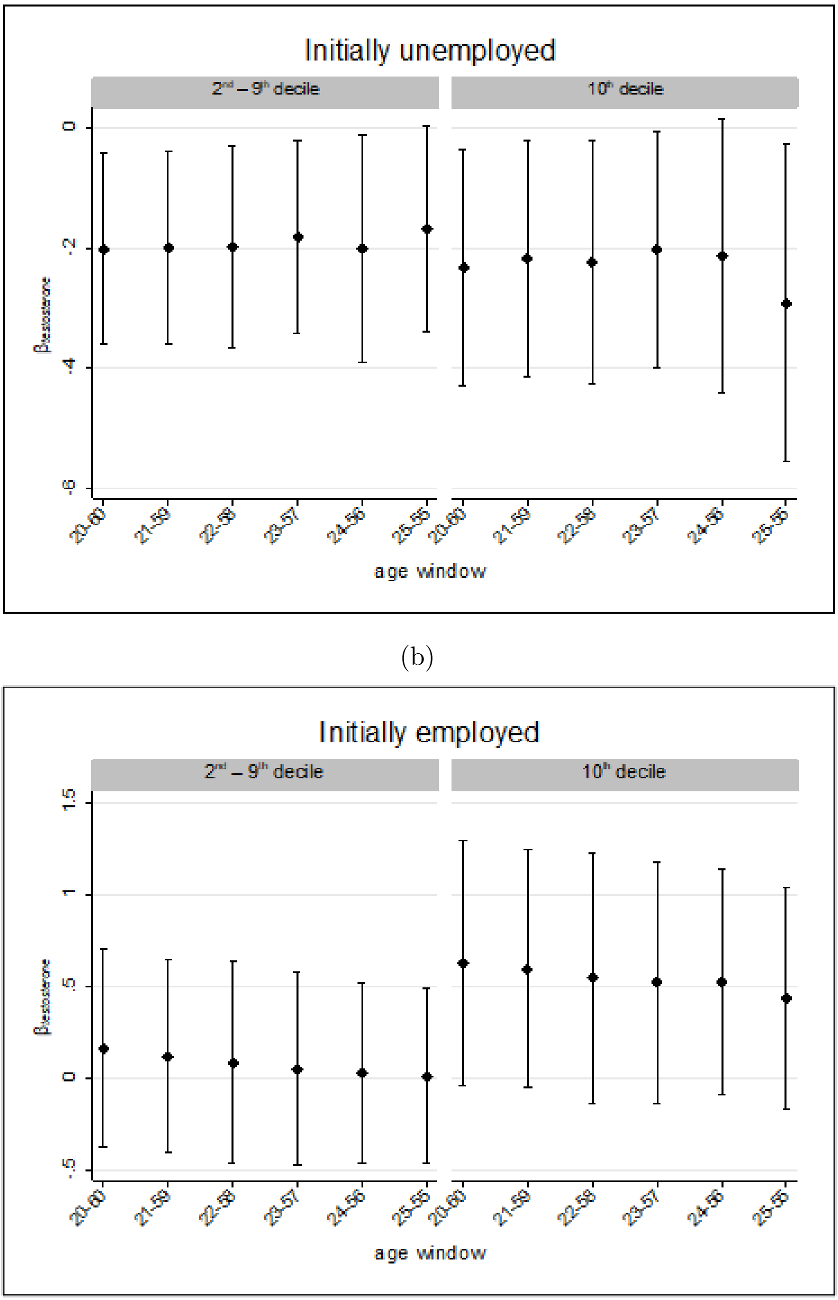
D.3 Numeric ability and fluid reasoning

Table D.2: Numerical ability (proxied by number ability-test) and testosterone level

	Score numerical ability	Score fluid reasoning
Testosterone		
Low (1st quintile)	reference category	
Medium (2nd 4th quintile)	1.29*** (0.10)	1.28** (0.14)
High (5th quintile)	1.01 (0.003)	1.22 (0.018)
Age	1.01*** (0.003)	0.99 (0.004)
Observations	3,123	1,540
LogLikelihood	-3844.41	-2358.37

Notes: Author's own calculations using data from Understanding Society. ***,**,* refers to statistically significant at 1%, 5% and 10% level respectively. Coefficients refer to odds ratio (exp(β)).

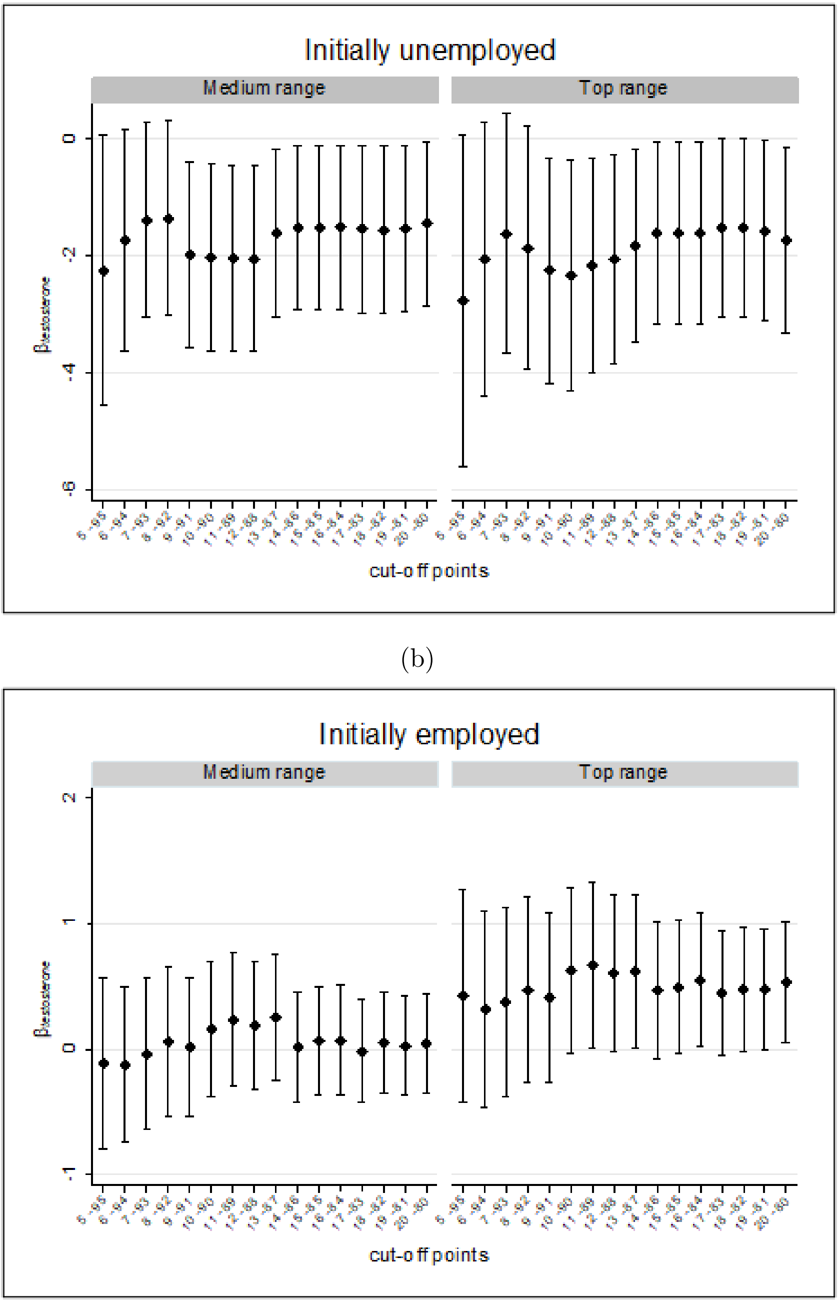
Figure D.1: Robustness check for different age ranges



(a)

(b)

Figure D.2: Robustness check for different testosterone cut-off points



(a)

(b)

Summary

This dissertation tackles the methodological challenges involved in estimating the effects of life -course transitions on economic well-being, and it investigates income, consumption, and subjective measures of well-being before and after these transitions. Additionally, new sources of data are taken into account to improve our understanding of why people move from one labor force status to another. The first chapter is a comparison of econometric methods applied to one dataset. The second chapter establishes an econometric framework, and discusses two applications. The third chapter generalizes the framework from the second chapter and applies it to German survey data. The fourth chapter uses existing labor market transition models, but adds a new source of data.

The first chapter provides a comprehensive comparison of methods for estimating equivalence scales, a metric that is routinely applied to adjust the incomes of households of different sizes and compositions. Drawing on German household expenditure data, we estimate equivalence scales using several parametric, semiparametric, and nonparametric approaches. We find that some approaches yield more plausible results than others, while implausible scales are mostly based on linear Engel curves. The results we consider plausible are close to the modified OECD scale, and to the square root scale for larger households.

The second chapter analyses changes in monetary needs when individuals retire. We propose a framework that can be used to assess what pension benefit levels can be considered adequate, and the economic costs or benefits of retiring. Applying a range of econometric techniques to data from the U.S. and Germany, we find that a net pension income of around 100% of the end-of-career after-tax income can be considered adequate.

The third chapter is concerned with the question of whether high earners need the same adequacy standard as low earners. I generalize the framework from the second chapter such that the adequate replacement rate is a function of income. Using longitudinal data from the German Socio-Economic Panel, and applying fixed-effects ordered logit models, I find that the benchmark replacement rate decreases with income. For singles, the metric decreases significantly which is consistent across many modifications of the

approach.

The fourth chapter looks at the variation in testosterone levels in the population to explain the probabilities of transitioning into and out of unemployment. By following the individual employment histories of British men, and by applying dynamic random-effects models, we find that individuals with high testosterone levels are more likely to become unemployed, but they are also more likely to exit unemployment. Based on previous studies and descriptive evidence, we argue that these effects are likely driven by the personality traits and the occupational sorting of men with high testosterone levels. Our findings suggest that latent biological processes not necessarily related to illness and disability can affect job search behavior and labor market outcomes.

German summary

Was hat der Wechsel von einem Lebensabschnitt zum nächsten für einen Einfluss auf unseren Lebensstandard und wie verhalten sich Einkommen, Konsum und subjektive Zufriedenheit vor und nach solchen Lebensabschnitten? Weshalb werden Menschen arbeitslos und wie kann man erklären, dass Manche schneller zurück in das Erwerbsleben finden? Diese Arbeit widmet sich den methodischen Herausforderungen, die mit diesen Fragen einhergehen und untersucht inwiefern neu verfügbare Datentypen neue kausale Rückschlüsse zulassen.

Das erste Kapitel der Arbeit umfasst einen breiten Methodenvergleich einiger, in der Literatur vorhandene, Schätzmethoden. Das zweite Kapitel leitet ein ökonometrisches Schätzmodell her und stellt zwei Anwendungen vor. Das dritte Kapitel, erweitert das Schätzmodell des zweiten Kapitels und demonstriert ebenfalls zwei Anwendungen. Das vierte Kapitel nutzt vorhandene Identifikationsstrategien, führt aber eine in dem Zusammenhang neue Variable ein.

Im ersten Kapitel werden verschieden Methoden zur Schätzung von Äquivalenzskalen mit einander verglichen. Äquivalenzskalen werden zur Standardisierung von Einkommen zwischen unterschiedlichen Haushaltsgrößen- und Typen herangezogen. An Hand der Einkommens- und Verbrauchsstichprobe (EVS) des Statistischen Bundesamtes werden parametrische, semi-parametrische und nicht-parametrische Methoden zur Schätzung der Skalen durchgeführt. Dabei liefern Ansätze, welche auf der Annahme linearer Engelkurven basieren, fragwürdige Äquivalenzskalen, während der Rest der Skalen ähnlich zur modifizierten OECD-Skala ausfällt (oder zu den Skalen der Quadratwurzelmethode bei größeren Familien).

Im zweiten Kapitel geht es um den Renteneintritt und wie sich dadurch Einnahmen und Ausgaben verändern. Ein ökonometrisches Schätzmodell wird vorgestellt, an Hand dessen sich ein „adäquates“ Sicherungsniveau empirisch herleiten lässt. Es werden dazu ein Datensatz aus den USA und ein Datensatz aus Deutschland untersucht. Unter der Verwendung einer Vielzahl von ökonometrische Methoden kommen wir zu dem Ergebnis, dass ein Nettorenteneinkommen, welches etwa 100% des letzten Nettoarbeitseinkommen entspricht, als „adäquat“ bezeichnet werden kann.

Das dritte Kapitel widmet sich der Frage, ob die oben berechneten Sicherungsziele gleichermaßen für niedrige bzw. hohe Einkommensklassen gelten. Dazu wird das

Modell aus dem zweiten Kapitel erweitert, so dass das Sicherungsziel in Abhängigkeit vom Einkommen identifiziert wird. Ein derartiges Sicherungsziel wird für Deutschland mit dem SOEP berechnet. Es stellt sich heraus, dass eine Ersatzrate, welche die Einkommenszufriedenheit über den Renteneintritt aufrecht erhält, mit dem Einkommen sinkt – bei Single-Haushalten sogar signifikant.

Im vierten Kapitel werden Bevölkerungsunterschiede im Testosteronspiegel genutzt um den Übergang in die Arbeitslosigkeit sowie zurück in die Erwerbstätigkeit besser zu verstehen. Dazu wird die Erwerbshistorie einer Reihe von britischen Männer untersucht. Mit Hilfe von dynamischen *Random-effect*-Schätzmodellen stellen wir fest, dass Menschen mit hohen Testosteronspiegel mit höherer Wahrscheinlichkeit arbeitslos werden, sie aber gleichwohl schneller in die Erwerbstätigkeit zurück finden. Einschlägige Literatur sowie einige deskriptive Auswertungen ergaben, dass die Heterogenität durch Persönlichkeitsmerkmale und durch berufliche Selbstselektion erklärt werden kann. Die Ergebnisse zeigen, dass biologische Prozesse einen Einfluss auf das individuelle Jobsuchverhalten nehmen kann, ohne sich auf gesundheitliche Probleme zu beschränken.

Bibliography

Abadie, A. and Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–257.

Abbring, J. H. and Heckman, J. J. (2007). Econometric evaluation of social programs, Part III. In *Handbook of Econometrics*, volume 6B, pages 5146–5303. Elsevier.

Aguiar, M. and Hurst, E. (2005). Consumption versus expenditure. *Journal of Political Economy*, 113:919–948.

Aguiar, M. and Hurst, E. (2007). Life-cycle prices and production. *American Economic Review*, 97:1533–1559.

Akay, A. (2012). Finite-sample comparison of alternative methods for estimating dynamic panel data models. *Journal of Applied Econometrics*, 27(7):1189–1204.

Angelini, V., Brugiavini, A., and Weber, G. (2014). The dynamics of homeownership among the 50+ in Europe. *Journal of Population Economics*, 27:797–823.

Apicella, C. L., Dreber, A., Campbell, B., Gray, P. B., Hoffman, M., and Little, A. C. (2008). Testosterone and financial risk preferences. *Evolution and Human Behavior*, 29(6):384–390.

Archer, J. (2006). Testosterone and human aggression: An evaluation of the challenge hypothesis. *Relationship between the Brain and Aggression*, 30(3):319–345.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.

Arulampalam, W. (2001). Is unemployment really scarring? Effects of unemployment experiences on wages. *The Economic Journal*, 111(475):585–606.

Atkinson, A. B., Rainwater, L., Smeeding, T. M., et al. (1995). Income distribution in OECD countries: Evidence from the Luxembourg Income Study. LIS Working Paper Series 120, Cross-national Data Center in Luxembourg.

Attanasio, O. P., Kitao, S., and Violante, G. L. (2007). Global demographic trends and social security reform. *Journal of Monetary Economics*, 54:144–198.

Attanasio, O. P. and Weber, G. (2010). Consumption and saving: Models of intertemporal allocation and their implications for public policy. *Journal of Economic Literature*, 48:693–751.

Baetschmann, G., Staub, K. E., and Winkelmann, R. (2015). Consistent estimation of the fixed effects ordered logit model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):685–703.

Balli, F. and Tiezzi, S. (2010). Equivalence scales, the cost of children and household consumption patterns in Italy. *Review of Economics of the Household*, 8(4):527–549.

Banks, J., Blundell, R., Levell, P., and Smith, J. P. (2019). Life-cycle consumption patterns at older ages in the united states and the united kingdom: Can medical expenditures explain the difference? *American Economic Journal: Economic Policy*, 11(3):27–54.

Banks, J., Blundell, R., and Lewbel, A. (1997). Quadratic engel curves and consumer demand. *The Review of Economics and Statistics*, 79(4):527–539.

Bann, D., Hardy, R., Cooper, R., Lashen, H., Keevil, B., Wu, F. C., Holly, J. M., Ong, K. K., Ben-Shlomo, Y., and Kuh, D. (2015). Socioeconomic conditions across life related to multiple measures of the endocrine system in older adults: Longitudinal findings from a British birth cohort study. *Social Science & Medicine*, 147:190–199.

Barnett, W. A. and Serletis, A. (2008). Consumer preferences and demand systems. *Journal of Econometrics*, 147:210–224.

Battistin, E., Brugiavini, A., Rettore, E., and Weber, G. (2009). The retirement consumption puzzle: Evidence from a regression discontinuity approach. *American Economic Review*, 99:2209–2226.

Behagel, L. and Blau, D. M. (2012). Framing social security reform: Behavioral responses to changes in the full retirement age. *American Economic Journal: Economic Policy*, 4:41–67.

Bellemare, C., Melenberg, B., and van Soest, A. (2002). Semi-parametric models for satisfaction with income. *Portuguese Economic Journal*, 1:181–203.

Benzeval, M., Davillas, A., Kumari, M., and Lynn, P. (2014). Understanding Society: The UK Household Longitudinal Study. Biomarker User Guide and Glossary. Technical report, University of Essex, Colchester.

Bernheim, B. (1992). Is the baby boom generation preparing adequately for retirement? Technical Report, Princeton NJ, Merrill Lynch.

Bernheim, B. D., Skinner, J., and Weinberg, S. (2001). What accounts for the variation in retirement wealth among U.S. households? *American Economic Review*, 91:832–857.

Bernheim, D. A. and Garret, D. M. (2003). The effects of financial education in the workplace: evidence from a survey of households. *Journal of Public Economics*, 87:1487–1519.

Bertrand, M. and Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. *American Economic Review*, 91(2):67–72.

Bhuller, M., Brinch, C. N., and Königs, S. (2017). Time aggregation and state dependence in welfare receipt. *The Economic Journal*, 127(604):1833–1873.

Biewen, M. and Juhasz, A. (2017). Direct estimation of equivalence scales and more evidence on independence of base. *Oxford Bulletin of Economics and Statistics*, 79:875–905.

Biewen, M. and Steffes, S. (2010). Unemployment persistence: Is there evidence for stigma effects? *Economics Letters*, 106(3):188–190.

Bijlsma, M. J., Tarkiainen, L., Myrskylä, M., and Martikainen, P. (2017). Unemployment and subsequent depression: A mediation analysis using the parametric G-formula. *Social Science & Medicine*, 194:142–150.

Binswanger, J. and Schunk, D. (2012). What is an adequate standard of living during retirement? *Journal of Pension Economics and Finance*, 11:203–222.

Blacklow, P., Nicholas, A., and Ray, R. (2010). Demographic demand systems with application to equivalence scales estimation and inequality analysis: The Australian evidence. *Australian Economic Papers*, 49(3):161–179.

Blackorby, C. and Donaldson, D. (1993). Adult-equivalence scales and the economic implementation of interpersonal comparisons of well being. *Social Choice and Welfare*, 10:335–361.

Blanchflower, D. G. and Piper, A. (2022). There is a mid-life low in well-being in Germany. *Economics Letters*, 214:110430.

Blundell, R., Crawford, R., French, E., and Tetlow, G. (2016). Comparing retirement wealth trajectories on both sides of the pond. *Fiscal Studies*, 37:105–130.

Blundell, R., Duncan, A., and Pendakur, K. (1998). Semiparametric estimation and consumer demand. *Journal of Applied econometrics*, 13(5):435–461.

Blundell, R. and Lewbel, A. (1991). The information content of equivalence scales. *Journal of Econometrics*, 50:49–68.

Bönke, T., Kemptner, D., and Lüthen, H. (2018). Effectiveness of early retirement disincentives: Individual welfare, distributional and fiscal implications. *Labour Economics*, 51(1):25–37.

Bönke, T., Schröder, C., and Schulte, K. (2010). Incomes and inequality in the long run: The case of German elderly. *German Economic Review*, 11:487–510.

Booth, A., Granger, D. A., Mazur, A., and Kivlighan, K. T. (2006). Testosterone and social behavior. *Social Forces*, 85(1):167–191.

Booth, A., Johnson, D. R., and Granger, D. A. (1999). Testosterone and men's depression: The role of social behavior. *Journal of Health and Social Behavior*, 40(2):130–140.

Borah, M., Keldenich, C., and Knabe, A. (2019). Reference income effects in the determination of equivalence scales using income satisfaction data. *Review of Income and Wealth*, 65(4):736–770.

Börsch-Supan, A. (2000). Incentive effects of social security on labor force participation: evidence in Germany and across Europe. *Journal of Public Economics*, 78(1):25 – 49.

Börsch-Supan, A. and Wilke, C. B. (2004). The German public pension system: How it was, how it will be. NBER Working Paper No. 10525.

Bosch-Domènech, A., Brañas-Garza, P., and Espín, A. M. (2014). Can exposure to prenatal sex hormones (2D:4D) predict cognitive reflection? *Psychoneuroendocrinology*, 43:1–10.

Boskin, M. J. and Shoven, J. B. (1984). Concepts and measures of earnings replacement during retirement. NBER Working Paper No. 1360.

Brenke, K. and Zimmermann, K. (2009). Ostdeutschland 20 Jahre nach dem Mauerfall: Was war und was ist heute mit der Wirtschaft? *Quarterly Journal of Economic Research*, 78(2):32–62.

Brookes, H., Neave, N., Hamilton, C., and Fink, B. (2007). Digit Ratio (2D:4D) and lateralization for basic numerical quantification. *Journal of Individual Differences*, 28(2):55–63.

Brosnan, M., Gallop, V., Iftikhar, N., and Keogh, E. (2011). Digit ratio (2D:4D), academic performance in computer science and computer-related anxiety. *Personality and Individual Differences Research*, 51(4):371–375.

Browning, M. and Crossley, T. F. (2001). The life-cycle model of consumption and saving. *Journal of Economic Perspectives*, 15:3–22.

Bütikofer, A., Figlio, D. N., Karbownik, K., Kuzawa, C. W., and Salvanes, K. G. (2019). Evidence that prenatal testosterone transfer from male twins reduces the fertility and socioeconomic success of their female co-twins. *Proceedings of the National Academy of Sciences*, 116(14):6749.

Bucher-Koehnen, T. and Lusardi, A. (2011). Financial literacy and retirement planning in Germany. *Journal of Pension Economics and Finance*, 10:565–584.

Buhmann, B., Rainwater, L., Schmaus, G., and Smeeding, T. M. (1988). Equivalence scales, well-being, inequality, and poverty: Sensitivity estimates across ten countries using the Luxembuorg Income Study (LIS) database. *Review of Income and Wealth*, 34:115–142.

Burgess, S. and Thompson, S. G. (2015). *Mendelian randomization: methods for using genetic variants in causal estimation*. CRC Press.

Burkhauser, R. V., Smeeding, T. M., and Merz, J. (1996). Relative inequality and poverty in Germany and the United States using alternative equivalence scales. *Review of Income and Wealth*, 42:381–400.

Butler, J., Anderson, K. H., and Burkhauser, R. V. (1989). Work and health after retirement: A competing risks model with semiparametric unobserved heterogeneity. *Review of Economics and Statistics*, 71(1):46–53.

Caliendo, M., Cobb-Clark, D. A., and Uhlendorff, A. (2015). Locus of control and job search strategies. *The Review of Economics and Statistics*, 97(1):88–103.

Cambanis, S., Simons, G., and Stout, W. (1976). Inequalitites for $ek(x, y)$ when the marginals are fixed. *Probability Theory and Related Fields*, 36:285–294.

Carré, J. M. and McCormick, C. M. (2008). Aggressive behavior and change in salivary testosterone concentrations predict willingness to engage in a competitive task. *Hormones and Behavior*, 54(3):403–409.

Cartwright, D. I. and Field, M. J. (1978). A refinement of the arithmetic mean-geometric mean inequality. *Proceedings of the American Mathematical Society*, 71:36–38.

Chamberlain, G. (1979). Analysis of covariance with qualitative data. NBER Working Paper, National Bureau of Economic Research.

Chamberlain, G. (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics*, 10(1):20–26.

Chance, S. E., Brown, R. T., Dabbs, J. M., and Casey, R. (2000). Testosterone, intelligence and behavior disorders in young boys. *Personality and Individual Differences*, 28(3):437–445.

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., and Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology*, 104(1):103–125.

Cherchye, L., De Rock, B., and Vermeulen, F. (2012). Economic well-being and poverty among the elderly: An analysis based on a collective consumption model. *European Economic Review*, 56:985–1000.

Chiappori, P.-A. (2016). Equivalence versus indifference scales. *The Economic Journal*, 126(592):523–545.

Coates, J. M., Gurnell, M., and Rustichini, A. (2009). Second-to-fourth digit ratio predicts success among high-frequency financial traders. *Proceedings of the National Academy of Sciences*, 106(2):623.

Coates, J. M. and Herbert, J. (2008). Endogenous steroids and financial risk taking on a London trading floor. *Proceedings of the National Academy of Sciences*, 105(16):6167.

Colby, S. L. and Ortman, J. M. (2015). Projections of the size and composition of the U.S. population: 2014 to 2060. Washington, D.C.: U.S. Census Bureau.

Crawford, R. and O'Dea, C. (2012). The adequacy of wealth among those approaching retirement. IFS Report R72, Institute for Fiscal Studies, London.

Crossley, T. F. and O'Dea, C. (2010). The wealth and saving of UK families on the eve of the crisis. IFS Reports R71, Institut for Fiscal Studies, London.

Dabbs, J. M. (1992). Testosterone and Occupational Achievement. *Social Forces*, 70(3):813–824.

Dabbs, J. M., Bernieri, F. J., Strong, R. K., Campo, R., and Milun, R. (2001). Going on stage: Testosterone in greetings and meetings. *Journal of Research in Personality*, 35(1):27–40.

Dabbs, J. M., Strong, R., and Milun, R. (1997). Exploring the mind of testosterone: A beeper study. *Journal of Research in Personality*, 31(4):577–587.

Dabbs Jr., J. M., de la Rue, D., and Williams, P. M. (1990). Testosterone and occupational choice: Actors, ministers, and other men. *Journal of Personality and Social Psychology*, 59(6):1261–1265.

De Nardi, M., French, E., and Jones, J. B. (2010). Why do the elderly save? The role of medical expenses. *Journal of Political Economy*, 118:39–75.

De Ree, J., Alessie, R., and Pradhan, M. (2013). The price and utility dependence of equivalence scales: Evidence from Indonesia. *Journal of Public Economics*, 97:272–281.

Deaton, A. and Muellbauer, J. (1980a). *Economics and consumer behavior*. Cambridge University Press, Cambridge.

Deaton, A. and Muellbauer, J. (1986). On measuring child costs: With application to poor countries. *Journal of Political Economy*, 94:720–744.

Deaton, A. S. (1975). *Models and Projections of Demand in Post-War Britain*. Number 1 in Cambridge Studies in Applied Econometrics. Springer, Cambridge.

Deaton, A. S. and Muellbauer, J. (1980b). An almost ideal demand system. *The American Economic Review*, 70(3):312–326.

Delavande, A., Rohwedder, S., and Willis, R. (2008). Preparation for Retirement, Financial Literacy and Cognitive Resources. Technical report, University of Michigan, Michigan Retirement Research Center.

DellaVigna, S. and Paserman, M. D. (2005). Job search and impatience. *Journal of Labor Economics*, 23(3):527–588.

Dolls, M., Doerrenberg, P., Peichl, A., and Stichnoth, H. (2018). Do retirement savings increase in response to information about retirement and expected pensions? *Journal of Public Economics*, 158:168–179.

Donaldson, D. and Pendakur, K. (2004). Equivalent-expenditure functions and expenditure-dependent equivalence scales. *Journal of Public Economics*, 88:175–208.

Donaldson, D. and Pendakur, K. (2006). The identification of fixed costs from consumer behaviour. *Journal of Business and Economic Statistics*, 24:255–265.

Dreher, J.-C., Dunne, S., Pazderska, A., Frodl, T., Nolan, J. J., and O'Doherty, J. P.

(2016). Testosterone causes both prosocial and antisocial status-enhancing behaviors in human males. *Proceedings of the National Academy of Sciences*, 113(41):11633.

Dudel, C. (2015). Nonparametric bounds on equivalence scales. *Economics Bulletin*, 35:2161–2165.

Dudel, C., Garbuszus, J. M., Ott, N., and Werding, M. (2017a). Matching as nonparametric preprocessing for the estimation of equivalence scales. *Journal of Economics and Statistics*, 237(2):115–141.

Dudel, C., Garbuszus, J. M., Ott, N., and Werding, M. (2017b). Regelbedarfsermittlung für die Grundsicherung: Perspektiven für die Weiterentwicklung. *Sozialer Fortschritt*, 66(6):433–450.

Dudel, C., Ott, N., and Werding, M. (2016). Maintaining one's living standard at old age: What does that mean? *Empirical Economics*, 51:1261–1279.

Dudel, C. and Schmied, J. (2019). Pension adequacy standards: An empirical estimation strategy and results for the United States and Germany. MPIDR Working Paper 3, Max Planck Institute for Demographic Research.

Dudel, C., Schmied, J., and Werding, M. (2020). Sicherungsziele für die Rente: Empirische Messung und Ergebnisse. *Wirtschaftsdienst*, 100(3):185–193.

Eibich, P. (2015). Understanding the effect of retirement on health: Mechanisms and heterogeneity. *Journal of Health Economics*, 43:1–12.

Elagizi, A., Köhler, T. S., and Lavie, C. J. (2018). Testosterone and cardiovascular health. *Mayo Clinic Proceedings*, 93(1):83–100.

Engel, E. (1895). Die Productions- und Consumtionsverhältnisse des Königreichs Sachsen. In *Die Lebenskosten Belgischer Arbeiter-Familien früher und jetzt*. C. Heinrich, Dresden.

Faik, J. (2011). Der Zerlegungs-Ansatz – ein alternativer Vorschlag zur Messung von Armut. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 4(4):293–315.

Fan, Y., Guerre, E., and Zhu, D. (2017). Partial identification of functionals of the joint distribution of "potential outcomes". *Journal of Econometrics*, 197:42–59.

Ferrer-i-Carbonell, A. (2005). Income and well-being: An empirical analysis of the comparison income effect. *Journal of Public Economics*, 89:997–1019.

Ferrer-i Carbonell, A. and Frijters, P. (2004). How important is methodology for the estimates of the determinants of happiness? *The Economic Journal*, 114(497):641–659.

Finkelstein, A., Luttmer, E. F. P., and Notowidigdo, M. J. (2013). What good is wealth without health? The effect of health on the marginal utility of consumption. *Journal of the European Economic Association*, 11:221–258.

Fisher, G. M. (2007). An overview of recent work on standard budgets in the United States and other Anglophone countries.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42.

Friedman, M. (1957). *A Theory of the Consumption Function.* Princeton University Press.

Garbuszus, J. M. (2017). Schätzung nichtlinearer Gleichungssysteme in R mit nlsur. Mimeo. Ruhr-Universität Bochum.

Garbuszus, J. M. (2018). *Quadratische Engelkurven. Schätzung von nutzenabhängigen Äquivalenzskalen*, volume 86 of *Wissenschaftliche Beiträge aus dem Tectum Verlag. Reihe: Sozialwissenschaften.* Tectum Verlag, Baden-Baden.

García, J. and Labeaga, J. M. (1996). Alternative approaches to modelling zero expenditure: An application to Spanish demand for tobacco. *Oxford Bulletin of Economics and statistics*, 58(3):489–506.

Geyer, J. and Steiner, V. (2014). Future public pensions and changing employment patterns across birth cohorts. *Journal of Pension Economics & Finance*, 13(2):172–209.

Gielen, A. C., Holmes, J., and Myers, C. (2016). Prenatal testosterone and the earnings of men and women. *Journal of Human Resources*, 51(1):30–61.

Grabka, M. (2020). *SOEP Core v36: Codebook for PEQIVAL.*

Greene, F. J., Han, L., Martin, S., Zhang, S., and Wittert, G. (2014). Testosterone is associated with self-employment among Australian men. *Economics & Human Biology*, 13:76–84.

Greene, W. H. (2012). *Econometric Analysis.* Pearson Education Limited, Edinburgh Gate, 7 edition.

Gregg, P. (2001). The impact of youth unemployment on adult unemployment in the NCDS. *The Economic Journal*, 111(475):626–653.

Haag, B. R., Hoderlein, S., and Pendakur, K. (2009). Testing and imposing Slutsky symmetry in nonparametric demand systems. *Journal of Econometrics*, 153(1):33–50.

Haan, P. and Prowse, V. (2014). Longevity, life-cycle behavior and pension reform. *Journal of Econometrics*, 178:582–601.

Hagenaars, A. J., Vos, K., and Zaidi, M. A. (1994). *Poverty statistics in the late 1980s.* Theme / Statistical Office of the European Communities : 3, Population and social conditions : Series C, Accounts, surveys and statistics. Off. of Official Publ. of the Europ. Communities, Luxembourg.

Hall, P., Racine, J. S., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99:1015–1026.

Hamermesh, D. (1984). Consumption during retirement: The missing link in the life cycle. *Review of Economics and Statistics*, 66:1–7.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21:1926–1947.

Haveman, R., Holden, K., Romanov, A., and Wolfe, B. (2007). Assessing the maintenance of savings sufficiency over the first decade of retirement. *International Tax and Public Finance*, 14:481–502.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27:1–32.

Heckman, J. J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic processs. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications*, pages 179–195. MIT Press, Cambridge, Massachusetts.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3):411–482.

Hell, B. and Päßler, K. (2011). Are occupational interests hormonally influenced? The 2D:4D-interest nexus. *Personality and Individual Differences Research*, 51(4):376–380.

Henle, P. (1972). Recent trends in retirement benefits related to earnings. *Monthly Labor Review*, 95:12–20.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960.

Horowitz, J. and Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95:77–84.

Howe, H., Pollak, R. A., and Wales, T. J. (1979). Theory and time series estimation of the quadratic expenditure system. *Econometrica*, 47(5):1231–1247.

Hsu, C.-C., Su, B., Kan, N.-W., Lai, S.-L., Fong, T.-H., Chi, C.-P., Chang, C.-C., and Hsu, M.-C. (2015). Elite collegiate tennis athletes have lower 2D:4D Ratios than those of nonathlete controls. *The Journal of Strength & Conditioning Research*, 29(3):822–825.

Hughes, A. and Kumari, M. (2019). Testosterone, risk, and socioeconomic position in British men: Exploring causal directionality. *Social Science & Medicine*, 220:129–140.

Imbens, G. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635.

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

Juster, F. T. and Suzman, R. (1995). An overview of the Health and Retirement Study. *Journal of Human Resources*, 30:S7–S56.

Kahneman, D. and Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *Journal of Economic perspectives*, 20(1):3–24.

Kämpfen, F. and Maurer, J. (2016). Time to burn (calories)? The impact of retirement on physical activity among mature Americans. *Journal of Health Economics*, 45:91 – 102.

Kitao, S. (2014). Sustainable social security: Four options. *Review of Economic Dynamics*, 17:756–779.

Kluth, S. and Gasche, M. (2015). Ersatzraten in der gesetzlichen rentenversicherung/replacement rates in the German statutory pension system. *Jahrbücher für Nationalökonomie und Statistik*, 235(6):553–583.

Knoef, M., Been, J., Alessie, R., Caminada, K., Goudswaard, K., and Kalwij, A. (2016). Measuring retirement savings adequacy: Developing a multi-pillar approach in the Netherlands. *Journal of Pension Economics and Finance*, 15:55–89.

Kohn, K. and Missong, M. (2003). Estimation of quadratic expenditure systems using German household budget data. *Jahrbücher für Nationalökonomie und Statistik*, 223:422–448.

Koulovatianos, C., Schröder, C., and Schmidt, U. (2005). On the income dependence of equivalence scales. *Journal of Public Economics*, 89(5-6):967–996.

Kroft, K., Lange, F., and Notowidigdo, M. J. (2013). Duration dependence and labor market conditions: Evidence from a field experiment. *The Quarterly Journal of Economics*, 128(3):1123–1167.

Krueger, A. B. and Schkade, D. A. (2008). The reliability of subjective well-being measures. *Journal of Public Economics*, 92(8-9):1833–1845.

Lancaster, G. and Ray, R. (1998). Comparison of alternative models of household equivalence scales: The Australian evidence on unit record data. *Economic Record*, 74:1–14.

Lancaster, G., Ray, R., and Valenzuela, M. R. (1999). A cross-country study of equivalence scales and expenditure inequality on unit record household budget data. *Review of Income and Wealth*, 45:455–482.

Layard, R., Mayraz, G., and Nickell, S. (2008). The marginal utility of income. *Journal of Public Economics*, 92(8-9):1846–1857.

Lee, D. S. and Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655 – 674.

Leser, C. E. V. (1963). Forms of Engel Functions. *Econometrica*, 31(4):694–703.

Lewbel, A. (1989a). Household equivalence scales and welfare comparisons. *Journal of Public Economics*, 39:377–391.

Lewbel, A. (1989b). Nesting the AIDS and translog demand systems. *International Economic Review*, 30(2):349–356.

Lewbel, A. (2010). Shape-invariant demand functions. *Review of Economics and Statistics*, 92:549–556.

Lluch, C. (1973). The extended linear expenditure system. *European Economic Review*, 4:21–32.

Love, D. A., Smith, P. A., and McNair, L. C. (2008). A new look at the wealth adequacy of older U.S. households. *Review of Income and Wealth*, 54:616–642.

Luengo-Prado, M. J. and Sevilla, A. (2013). Time to cook: Expenditure at retirement in Spain. *Economic Journal*, 123:764–789.

Lundberg, S., Startz, R., and Stillman, S. (2003). The retirement-consumption puzzle: A marital bargaining approach. *Journal of Public Economics*, 87:1199–1218.

Lusardi, A. and Mitchell, O. S. (2007). Baby Boomer retirement security: The roles of planning, financial literacy, and housing wealth. *Journal of Monetary Economics*, 54:205–224.

Manski, C. (2003). *Partial Identification of Probability Distributions.* Springer.

Marcus, J. (2014). Does job loss make you smoke and gain weight? *Economica*, 81(324):626–648.

Mazur, A. (1985). A biosocial model of status in face-to-face primate groups. *Social Forces*, 64(2):377–402.

McArdle, J. J., Smith, J. P., and Willis, R. (2009). Cognition and economic outcome in the health and retirement survey. *National Bureau of Economic Research Working Paper Series*, No. 15266(published as John J. McArdle, James P. Smith, Robert Willis. "Cognition and Economic Outcomes in the Health and Retirement Survey," in David A. Wise, editor, "Explorations in the Economics of Aging" University of Chicago Press (2011)).

McFall, S. (2013). Understanding Society- UK Household Longitudinal Study: Cognitive ability measures. Understanding Society User Manual.

Mehta, P. H. and Prasad, S. (2015). The dual-hormone hypothesis: A brief review and future research agenda. *Social Behavior*, 3:163–168.

Merz, J. and Faik, J. (1995). Equivalence scales based on revealed preference consumption expenditures: The case of Germany. *Journal of Economics and Statistics*, 214(4):425–447.

Michelini, C. (2001). Estimating the cost of children from New Zealand quasi-unit record data of household consumption. *The Economic Record*, 77(239):383–392.

Mitchell, O. and Moore, J. (1998). Can Americans afford to retire? New evidence on retirement savings adequacy. *The Journal of Risk and Insurance*, 65:371–400.

Modigliani, F. and Brumberg, R. (1954). *Utility analysis and the consumption function: An interpretation of cross-section data*. Rutgers University Press.

Moran, P., O'Connell, M., O'Dea, C., Parodi, F., Submitter, M. R., et al. (2021). Heterogeneity in household spending and well-being around retirement. *University of Michigan Retirement and Disability Research Center (MRDRC) Working Paper*.

Moreau, N. and Stancanelli, E. (2015). Household consumption at retirement: A regression discontinuity study on French data. *Annals of Economics and Statistics*, 117/118:253–276.

Muellbauer, J. and van de Ven, J. (2004). Equivalence scales and taxation: A simulation analysis. In Dagum, C. and Ferrari, G., editors, *Household Behaviour, Equivalence Scales, Welfare and Poverty*, pages 85–106. Physica.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

Munnell, A. H. and Soto, M. (2005). What replacement rates do households actually experience in retirement? *Canadian Rights Reporter Paper*, (2005-10).

Murtagh, M. J., Blell, M. T., Butters, O. W., Cowley, L., Dove, E. S., Goodman, A., Griggs, R. L., Hall, A., Hallowell, N., Kumari, M., et al. (2018). Better governance, better access: practising responsible data sharing in the metadac governance infrastructure. *Human Genomics*, 12(1):1–12.

Nicholson, J. L. (1976). Appraisal of different methods of estimating equivalence scales and their results. *Review of Income and Wealth*, 22(1):1–11.

Nicolaou, N., Patel, P. C., and Wolfe, M. T. (2017). Testosterone and tendency to engage in self-employment. *Management Science*.

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

Nye, J. V., Bryukhanov, M., Kochergina, E., Orel, E., Polyachenko, S., and Yudkevich, M. (2017). The effects of prenatal testosterone on wages: Evidence from Russia. *Economics & Human Biology*, 24:43–60.

OECD (2008). Growing unequal? Income disbtribution and poverty in OECD countries.

OECD (2015). Pensions at a glance 2015. Available online at `http://www.oecd-ilibrary.org/social-issues-migration-health/pensions-at-a-glance-2015_pension_glance-2015-en`.

Ohlsson, C., Wallaschofski, H., Lunetta, K. L., Stolk, L., Perry, J. R., Koster, A., Petersen, A.-K., Eriksson, J., Lehtimäki, T., Huhtaniemi, I. T., et al. (2011). Genetic determinants of serum testosterone concentrations in men. *PLoS genetics*, 7(10):e1002313.

Palmer, B. A. (1989). Tax reform and retirement income replacement ratios. *The Journal of Risk and Insurance*, 56(4):702–725.

Pantoja, P., Hurd, M., Martin, C., Meijer, E., Rohwedder, S., and St Clair, P. (2017). Tax calculations for hrs 2000 and 2014. RAND Documentation.

Parslow, E., Ranehill, E., Zethraeus, N., Blomberg, L., von Schoultz, B., Hirschberg, A. L., Johannesson, M., and Dreber, A. (2019). The digit ratio (2D:4D) and economic preferences: No robust associations in a sample of 330 women. *Journal of the Economic Science Association*, 5(2):149–169.

Pendakur, K. (1999). Semiparametric estimates and tests of base-independent equivalence scales. *Journal of Econometrics*, 88:1–40.

Pendakur, K. (2002). Taking prices seriously in the measurement of inequality. *Journal of Public Economics*, 86(1):47–69.

Pendakur, K. and Sperlich, S. (2010). Semiparametric estimation and consumer demand systems in real expenditure. *Journal of Applied Econometrics*, 25:420–457.

Phipps, S. A. and Burton, P. S. (1995). Sharing within families: Implications for the measurement of poverty among individuals in Canada. *The Canadian Journal of Economics*, 28:177–204.

Phipps, S. A. and Garner, T. I. (1994). Are equivalence scales the same for the United States and Canada. *Review of Income and Wealth*, 40(1).

Pollak, R. A. (1991). Welfare comparisons and situation comparisons. *Journal of Econometrics*, 50:31–48.

Pollak, R. A. and Wales, T. J. (1979). Welfare comparisons and equivalence scales. *American Economic Review*, 69:216–221.

Ponzi, D., Zilioli, S., Mehta, P. H., Maslov, A., and Watson, N. V. (2016). Social network centrality and hormones: The interaction of testosterone and cortisol. *Psychoneuroendocrinology*, 68:6–13.

Pradhan, M. and Ravallion, M. (2000). Measuring poverty using qualitative perceptions of consumption adequacy. *Review of Economics and Statistics*, 82(3):462–471.

Pudney, S. (2011). Perception and retrospection: The dynamic consistency of responses to survey questions on wellbeing. *Journal of Public Economics*, 95(3-4):300–310.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria.

Rabe-Hesketh, S. and Skrondal, A. (2013). Avoiding biased versions of Wooldridge's simple solution to the initial conditions problem. *Economics Letters*, 120(2):346–349.

Racine, J. S. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119:99–130.

Ray, R. (1983). Measuring the costs of children: An alternative approach. *Journal of Public Economics*, 22(1):89–102.

Rothbarth, E. (1943). Note on a method of determining equivalent income for families of different composition. In Madge, C., editor, *War-time Pattern of Saving and Spending*, pages 123–130. Cambridge University press, Cambridge.

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society*, 16(5):754–760.

Samwick, A. A. (1998). New evidence on pensions, social security, and the timing of retirement. *Journal of Public Economics*, 70:207–236.

Sapienza, P., Zingales, L., and Maestripieri, D. (2009). Gender differences in financial risk aversion and career choices are affected by testosterone. *Proceedings of the National Academy of Sciences*, 106(36):15268.

Schaal, B., Tremblay, R. E., Soussignan, R., and Susman, E. J. (1996). Male testosterone linked to high social dominance but low physical aggression in early adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry*, 35(10):1322–1330.

Scheffter, M. (1991). *Haushaltsgröße und privater Verbrauch: Zum Einfluss einer steigenden Kinderzahl auf den privaten Verbrauch.* Lang, Frankfurt.

Schnabel, R. (2003). *Die neue Rentenreform: Die Nettorenten sinken.* Deutsches Institut für Altersvorsorge, Köln.

Scholz, J., Seshadri, A., and Khitatrakun, S. (2006). Are Americans saving 'optimally' for retirement? *Journal of Political Economy*, 114:607–643.

Schröder, C. (2004). *Variable Income Equivalence Scales.* Contributions to Economics. Physica-Verlag HD, Heidelberg.

Schröder, C. and Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, 92:337–358.

Schulz, J. and Carrin, G. (1972). The role of savings and pension systems in maintaining living standards in retirement. *Journal of Human Resources*, 7:343–365.

Schurer, S. (2017). Bouncing back from health shocks: Locus of control and labor supply. *Journal of Economic Behavior & Organization*, 133:1–20.

Schwarze, J. (2003). Using panel data on income satisfaction to estimate equivalence scale elasticity. *Review of Income and Wealth*, 49:359–372.

Schwerdt, G. (2005). Why does consumption fall at retirement? evidence from Germany. *Economics Letters*, 89(3):300–305.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the Matching package for R. *Journal of Statistical Software*, 42(7):1–52.

Skinner, J. (2007). Are you sure you're saving enough for retirement? *Journal of Economic Perspectives*, 21(3):59–80.

Smith, J. P. (2003). Trends and projections in income replacement during retirement. *Journal of Labor Economics*, 21(4):755–781.

Smith, S. (2006). The retirement-consumption puzzle and involuntary early retirement: Evidence from the British Household Panel Survey. *Economic Journal*, 116:C130–C148.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stengos, T., Sun, Y., and Wang, D. (2006). Estimates of semiparametric equivalence scales. *Journal of Applied Econometrics*, 21:629–639.

Stenstrom, E., Saad, G., Nepomuceno, M. V., and Mendenhall, Z. (2011). Testosterone and domain-specific risk: Digit ratios (2D:4D and rel2) as predictors of recreational, financial, and social risk-taking behaviors. *Personality and Individual Differences Research*, 51(4):412–416.

Stewart, M. B. (2007). The interrelated dynamics of unemployment and low-wage employment. *Journal of Applied Econometrics*, 22(3):511–531.

Stone, R. (1954). Linear expenditure systems and demand analysis: An application to the pattern of British demand. *The Economic Journal*, 64(255):511–527.

Sumner, R. C., Bennett, R., Creaven, A.-M., and Gallagher, S. (2020). Unemployment, employment precarity, and inflammation. *Brain, Behavior, and Immunity*, 83:303–308.

Szelky, M., Lustig, N., Cumpa, M., and Meja, J. A. (2004). Do we know how much poverty there is? *Oxford Development Studies*, 32:523–558.

Szulc, A. (2009). A matching estimator of household equivalence scales. *Economics Letters*, 103:81–83.

UK Pension Commission (2004). Pensions: Challenges and choices; the first report of the pensions commission.

United Nations (2015). World population ageing 2015, highlights. Department of Economic and Social Affairs, `http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015`.

Uysal, S. D. and Pohlmeier, W. (2011). Unemployment duration and personality. *Journal of Economic Psychology*, 32(6):980–992.

van Praag, B. M. S. (1991). Ordinal and cardinal utility. An integration of the two dimensions of the welfare concept. *Journal of Econometrics*, 50:69–89.

van Praag, B. M. S. and Kapteyn, A. (1973). Further evidence on the individual welfare function of income. *European Economic Review*, 4:32–62.

VanDerhei, J. and Copeland, C. (2010). The EBRI retirement readiness rating: Retirement income preparation and future prospects. EBRI Issue Brief 344.

Viinikainen, J. and Kokko, K. (2012). Personality traits and unemployment: Evidence from longitudinal data. *Journal of Economic Psychology*, 33(6):1204–1222.

Welker, K. M. and Carré, J. M. (2015). Individual differences in testosterone predict persistence in men. *European Journal of Personality*, 29(1):83–89.

Werding, M. (2016). One pillar crumbling, the others too short: Old-age provision in Germany. *National Institute Economic Review*, 237(1):R13–R21.

Wilke, F. (2014). Abschied von der Lebensstandardsicherung. Altersvorsorge im Spannungsfeld zwischen Unsicherheit und langfristiger Zielsetzung. *Sozialer Fortschritt*, 3/2014:58–65.

Wilke, R. A. (2006). Semi-parametric estimation of consumption-based equivalence scales: The case of Germany. *Journal of Applied Econometrics*, 21(6):781–802.

Wolfson, M. C. (2011). Projecting the adequacy of Canadians' retirement income. IRPP Study No. 17.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20(1):39–54.

Working, H. (1943). Statistical laws of family expenditure. *Journal of the American Statistical Association*, 38:43–56.

Declarations

Erklärung gem. §4 Abs. 2 der Promotionsordnung

Hiermit erkläre ich, dass ich mich noch keinem Promotionsverfahren unterzogen oder um Zulassung zu einem solchen beworben habe, und die Dissertation in der gleichen oder einer anderen Fakultät, einem Prüfungsausschuss oder einem Fachvertreter an einer anderen Hochschule nicht bereits zur Überprüfung vorgelegen hat.

Berlin, Dezember 2022

Julian Schmied

Erklärung gem. §10 Abs. 3 der Promotionsordnung

Hiermit erkläre ich, dass ich für die Dissertation folgende Hilfsmittel und Hilfen verwendet habe: Stata, R, Latex und Microsoft Excel.

Berlin, Dezember 2022

Julian Schmied